The Evolution and Development of Left / Right Asymmetry in the Lophotrochozoa

Nathan James Kenny
St Cross College

DPhil Thesis
Department of Zoology
The University of Oxford

2013

Abstract

The Evolution and Development of Left / Right Asymmetry in the Lophotrochozoa

Nathan Kenny, St Cross College, DPhil Zoology, Michaelmas Term 2013

Left/right (L/R) asymmetries, differences in morphology between the otherwise mirrored left- and right-hand sides of the body, are found in animals across the Bilateria. For many years it was thought that the mechanisms for establishment of these asymmetries had evolved separately in the three superphyla that constitute the Bilateria, but the discovery in 2009 that the TGF β ligand *Nodal* shares a conserved role in the Deuterostomia and Lophotrochozoa has re-ignited debate and interest in this field. In this thesis, work examining the establishment and maintenance of L/R asymmetries in the lophotrochozoan superphylum is presented, aimed at uncovering the wider conservation of these pathways across the Bilateria.

Illumina sequencing and a range of *de novo* assembly techniques were used to derive genomic and transcriptomic data respectively for two primary model organisms, the limpet *Patella vulgata* and the serpulid annelid *Pomatoceros lamarckii*. Additionally, collaborative work lead to the derivation of transcriptomes for two other mollusc species and the genome of the monogont rotifer *Brachionus plicatilis*. A range of analysis was performed on these novel resources and is detailed here, with particular reference to the transcription factor cassettes contained in these datasets.

These sequence resources formed the basis for examination of the breaking of initial symmetry in these model organisms. Known read-outs of correct establishment of L/R asymmetry, the expression of genes *Nodal* and *Pitx* on the right of the body, were codified in the course of normal development in *P. vulgata*. Pharmacological inhibitors of genes implicated in the establishment of L/R asymmetry, particularly ATPase ion channels, were then applied to embryos. After development, markers of normal development were assayed for signs of bilateral inversion. Although radialised phenotypes were observed, it is unclear whether these are specifically the result of L/R asymmetry defects. The localisation of ATPase mRNA and serotonin, often posited as a small molecule potential morphogen, were also assayed, although no conclusions could be drawn as to a role in the establishment of L/R asymmetry for these molecules, counter to some evidence from vertebrates.

Once symmetry is broken, the TGF β pathway is responsible for the communication, specification and maintenance of tissue identity across the L/R axis. The novel sequence resources described in this thesis provided a comprehensive window into this signalling cassette, and detailed here is a treatment of the TGF β pathway within the Lophotrochozoa. Ligand diversity has increased markedly in some clades, while signal transduction and regulatory steps are relatively unchanged.

This work has increased our knowledge of lophotrochozoan biology and particularly the mechanisms underpinning the establishment of asymmetry in this under-researched clade, however, much remains to be discovered about the ultimate origin of asymmetry itself.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Development of Left/Right Asymmetry in the Lophotrochozoa

The term 'left-right (L/R) asymmetry' refers to how organisms differ in growth and form on the left and right halves of their body. These asymmetries can be difficult to detect by external examination, but internally they can be quite obvious. For example, in mammals, the liver is often offset to one side, and the heart is enlarged on the left in order to pump blood more efficiently around the body (Romer, 1949). When L/R asymmetry is perturbed a variety of problems, including (in humans, clinically relevant) defects in growth and development can occur, to the detriment of the organism containing them (Ramsdell, 2005; Cohen Jr., 2012). The correct patterning of L/R asymmetry is therefore a crucial part of normal development.

For some time the identity of the factor or factors which determined asymmetry has been a mystery. In 1990 Brown and Wolpert proposed a theoretical solution to how organisms could reliably develop L/R asymmetries - the F molecule (Brown & Wolpert, 1990). This F molecule was hypothesised to be a chiral molecule, which would be oriented in relationship with previously existing (anterior-posterior and dorso-ventral) axes, and then would provide the directionality needed to determine the left and right sides of the body - perhaps by directionally transporting another molecule. The F molecule (if such an entity exists) has not yet been identified, but a number of candidates for such an effect have been postulated. While we have not identified the root cause of the establishment of asymmetry, we now understand how L/R asymmetry is patterned once established, and laboratories worldwide are making considerable strides towards understanding where it begins in the first

place.

It has recently been discovered that a key cell signalling molecule, known as *Nodal*, plays an important role in controlling the directionality of L/R asymmetry in vertebrates. Beginning with the discovery of left sided expression of *Nodal* in the chicken (Levin et al., 1995) a fraction under twenty years ago, a range of molecular evidence has shown how asymmetry is patterned and maintained. In all vertebrates thus examined, *Nodal*, and the transcription factor known as *Pitx2* (which is the downstream effector turned on by *Nodal*), act to pattern asymmetry on the left hand side of the body.

More recently, not just the mechanisms of patterning asymmetry but also potential mechanisms for breaking initial symmetry have been discovered. The discovery of fluid flow driven by cilia in a structure known as the "node" in vertebrates has revolutionised work in this regard in deuterostomes (a fact first noted by Nonaka et al. (1998)), but such a node is far from universally found across the Bilateria. Indeed, many apparent mechanisms of L/R asymmetry-breaking predate this event in embryogenesis in some species, and it seems a variety of mechanisms must exist and interact to break initial symmetry, with some elements conserved more widely than others. Due to a paucity of present research, however, we know little of the patterning of L/R asymmetry outside the narrow frame of reference provided by established laboratory models.

Recently, the discovery and publication of the existence of *Nodal* in snail species (Grande & Patel, 2009), has reawakened broad interest in the evolution of L/R asymmetry. This has lead to reexamination of the question of whether directional L/R asymmetry evolved independently in each of the three superphyla that constitute the Bilateria, as previously suspected due to evidence from the Ecdysozoa (Palmer, 1996, 2004), or whether the mechanisms of establishment of L/R asymmetry are shared ancestrally.

Snails are members of the Lophotrochozoa, a protostome grouping still underrepresented in the literature compared to those superphyla with more established model organisms in their midst. The Lophotrochozoa is the most phyla-rich and perhaps the most morphologically diverse superphylum, and as such provides a

wide range of potential organisms for study. For this work, representatives of the Annelida and Mollusca, the two most widely spread and speciose phyla within the Lophotrochozoa, were chosen for study.

Studying the mechanisms involved in establishing asymmetry in the Lophotrochozoa will give us a broader understanding of true conservation of mechanisms for breaking, establishing and maintaining asymmetry. Furthermore, it will provide an insight into how a fundamental aspect of body plan establishment may have affected the divergence and evolution of the bilaterian bauplan.

1.1.1 The Structure of this Introduction

This introductory section aims to give a broad overview of the general themes incorporated into the more specific work found in the chapters of this thesis. It begins by considering the three major themes which underly my substantive work in this field:

- What is asymmetry *sensu lato*, and how does it interrelate with embryonic anatomy?
- In what way could initial symmetry be broken?
- How is asymmetry then patterned and maintained?

The evolutionary considerations of the above are then introduced, with a perspective on historical understanding of L/R asymmetry. The relatively recent grouping of the Lophotrochozoa (Halanych et al., 1995), and the characteristic development of that clade, are then briefly explained, along with the asymmetries that can be found within these organisms. This introduction then concludes with an overview of the phylogeny, anatomy and development of the species used as models in this thesis to provide a framework for reference when interpreting the results found later in this work.

1.2 What are Asymmetries?

In order to understand how L/R asymmetry forms, we must first consider how it is distinct from more transient asymmetries which can occur in some organisms. We must also determine how it relates to the basic axes which define the form of a bilaterian animal, and the structures and tissues which act to construct them during embryogenesis.

By far the majority of metazoan life possesses bilateral symmetry, with only the most early branching phyla within Animalia - the likes of sponges, comb jellies and Cnidarians - exhibiting what has classically been referred to as radial symmetry (although it should be noted that a variety of work is now showing at least molecular signatures of bilaterality in some species in some of these clades). The development of true bilateral symmetry is thought to have coincided with the development of a head-like structure, a process known as cephalisation. In the course of this process, a centralised concentration of sensory structures developed around the mouth, which allowed these organisms to react quickly and efficiently, both to feed and to avoid danger (Moroz, 2012).

True bilateral symmetry is, however, rather rare. While organisms may appear to be symmetrical, this is often only skin deep - internally, the arrangement of organs often varies greatly from left to right (Neville, 1976). Formally, three types of asymmetry exist: (reviewed in Palmer (2004))

- Fluctuating asymmetries, which come about through environmental changes during development which alter the appearance of the organism on the left and right side of its midline. These are not heritable, but may be biased by environmental cues, for example; facial structures and brain lateralisation in humans may be influenced in this way (Harnad, 1977).

- Antisymmetry, in which sidedness is determined randomly from generation to generation, but the state of having a difference between the two sides is itself genetically determined. This has been noted in a wide range of organisms, from cichlid fish to barnacles (Palmer, 1996).

- Directional asymmetry refers to asymmetries which are fixed and heritable, sometimes within populations (such as isolated groups of snails (Schilthuizen & Davison, 2005)), but generally across members of a species. These asymmetries require a constant and tightly regulated developmental process to maintain.

The latter, directional asymmetry, is the focus of this work, as these are the ones that are thought to be homologous across the Bilateria. In some organisms directional L/R asymmetries are obvious, for example; shell coiling in gastropods, while other examples require internal examination - the hearts of mammals, for example, are larger on one side than the other. Directional asymmetries likely evolved from antisymmetries when an advantage was gained by having a particular asymmetry consistently on one side over many generations. In some species (such as humans) the directionality of asymmetry is almost totally fixed, with the exception of some individuals, who possess "*situs inversus*", the total mirror image inversion of asymmetry along the L/R axis.

Representatives of each of the three superphyla which comprise the Bilateria have evolved fixed directional (L/R) asymmetries, but all three possess markedly different patterns of initial growth and development. We will therefore cover what is known of the establishment of asymmetry in the more established model systems, before moving to the Lophotrochozoa and the peculiarities of cleavage and development in this clade.

1.3 Breaking Initial Symmetry

In order for L/R asymmetry to be reliably and repeatably established in an organism it is generally believed that a number of linked phenomena must proceed in concert in three progressive steps (Burdine & Schier, 2000).

- Firstly, initial symmetry must be broken in some way, differentiating and establishing the L/R axis in respect to the presumptive A/P and D/V axes.

- Secondly, asymmetric expression of genes begins, as positional information

must be established and converted into a form which can be widely communicated.

- Finally, transfer of this information to the wider organism can take place, and asymmetric growth begins in earnest.

It is the first of these steps which is considered in this section of the introduction. While the latter two parts of this process are relatively well described, as will become apparent later in this work, the steps that occur during the initial stage are only now starting to be investigated, and are the topic of sometimes heated debate. As with most fields of biological research, initial investigations focussed on model organisms, and in particular the mouse (*Mus musculus*), chick (*Gallus gallus*) and zebrafish (*Danio rerio*). While this has provided some vital insights, caution must be exercised in drawing more general conclusions across the wider animal kingdom from findings in these organisms, as is discussed later in this report with reference to the peculiarities of development within the Spiralia.

Many initial examinations into the establishment of L/R asymmetry looked for candidate pathways via a variety of drug treatments. These screens represented a straightforward and practical means of doing functional work at a time when more targeted approaches were unavailable, and a number of examples of these studies can be seen in Table 1.1. While many of these potential targets are still posited today as possible underlying portions of the L/R establishment cascade, the mechanisms by which these drugs had an effect are often yet to be explained. Some pharmacological effects, however, have been examined in more detail, and helped provide the first molecular-level understanding of the process. The actions of inhibitors of H^+/K^+ ATPases, for example, are described in more detail later in this report, and were one of the first well-understood mechanisms for establishing the differences between left and right hand sides of organisms.

More recently, the discovery that the movement of cilia within the node of mice and some other vertebrates could create an asymmetric fluid movement, along with compelling evidence from knock-down of genes involved in cilial motion, then led to the supposition that this 'Nodal flow' was the key symmetry breaking step, from

Table 1.1: Examples of Early Pharmacological Investigation into Left/Right Asymmetry (after Levin (2005))

Drug	Function	Species	Result	Reference
Cadmium	Heavy metal	Rat	Left limb deformities	Barr (1973)
Retinoic acid	Teratogen	Hamster	Heterotaxia	Shenefelt (1972)
Phenylephrine	Adrenergic agonist	Rat	Heterotaxia	Fujinaga & Baden (1991)
Nitrofurazone	Antibiotic	Rat	Right-sided hypoplasia	Greenaway et al. (1986)
A23187	Ca^{2+} ionophore	Xenopus	Heterotaxia	Toyoizumi et al. (1997)
RGD polypeptides	Block ECM interactions	Frog	Heterotaxia	Yost (1992)

which all later asymmetry flowed. If this were true, it was argued, chiral cilia motor proteins could represent the 'F' molecule, leading to downstream L/R asymmetry. The 'node' or 'primitive knot', is also called 'Hensen's node' in birds and 'Spemann's organiser' in *Xenopus laevis*. In some species it has been shown that the node forms directly opposite the site of entry of the sperm which fertilised the embryo (Lane & Sheets, 2006).

The cilial hypothesis is generally accepted as the primary means of establishing L/R asymmetry in at least the majority of vertebrates, at least as a means of propagating signal widely across the embryo. However, a number of competing hypotheses have emerged, as the node and the cilia within it do not universally seem to be involved in the establishment of asymmetry, and in many cases L/R asymmetry seems to be established (perhaps even intracellularly) well before cilial movement and Nodal flow takes place. While Nodal flow certainly seems to at least amplify the messages leading to whole-body L/R asymmetry in those organisms that exhibit it, many other factors also seem to be implicated in breaking asymmetry in the first instance (Vandenberg & Levin, 2010).

Some of these factors, such as the aforementioned H^+/K^+ ATPases, now seem to be more widely implicated in the break of initial symmetry across the Metazoa, while others have only been observed to have an affect in one or a handful of species. Three main hypotheses are proposed to act in the establishment of L/R asymmetry - Nodal flow as mentioned above, ion transport (via molecules such as H^+/K^+ ATPases) and, more fundamentally, a role for the cytoskeleton. These, along with other elements implicated into these broader schema, are the focus of much current research and as such are summarised below.

1.3.1 The Cilial Model

Perhaps the most popular theory for the origin of L/R asymmetry, courtesy of a number of remarkable papers in the last decade, is the cilial model of symmetry breaking. While some problems exist with generalising the findings of this model to the wider Metazoa there is no doubt that cilia and Nodal flow play a part in patterning and maintaining L/R asymmetry in at least some species.

Cilia were first implicated in the establishment of L/R asymmetry in the 1970s, when sufferers of Kartagener syndrome, which is characterised by *situs inversus* (where the left and right hand sides of the body are mirrored compared to the general population) and a range of respiratory faults were also found to have structural faults in their airway cilia (Afzelius, 1976; Pedersen & Mygind, 1976). It has also been known for some years that primary cilia defects can lead to L/R anatomic defects such as congenital heart disease (Kennedy et al., 2007; McGrath & Brueckner, 2003). The first understanding of how components of cilia are necessary for correct L/R asymmetry to form was gained in the mouse (for example Supp et al. (1997); Nonaka et al. (1998)).

Cilia are found on many eukaryotic cells, in either motile or non-motile forms. They are formed around a microtubule-based skeleton, called an axoneme, with non motile forms possessing 9 outer microtubule pairs (made up of a smaller α tubule joined to a semi-circular β tubule) and no inner pairs, and thus referred to as '9+0'. Motile forms generally also possess two singlet inner microtubules, which contribute to their ability to move (and are thus called '9+2' axonemes). While motile forms of cilia generally are of the 9+2 arrangement, cells in the ventral portion of the mouse node possess a single cilium with a 9+0 arrangement, which are nonetheless motile.

Microtubules provide an anchor point for a variety of other binding proteins, and molecular motors can use them as tracks to move up and down (Basu & Brueckner, 2008). To move, cilia use Dynein, a large protein which changes conformation upon ATP binding and hydrolysis. Dynein is attached firmly to the alpha tubules of the outer axoneme ring and attaches transiently to the β tubule found

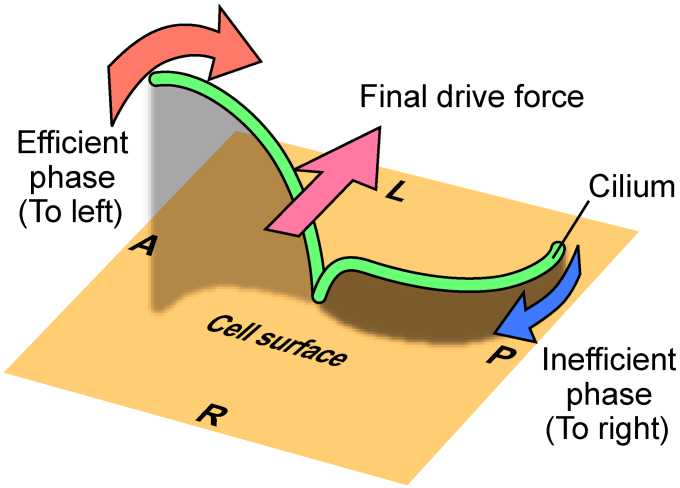


Figure 1.1: *By tilting the orientation of a cilia posteriorly, its clockwise motion can produce leftward flow. A cilium travelling leftwards (red arrow) moves fluid more efficiently than the returning rightward (blue arrow) movement, and net leftward fluid flow results. Figure reproduced from Nonaka et al. (2005)*

adjacent to it. As ATP is sequentially bound, hydrolysed and released the microtubules bend relative to one another, driving the movement of the cilia (Burgess et al., 2003).

The first studies to confirm the involvement of cilia in the establishment of L/R asymmetry knocked out or created mutations within important cilial proteins linked to cilia movement, such as Dynein axonemal heavy chain 11 (contracted to Dnahc11 or Lrd) and KIF3 (Supp et al., 1997; Nonaka et al., 1998; Okada et al., 1999; Supp et al., 1999; Takeda et al., 1999). These studies showed that cilial movement somehow affected the establishment of L/R asymmetry. Nonaka et al. (2002) were able to rescue and modulate correct L/R asymmetry by artificially providing a flow of fluid across the node in mice with ciliary defects, demonstrating that it was the flow itself that played a role in establishing L/R asymmetry.

How this flow was generated was discerned in Cartwright et al. (2004) and demonstrated by Nonaka et al. (2005), and can be seen in Fig. 1.1. By tilting the orientation of a cilia posteriorly, its clockwise motion can produce leftward flow. A cilium travelling leftwards, almost at right angles to the surface of the cell, moves fluid more efficiently than the returning rightward portion of its movement, which travels close to the cell surface, and net leftward fluid flow results.

While it is now known how ciliary movement creates Nodal flow, how this acts to activate Nodal signalling and induce L/R asymmetry has been the subject of some debate. Two kinds of cilia are observed in the mouse node, the motile forms

responsible for the flow noted above, and a second, immotile form, which some evidence suggests are actually responsible for establishing asymmetry, in spite of the elegant solution suggested by Cartwright et al. (2004) and Nonaka et al. (2005). This has resulted in two alternative explanations for the role of Nodal cilia in establishing asymmetry, each with evidence for and against its veracity.

- The "Chemical Gradient Model", which suggests that a unidirectional concentration gradient is produced by the action of Nodal cilia. This gradient then functions to turn on and off various downstream targets, which set up the differences which determine left/right asymmetry (Nonaka et al., 1998; Okada et al., 1999). 20kDa - 40kDa proteins have been observed to form a gradient concentration via Nodal flow (Okada et al., 2005). This has lead to the slightly modified form of this model, by which extracellular particles called Nodal vesicular parcels (NVPs) act as morphogen transporters (containing lipophilic granules, sonic hedgehog and retinoic acid) (Tanaka et al., 2005). The absorption of these at the left of the node has been noted, and would further confirm the gradient model. However, this theory fails to explain the role of immotile cilia and their constituent components, which are found in the node and seem to be intimately involved in the process of L/R asymmetry establishment.

- The "Two Cilia Model", which is more recent (McGrath et al., 2003; Tabin & Vogan, 2003), posits that immotile cilia are also involved in the process, as suggested by altered *Pitx* and downstream gene expression when immotile cilia are absent from the mouse node (Nonaka et al., 1998; Marszalek et al., 1999; Takeda et al., 1999; Murcia et al., 2000) and states that the left sided pressure created by motile cilia is sensed in turn by immotile cilia. This explains why some L/R asymmetry, although randomised, is seen in mice lacking only the motile cilia. The immotile cilia still present register random fluctuations in flow, and translate them into L/R asymmetry, either on the left, right, or potentially (although rarely) on both sides (Supp et al., 1997; Collignon et al., 1996; Lowe et al., 1996). In contrast, when both motile and immotile cilia are

absent from the node, genes such as *Pitx* are expressed bilaterally or not at all (Nonaka et al., 1998; Marszalek et al., 1999; Takeda et al., 1999; Murcia et al., 2000).

Recent evidence may support the two cilia model, as mutations in *Pkd2*, a calcium ion release channel controlled by sensory cilia, cause defects in L/R patterning, and intracellular calcium ion release can only be observed on the left side of the node (for examples, see Raya & Belmonte (2006); Field et al. (2011)). Cilia have also been observed in a homologous position in non-vertebrate chordates (Nishide et al., 2012; Thompson et al., 2012). This may indicate that a role for cilia is ancestral in the establishment of L/R asymmetry in chordates.

Cilia in Nonvertebrate Chordates

It has recently been shown that the embryos of *Ciona intestinalis* develop monocilia on endodermal cells immediately preceding the onset of asymmetric gene expression (Konno et al., 2010; Thompson et al., 2012). The axoneme structure of these cilia resembles the 9+0 arrangement found in cilia in the mouse node, although examination suggests that they are non-motile. They are also positioned relative to the AP axis, towards the rear of each cell. Coupled with evidence that cilia are involved functionally in driving consistently counter-clockwise rotation within ascidians, an event that is essential for correct establishment of L/R asymmetry, and may sense the resulting forces from this event (Nishide et al., 2012), it seems likely that a role for cilia in establishing asymmetry was present at least as far back as the common ancestor of vertebrates and ascidians.

There have been no studies to date which have implicated cilia in establishing asymmetry in other deuterostomes, such as the sea urchin and amphioxus, despite a long history of embryonic research in these organisms (e.g Garcia-Fernandez & Benito-Gutierrez (2009); Warner et al. (2012)). Cilia are found in the first asymmetric morphological structure to form in sea urchin larvae, the hydroporic canal, but this is specified downstream of already extant molecular asymmetries (Luo & Su, 2012). It is therefore possible that a role for cilia in establishing L/R asymmetry is a

Olfactores innovation, and the ancestral role of cilia in establishing L/R asymmetry could be a sensory one - a role played by cilia in many other contexts (Satir & Christensen, 2007) - or motile, as evidenced by their role in moving the embryo within the vitelline membrane.

Evidence Against an Ancestral Role

In spite of the overwhelming evidence for a role for cilia in propagating asymmetry, there remain two major arguments against the cilia model as the primary and ancestral method for performing the initial break in symmetry across the Metazoa. The first is that the 'Node' structure is far from universal across animal phylogeny, and vertebrates such as *Xenopus*, deuterostomes like the Sea Urchin *S. purpuratus*, as well as all examined protostomes break symmetry either prior to the formation of, or completely without, a Node structure (Levin, 2005).

Cilia are also by no means ubiquitous in the node. Both chickens and pigs (Gros et al., 2009) lack Nodal flow entirely or lack motile cilia in the region where this would occur. Ciliary movement cannot therefore be universally responsible for breaking L/R asymmetry, although it might have superseded some aspects of an ancestral process found in other organisms.

The second major argument against the ciliary model as the method of breaking symmetry is that many signs of L/R asymmetry occur before motile cilia do (Levin & Palmer, 2007). Molecular asymmetries are particularly well noted in early embryos prior to the development of the node, in a range of organisms including *Xenopus*, chick and zebrafish (Tabin, 2005; Kawakami et al., 2005; Levin, 2005; Qiu et al., 2005; Spéder et al., 2007). In chicks, for example, activin receptors are found on the right of the primitive streak - the precursor of the node- before cilia are motile (Tabin, 2006) and the asymmetric expression of ATPases and serotonin has been noted in *X. laevis* as early as the four cell stage (Levin et al., 2002).

If we are to understand the establishment of L/R asymmetry outside the vertebrates and the evolution of this process, we must therefore look beyond the motile ciliary paradigm. While there is unquestionably a role for cilia in the establishment of asymmetry in vertebrates, it does not tell the whole story, and the initial break in

symmetry must occur elsewhere. Given the known evidence of asymmetric gene expression and ion movement prior to the formation of the node (Pagan-Westphal & Tabin, 1998; Levin & Mercola, 1999; Levin et al., 2002), a number of alternative models for the establishment of L/R asymmetry have been proposed, and should be considered.

1.3.2 The Cytoskeleton and L/R Asymmetry

Even more fundamentally than the operation of cilia, the cytoskeleton may provide the basic information required for breaking symmetry or establishing L/R asymmetry. Whether it does this due to some inherent asymmetry (perhaps with some structure representing the 'F' molecule as defined by Wolpert and explained in Section 1.1), or alternatively through some aspect of its operations remains to be seen and fully understood.

The cytoskeleton has been linked to L/R asymmetry in a variety of organisms, even those as simple as Protozoans, whose parapodia extend preferentially leftwards by virtue of the intrinsic structure of the cytoskeleton (Nelsen et al., 1989). It seems that the cytoskeleton itself imposes as a result of microtubule structure an inherent L/R bias on cells, one which is disrupted when microtubule and dynein functionality is blocked (Xu et al., 2007). It has been known for quite some time that the underlying cytoskeletal architecture could be crucial for L/R asymmetry to be formed correctly in vertebrates (Yost, 1991). Evidence has supported a role for microtubules (as stated above), spindles and actins, with their associated molecular motors, although how these interrelate remains a mystery. It has remained difficult to disentangle the role of the cytoskeleton from the effects of the proteins which are supported and transported by it, and its exact role in the establishment of L/R asymmetry remains an open question (Adams et al., 2006; Vandenberg & Levin, 2010). The underlying structure and organisation of the cytoskeleton, however, seems to lend itself well to an organisational role if the identity of polarising signals can be found (Qiu et al., 2005; Vignaud et al., 2012)

Actin, via myosin, has been noted as playing a role in the establishment of L/R asymmetries in *Drosophila melanogaster*, with the conserved myosin 31DF (My-

oID) gene identified as the factor responsible for dextral looping of genitalia in flies (Spéder et al., 2006). These myosins act as molecular motors, moving towards the barbed end of F-actin, and thus only in a single direction relative to these. Myosins have also been shown to be involved in intrinsic asymmetry in *X. laevis* (Danilchik et al., 2006), and actins have been shown necessary for the correct arrangement of ATPases in this species (Adams et al., 2006). If actin is somehow organised relative to the A/P and D/V axes, perhaps by sperm entry point (Aw et al., 2008), this may mean that either actin or MyoID represent the hypothetical 'F' molecule (Spéder et al., 2007) - although the veracity of this claim is yet to be properly tested.

Intriguingly for the current research project, actin cytoskeletons have been linked with the establishment of L/R asymmetry in snails for several years (Shibazaki et al., 2004). The observations of Kuroda et al. (2009), that physical alteration of the direction of cell cleavage, and thus alteration of the cytoskeletal orientation, has a direct downstream effect on the expression of *Nodal* and thus the direction of L/R asymmetry have only leant further support for a role for the cytoskeleton. The development and genetic control of the two alternate L/R chiral forms of snails is discussed further in section 1.7.5, but the implications are clear for the present investigation, and it seems likely that actin has some conserved underlying role in establishing L/R asymmetry, as an enabler if not as a driver.

Microtubules have not been linked with the establishment of L/R asymmetry in snails (Shibazaki et al., 2004), but along with the evidence noted by Xu et al. (2007), microtubule use and spindle orientation has been shown to be important in L/R asymmetry establishment in a range of other species. In *C. elegans*, micromanipulation at the 6-cell stage has been shown to be able to reverse the normal L/R asymmetry (Wood, 1991). This has been shown to be a result of G proteins orienting the spindle (Bergmann et al., 2003), a finding that seems to be mirrored in *D. melanogaster* (Ahringer, 2003) without direct evidence for a role in L/R asymmetry. Microtubule-related motor proteins also seem to have a definite role in establishing L/R asymmetry in *X. laevis* (Qiu et al., 2005; Aw et al., 2008), along with playing a major role in cilia motion as noted earlier in this report. The answer to how actin and myosin interrelate with microtubules and potentially regulate L/R asymmetry

probably lies in the spindle, where F-actins seem to regulate spindle assembly and ensure correct functionality (Sandquist et al., 2011). This is consistent with some data from the likes of *X. laevis* (Qiu et al., 2005), which shows that disrupting microtubule orientation leads to errors in cytoplasmic localisation of potentially key proteins such as KIF3B and Left/Right Dynein, but the evidence for this is not always convincing and further study is needed to untangle the threads which link this to the establishment of L/R asymmetry.

It has therefore been postulated that the cytoskeleton either acts as the basis for chiral distribution of true determinant molecules (Qiu et al., 2005; Aw et al., 2008) or, perhaps alternatively, it itself is the chiral molecule from which other determinants take their cues (Spéder et al., 2007; Aw & Levin, 2009). Neither of these hypotheses has thus far been adequately proven. It seems likely that the cytoskeleton plays at least some part in the establishment of L/R asymmetry - but whether it is the primary source or merely the primary mechanism of propagating signal remains to be seen.

Chromatid Imprinting and Movement

Somewhat intertwined with cytoskeletal asymmetry is the alternative theory that differential chromatid imprinting provides the initial cue for establishing L/R asymmetry, with imprinted chromatids move in particular directions by an asymmetrically operating cytoskeleton (Armakolas & Klar, 2007; Armakolas et al., 2010; Klar, 1994). Yeast exhibits a well studied asymmetry in chromatid segregation, reviewed in Armakolas et al. (2010). Some structures within the cytokinesis apparatus of snails are asymmetrically organised, and suggest a mechanism by which differences in cleavage could come about - particularly as Left/Right Dynein, which has been implicated in asymmetry in a variety of vertebrate species, has a role in this structure (Meshcheryakov & Beloussov, 1975; Armakolas & Klar, 2007; Klar, 2008). It is especially possible that the microtubule organizing center, which is undoubtedly chiral given its complexity, could ultimately be the source of chiral transport of mRNA/protein and cilial localisation and functionality (Beisson & Jerka-Dziadosz, 1999). This possible ultimate mechanism for determining the chirality

of L/R asymmetry is fascinating, but has yet to be fully explored (Vandenberg & Levin, 2010).

1.3.3 Ions, Ion Channels and Gap Junctions

A contentious candidate for breaking asymmetry is a role for ions and ion channels in differentially distributing charge around embryos, as a result of differential movement or deposition of these molecules in embryos from early stages of development. Ion involvement was one of the first potential regulators of L/R asymmetry to be investigated, as discussed in Section 1.3. Early pharmacological investigations implicated a number of ions and ion channels in a role in establishing L/R asymmetry, but until recently many of these had not been investigated in detail.

One of the first of these re-examinations took place in the chicken, where a region of depolarisation to the left of the primitive streak is the first notable sign of L/R asymmetry (Levin et al., 2002). 'Gastric' H^+/K^+ ATPase ion channels have been shown to have a transient difference in activity between the left and right hand sides of the chick node, with slightly weaker H^+/K^+ ATPase activity creating a slight rise in the level of extracellular Ca^{2+} in that region. This then leads to raised left-sided levels of Notch signalling, Shh concentration, and ultimately *Nodal* expression, leading to L/R asymmetry (this is explained further in Section 1.3.4) (Levin et al., 2002; Raya et al., 2004).

Unlike *X. laevis* and chickens, zebrafish seem to utilise both H^+/K^+ and Na^+/K^+ channels to set up ion gradients for the establishment of L/R asymmetry (Kawakami et al., 2005; Ellertsdottir et al., 2006). Protostomes do not seem to possess H^+/K^+ ion channel orthologues, but as these only seem to be present in vertebrate chordates and a role for ion channels in establishing asymmetry in *Ciona intestinalis* and the sea urchin is known (Shimeld & Levin, 2006; Bessodes et al., 2012), it seems likely that Na^+/K^+ ion channels could perform this function in the Protostoma. H^+/V ATPases and potassium channels are also good candidates for an assisting role in L/R asymmetry determination (See Table 1, Spéder et al. (2007)), although evidence for their role is phylogenetically patchy compared to sodium and hydrogen potassium ATPases (Levin, 2005).

Protostomes have been shown to use ion channels in the context of L/R asymmetry, suggesting that this may be shared across the Bilateria. *C. elegans* utilise calcium ion channels and gap junctions to specify left and right olfactory neurons, each with their own functions and gene expression profiles (Poole & Hobert, 2006). H^+/K^+ ATPases in *Dugesia japonica* have also been implicated in the correct regeneration of the head and eye, providing information via endogenous differences in function on the left and right sides of the body (Nogi et al., 2005).

As gap junctions and an established ventral midline (with fewer gap junctions) are also known to play an important role in establishing L/R asymmetry (Levin & Mercola, 1998, 1999; Esser et al., 2006), it has been hypothesised that differential localisation of H^+/K^+ ATPases creates a cellular voltage potential skewed to one side of the very early cleavage stage embryo. Any charged molecules small enough to diffuse through the gap junctions will then become arranged in a L/R asymmetric distribution (Levin et al., 2002; Adams et al., 2006) If these molecules are morphogens, L/R asymmetry can then be patterned by them (Fukumoto et al., 2005b; Esser et al., 2006). This hypothesis is known as the "ion flux model" of asymmetry.

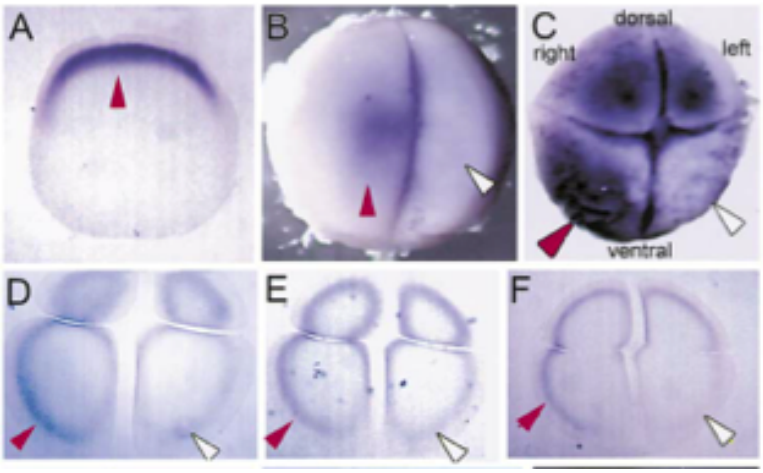


Figure 1.2: X. laevis H^+/K^+ ATPase α *mRNA distribution from Levin et al. (2002), at 1 cell (A), 2 cell (B) and 4 cell stages (C-F), visualised form the animal pole. mRNA is preferentially localised in the right ventral blastomere (red arrow) and depleted in the left ventral blastomere (white arrow)*

This model for H^+/K^+ ATPases in establishing L/R asymmetry was first proposed in *X. laevis*. mRNA coding for ATPases has been shown to be differentially expressed from the four cell stage, with some cells inheriting much greater quanti-

ties of H^+/K^+ *ATPase* α subunit mRNA, as can be seen in Fig. 1.2 (Levin et al., 2002). This has also been noted in the sea urchin (Hibino et al., 2006) but has not been observed in other vertebrates, and it has been suggested that the differential H^+/K^+ ATPase activity seen in other species instead results from translational modulation, rather than transcript - level transport (Raya & Belmonte, 2006). Zebrafish, for instance, have been found to suffer from asymmetry defects upon treatment with ATPase inhibitors, but no differential localisation of mRNA has been reported (Kawakami et al., 2005).

The ion channel/gap junction hypothesis does not provide an ultimate mechanism for the establishment of L/R asymmetry - some force or control must exist to arrange H^+/K^+ ATPases in such a fashion in the first place. Various explanations have been proposed for how this mechanism might operate, including asymmetric transport mechanisms (Qiu et al., 2005; Aw et al., 2008), as covered in more detail in Section 1.3.2 of this thesis.

However, while ATPases have been shown to have a role in establishing asymmetry in a variety of organisms, the model proposed in Levin et al. (2002) for this has proven controversial. In results reported in Beyer et al. (2012), investigations into the localisation of serotonin were unable to replicate the findings reported earlier in support of the ion flux model (Fukumoto et al., 2005b). Furthermore, Flachsova et al. (2013) could not find any evidence for asymmetric localisation of mRNA transcripts across the L/R axis of *Xenopus laevis* embryos through to the 32 cell stage. However, it should be noted that they did not look for ATPase transcripts in their qRT RT-PCR based approach. Applying this technique to the question of early ATPase localisation would seem to be an excellent test of the ion channel hypothesis, although definitive proof of localisation at the protein level would be the last word on this.

1.3.4 Serotonin and Other Potential 'Morphogens'

A variety of small, charged potential 'morphogens' have been proposed to act downstream of the charge gradient established by ion channel operation, or shown to have a wider role in the establishment of asymmetry independent of this hypoth-

esis. Some of those with the most evidence for their involvement include serotonin, inositol polyphosphates and Ca^{2+} ions, although others have also been proposed.

Wnt Signalling and the Serotonin Hypothesis

Serotonin, also known as 5-hydroxytryptamine (5-HT), is more often encountered in its role as a neurotransmitter but has been established as being asymmetrically distributed in the early embryos of *X. laevis*, as can be seen in Fig. 1.3 (Fukumoto et al., 2005b,a). Serotonin was also shown to be required for the normal establishment of L/R asymmetry, and when H^+/K^+ ATPase activity was blocked by chemical inhibition, asymmetric distribution of serotonin was not observed. Interestingly, it has recently been shown that serotonin may have a role in regulating the activity of Na^+/K^+ ATPases (Zhang et al., 2012b). This could mean that serotonin is the effector, rather than, as has previously been assumed, being the affected molecule downstream of ATPase activity, although further research is needed to establish the universality of this finding in other species.

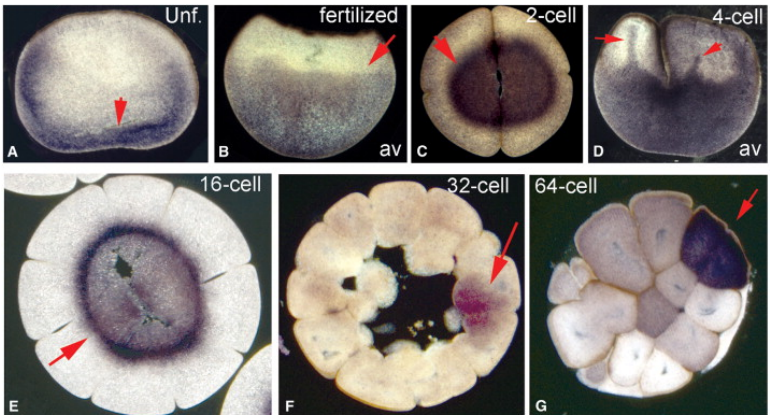


Figure 1.3: *The distribution of serotonin in the developing embryos of* X. laevis, *determined by Fukumoto et al. (2005b) using immunohistochemistry. Sections are shown perpendicular to the animal/vegetal axis.*

As noted above, Beyer et al. (2012) were unable to recapitulate the findings of Fukumoto et al. (2005b) with regards to serotonin localisation, finding it instead to be evenly distributed across the entirety of the embryo. This evidence therefore runs counter to the serotonin/ion channel hypothesis, as both exogenously provided and endogenous serotonin seemed to be stored in vesicles, and is thus not

available to diffuse across gap junctions. However, functional testing confirmed that serotonin is essential for the correct establishment of L/R asymmetry (Beyer et al., 2012), with interesting downstream effects via the Wnt pathway.

It has been known for some time that the *X. laevis Nodal* paralog *Xnr3* is a downstream target of canonical Wnt signalling (Smith et al., 1995; Glinka et al., 1996). It has been found that serotonin signalling induces the specification of superficial mesoderm in *Xenopus*. This mesoderm gives rise to the portion of the node in which Nodal flow occurs, the gastrocoel roof plate (Beyer et al., 2012), a process which is mediated by Wnt signalling. This provides a causative link between serotonin distribution and asymmetry, but, as noted above, the authors were unable to discern serotonin localisation in the same fashion as Fukumoto et al. (2005b). Instead it is proposed that serotonin and Wnt signalling act in concert, specifying the superficial mesoderm and aiding basic axis formation - from which the node structure is able to follow, with L/R asymmetry establishment following according to the cilial model detailed earlier.

Inositol polyphosphates

In zebrafish embryos, inositol polyphosphates could be transported by differences in charge post-production, thus creating a L/R asymmetric gradient alongside differences in Ca^{2+} levels, in partial concordance with the ion flux model of asymmetry establishment (Sarmah et al., 2005). However, inositol polyphosphates may actually lie directly upstream of Ca^{2+} function and thus be a link in the chain towards asymmetry, rather than an asymmetrically distributed morphogen itself (Sarmah et al., 2007).

Calcium Signalling

Calcium is known to play a part in various stages of the establishment of L/R asymmetry in vertebrates, including mice (McGrath et al., 2003) and chickens (Raya et al., 2004; Langenbacher & Chen, 2008). These often represent the first asymmetric signal following apparent Nodal flow, although whether asymmetric Ca^{2+} levels result from this or occur via another mechanism is still unclear. For example, in

chick embryos, where Nodal flow does not occur, elevated extracellular Ca^{2+} levels are found on the left side of the Hensen's node upstream of Nodal expression (Raya et al., 2004). It is hypothesised that this increase affects Notch signalling, through elevated expression of the Notch ligand Delta-like 1, which in turn activates Nodal signalling. A requirement for Notch upstream of Nodal signalling has also been reported in other vertebrates, as discussed in Section 1.5.1 (Krebs et al., 2003; Gourronc et al., 2007).

In zebrafish, elevated levels of Ca^{2+} are also found on the left of the node (Sarmah et al., 2005). These are not the first role for calcium in patterning zebrafish asymmetry, however. Fluxes of calcium levels appear to regulate Kuppfer's vesicle formation, (Schneider et al., 2008), and without a Kuppfer's vesicle to act as a node L/R asymmetry does not appear (Essner et al., 2005). Calcium levels also seem to play a role in proper cilial functionality, as when calcium homeostasis within cilia is disrupted, so is L/R asymmetry establishment (Shu et al., 2007; Sarmah et al., 2007).

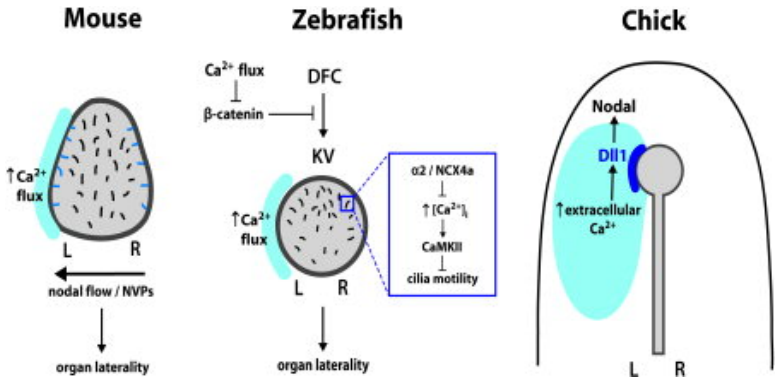


Figure 1.4: *Proposed mechanisms by which calcium is involved in the specification of asymmetry in mouse, zebrafish and chick. Respectively, calcium levels change in response to Nodal flow or NVP binding in mice, seem to pattern and control cilial operation in zebrafish, and regulate Notch signalling in the chicken. In all cases asymmetric levels of Ca^{2+} are then observed, with higher concentrations on the left. Figure reproduced from Langenbacher & Chen (2008).*

Elevated levels of calcium on the left of the node are also seen in mouse embryos, and are partially explained by the 'Two Cilia' model, which states that Pkd2 (a Ca^{2+}-permeable channel) expressing mechanosensory cilia release it asymmetrically in response to Nodal flow (Tabin & Vogan, 2003; McGrath et al., 2003). Tanaka

et al. (2005) showed that blocking the Nodal Vesicular Particle (NVP) signals but not Nodal flow itself is sufficient to prevent asymmetric Ca^{2+} levels from arising, although how NVPs would cause this is still unexplained.

The asymmetric localisation of calcium around the node is therefore well catalogued within vertebrates, and in many cases represents the first instance of molecular asymmetry to be catalogued in these species. More study is, however, required to investigate how calcium signalling activates Nodal, and whether its role, and the immediately downstream role for Notch signalling seen in the chicken, is seen more broadly across the animal kingdom.

Hedgehogs, Retinoic Acid and Juvenile Hormone

The common signalling molecules *Sonic Hedgehog* (Shh) and Retinoic Acid (RA) have for many years been noted as playing a positive role in the establishment of L/R asymmetry upstream of calcium signalling (Langenbacher & Chen, 2008), although their exact effects, and how they control asymmetric calcium release, are yet to be fully explained. It is mentioned elsewhere in this report (section 1.3.1) that small membrane bound vesicles (Nodal Vesicular Particles, henceforth NVPs) may be involved in transferring the signals from cilia and promoting *Nodal* expression on the left side of the developing embryo in mice (Tanaka et al., 2005). These NVPs contain Shh and RA, and their production is stimulated by FGF signalling (Langenbacher & Chen, 2008). When Fgf signalling is inhibited, the production of NVPs is impaired, and asymmetric calcium localisation and L/R asymmetry is prevented (Tanaka et al., 2005). These molecules are therefore likely to be closely linked to Nodal flow, and it is as yet unknown how conserved their action will thus be across the Bilateria.

The role of Shh upstream of calcium (and thus Nodal) signalling in the chick embryo was noted early in investigations into this subject (Levin et al., 1995). Other hedgehog molecules rather than Shh itself may play this role in the mouse, but this is still uncertain (Zhang et al., 2001; Tanaka et al., 2005; Raya & Belmonte, 2006). In Zebrafish, mice and *X. laevis* ectopic provision of Shh on the right resulted in bilateral *Nodal* expression (Levin et al., 1995; Sampath et al., 1997; Schilling et al.,

1999) and it can be confidently hypothesised that Shh plays a conserved upstream role in these species.

It has been noted that one of the targets of hedgehog signalling, *Gli* transcription factors, make use of cilia as a base for their actions (Cohen Jr., 2012). In the absence of hedgehog binding, Gli3, on the anterograde tip of cilia, is cleaved and acts as a transcriptional repressor, whereas when hedgehog signalling is present it switches to the retrograde side and activates downstream targets (Pan & Wang, 2007; Cohen Jr., 2010). The hedgehog targets Gli2 and Gli1 have also been implicated in these actions on cilia, although to a lesser degree (Cohen Jr., 2012). This suggests a mechanism by which hedgehog signalling, cilial sensation and calcium signalling may be linked, with Shh provision to cilia (through whichever model) leading directly to downstream activation.

RA also plays a role upstream of calcium signalling. In mice, unless maternal RA is cleared from the embryo, correct L/R asymmetry cannot be established (Uehara et al., 2009). In zebrafish development, RA has been shown to control aspects of L/R asymmetry (Kawakami et al., 2005; Huang et al., 2011), possibly through modulation of somitogenesis. RA inhibition interferes with the proper control of Nodal expression, which should be present in the left lateral plate mesoderm. When RA is knocked out before the '2-somite' stage, early in development, *spaw* (one of the three zebrafish homologues of Nodal) is expressed bilaterally in the lateral plate mesoderm, instead of purely on the left. This is due to increased Nodal cilia length, possibly as RA regulation of Fgf8, which controls this, is defective. Incorrect fluid flow then results in the bilateral expression of *Nodal*, as noted earlier, resulting in randomised L/R asymmetries in visceral growth. Later RA inhibition also has effects on Nodal signalling, but these are difficult to disentangle from its co-occurring effect on BMP4 signalling.

Interestingly, while *Nodal* is not found in *D. melanogaster*, the arthropod signalling molecule Juvenile Hormone has been shown to provide a signal 'bridge' between cytoskeletal cues and developing asymmetries (Adam et al., 2003; Okumura et al., 2008). Juvenile Hormone is similar in structure to RA, and it is interesting to consider whether this role could be a result of gradual divergence from an

ancestral state, or perhaps may have come about through convergent evolution.

1.3.5 Other Factors Implicated in L/R Asymmetry Establishment

Numerous other pathways and mechanisms have been implicated in the establishment of L/R asymmetry, as yet each only in a subset of model organisms. It is not always known how these mesh together, and it is often unclear when the effects noted below represent shared, possibly ancestral, states and when they represent derived conditions found only in a subset of species. The study of L/R asymmetry in Lophotrochozoa therefore represents fertile ground for discerning possible answers to such questions.

Planar Cell Polarity Genes

Planar cell polarity (PCP) genes, which control the orientation of cells and their structures, have in recent years been implicated in a role in establishing L/R asymmetry. How this occurs is still unclear, but a relationship between PCP, Wnt signalling and the cytoskeleton has been postulated as the link between intracellular chirality and L/R asymmetry (Aw & Levin, 2009).

Studies in a variety of vertebrates (including zebrafish, *Xenopus*, mice and chick embryos) have revealed a role for PCP genes such as *Van Gogh* and *Rock (Rho-associated kinase) 2b* in L/R asymmetry establishment, confirmed by a loss of L/R asymmetry after these genes are knocked out or down (Zhang & Levin, 2009; Antic et al., 2010; Borovina et al., 2010; May-Simera et al., 2010; Song et al., 2010; Wang et al., 2011). Following knock-out of *Vangl*, ciliary position in the node was randomised and Nodal flow became unorganised (Song et al., 2010). This provides an explanation for the role of PCP proteins in establishing L/R asymmetry in vertebrates which depend on Nodal flow to do so, but in other organisms, such as *X. laevis*, where asymmetry is in existence before motile cilia, and in those where motile cilia have no effect, an earlier role for PCP in establishing asymmetry has been posited.

PCP genes are regulated by non-canonical Wnt signalling (Aw & Levin, 2009;

Hashimoto et al., 2010). Non canonical Wnt signalling uses Dishevelled (along with other molecules) to activate the PCP protein Rock, which is a major regulator of the cytoskeleton. Dishevelled also interacts with Rac/JNK to control actin binding and polymerisation. A gradient of Wnt signalling during oogenesis or early embryogenesis could therefore produce asymmetric localisation of active PCP proteins, with downstream effects for the cytoskeleton. This could therefore produce the signals required for organising L/R asymmetry under the cytoskeletal model, through the arrangement of cytoskeletal components or through the differential division of imprinted chromatids (Aw & Levin, 2009). Asymmetric Wnt signalling could be provided maternally, or induced by the orientation of the A/P and D/V axes, and chirality in cytoskeletal molecules could do the remainder.

Alternatively, it has also been proposed that asymmetry could be established intracellularly in a single cell, and then communicated across a cell field (for instance, the node) or even an entire embryo (Aw & Levin, 2009). In evidence of this, the authors point to the fact that actin cytoskeletons in *Xenopus* oocytes are know to have set 'east/west' chirality, set relative to the known sperm entry point, which disrupts the actin cortex of the egg. This defines the dorsal point of the D/V axis, and, the authors propose, could act as the basis for an actin-determined chirality which would then be communicated by the PCP apparatus (Aw & Levin, 2009). Even without a direct connection to its establishment, PCP genes could amplify the signal of pre-existing L/R asymmetry (for example, from the microtubule organising centre) to the wider embryo (Vandenberg & Levin, 2010).

The conclusive experiments to discern the true role of PCP genes in the establishment of L/R asymmetry are yet to be performed. This hypothesis is, however, likely to come in for increasing scrutiny in the near future.

FGF Signalling

FGF signalling, via Fgf8 and the receptor Fgfr1, regulates *Nodal* in three separate ways in the mouse. Firstly, as noted in Section 1.3.4, FGF signalling is responsible for the production of NVPs in vertebrates, with implications for the cilial theory for *Nodal* regulation (Tanaka et al., 2005). Secondly, FGF signalling's role in spec-

ifying mesodermal fates via the Ras/MAPK cascade is upstream of the action of the Notch pathway - particularly *Dll* - in inducing perinodal *Nodal* expression (Oki et al., 2010). Finally, the FGF pathway is involved in correctly regulating the production of cilia, without which Nodal flow cannot occur in those organisms which exhibit it (Neugebauer et al., 2009; Hong & Dawid, 2009). FGF signalling therefore plays an important upstream role in the production of Nodal, although is likely not asymmetrically distributed itself.

MAPK signalling

Mitogen-activated protein-kinase (MAPK) signalling pathway activation has been implicated in a range of basic roles in controlling cell division and differentiation, as well as the intra-cellular control of transcription. It is therefore perhaps unsurprising that it has been implicated in establishing L/R asymmetry, however, in vertebrates it is a link in a signalling chain, downstream of FGF via the Ras/MAPK cascade to Notch, rather than an originator of asymmetric signal (Oki et al., 2010).

Perhaps more interestingly, the MAPK signalling pathway activation has been noted in the D-quadrant of snails and polyplacophorans, which is of interest due to the crucial role this quadrant plays in specifying cell fates in the Spiralia, as noted in Section 1.7.2 (Lambert & Nagy, 2001, 2003). When the MAPK pathway is inactivated *Nodal* is not expressed and radialised embryos result. Although there is no definitive functional link yet established between these molecules in the Lophotrochozoa, Nodal signalling and the MAPK pathway are known to communicate with one another in other contexts (Clements et al., 2011), and it would be unsurprising if MAPK might be involved in activating Nodal in some fashion in this clade.

Notch Signalling

Notch signalling has been functionally shown to be involved in the establishment of asymmetry in chick (Raya et al., 2004) and zebrafish (Kawakami et al., 2005) downstream of asymmetric Ca^{2+} signalling (Raya et al., 2004). It does this via the node specific 'NDE' Notch-responsive regulatory element found 9.5 to 8.7 kb upstream of the *Nodal* gene in mice (Adachi et al., 1999; Norris & Robertson, 1999;

Krebs et al., 2003). The Notch ligand Dll1, together with the Notch transcriptional mediator protein RBP-J, bind to this site and activate the expression of *Nodal* in a variety of vertebrates (Krebs et al., 2003; Raya et al., 2003; Gourronc et al., 2007). When BAPTA, a calcium chelator, is applied to chick embryos, both asymmetric *Nodal* and *Dll1* expression is lost. At least in these species, Notch signalling seems to be the final step in a pathway of signalling downstream of a variety of other asymmetric signals, rather than an initial breaker of symmetry itself, and is the final step before the *Nodal* gene itself takes over the patterning and maintenance of asymmetry through to adulthood.

1.4 Patterning and Maintaining Asymmetry Once Established

Once initial symmetry is broken, the asymmetry that results must then be transmitted to the relevant portions of the developing embryo and then maintained throughout growth and development. Two factors, *Nodal* and *Pitx*, play a variety of roles in this process (although are notably absent, so far, from sequenced ecdysozoan species), and it is impossible to understand how L/R asymmetry is established without careful consideration as to how these genes operate *in vivo*.

1.4.1 *Nodal*

Nodal was first discovered in 1993 (Zhou et al., 1993) and is a member of the transforming growth factor (TGF) β family of cytokines (secreted proteins). Its expression has been noted in a variety of locations, including pluripotent cells, the primitive streak, the node, and in the left lateral plate mesoderm (Conlon et al., 1994). One copy of Nodal is present in mammals thus studied, as well as lophotrochozoan models, while *Xenopus* possesses five copies (*Xnr1*, *2*, *4*, *5*, and *6*) and zebrafish three (*Cyclops*, *Squint*, and *Southpaw*). Thus far *Nodal* is conspicuous by its absence in the Ecdysozoa, with speculation that the more derived modes of mesendoderm development and L/R axis specification are responsible for its absence in this superphylum (Schier, 2009). Along with performing a key role in maintain-

ing pluripotency, Nodal plays an important part in embryogenesis at gastrulation, where it induces the formation of mesendoderm. This goes on to form mesoderm and endoderm, but plays a part in the specification of the A/P and L/R axes as it does so (Beck et al., 2002; Levin et al., 1995).

Nodal shares many of its structural characters with the wider TGF β group. In vertebrates it is transcribed as an approximately 347 amino acid long precursor protein, before being processed into three parts - a signal sequence of 26 amino acids, a propeptide around 200 residues in length, and the active Nodal protein itself, of around 110 amino acids (Zhou et al., 1993). When the propeptide is cleaved extracellularly by SPC1 (Furin) and SPC4 (Pace), Nodal is freed to its active state, where it becomes a homodimer with another copy of itself, linking through disulfide bonds (Le Good et al., 2005; Schier, 2009). It has been noted, however, that on occasion Nodal precursor protein can act without being cleaved first (Ben-Haim et al., 2006).

In vertebrates, *Nodal* is expressed in two waves around the node, firstly prior to and secondly post gastrulation. The first wave is perinodal, and in performing its role in inducing mesendoderm and maintaining pluripotency, sees *Nodal* expressed along the animal-vegetal margin in the blastula, where high levels of signalling induce endoderm, and lower levels induce mesoderm (Schier, 2003; Shen, 2007). In the absence of Nodal signalling, the primitive streak itself cannot form, and gastrulation is disrupted (Conlon et al., 1994).

Once gastrulation has occurred, the second wave of expression begins and the latter role for *Nodal* as a primary regulator of L/R asymmetry comes to the fore. In vertebrates, Nodal is expressed in the left lateral plate mesoderm at this point (Nonaka et al., 1998, 2002). In invertebrates where *Nodal* is found, it is expressed on the right of the body (Duboc et al., 2005; Grande & Patel, 2009) with the exception of ascidians, amphioxus and sinistral snails, where left sided expression is observed (Yasui et al., 2000; Boorman & Shimeld, 2002; Grande & Patel, 2009). Without this signal, organs either do not develop in an asymmetric distribution, or have their normal development significantly disrupted. The careful regulation of Nodal's signalling is therefore of crucial import, both to allow mesendoderm to be specified

correctly, and to allow for proper bilateral growth of the organism (Schier, 2009).

1.4.2 The Nodal and Wider TGF β Pathway

Nodal signalling is propagated much like any TGF β ligand, and the stereotypical TGF β pathway can be seen in Fig. 1.5, with the Nodal signalling pathway evident to the left of this figure (Schier & Talbot, 2001; Shen, 2007; Schier, 2009). TGF β signalling components are known by markedly different names in deuterostomes and ecdysozoans, and in the interest of consistency vertebrate names will be used unless stated, with ecdysozoan names provided in brackets when they are particularly well characterised.

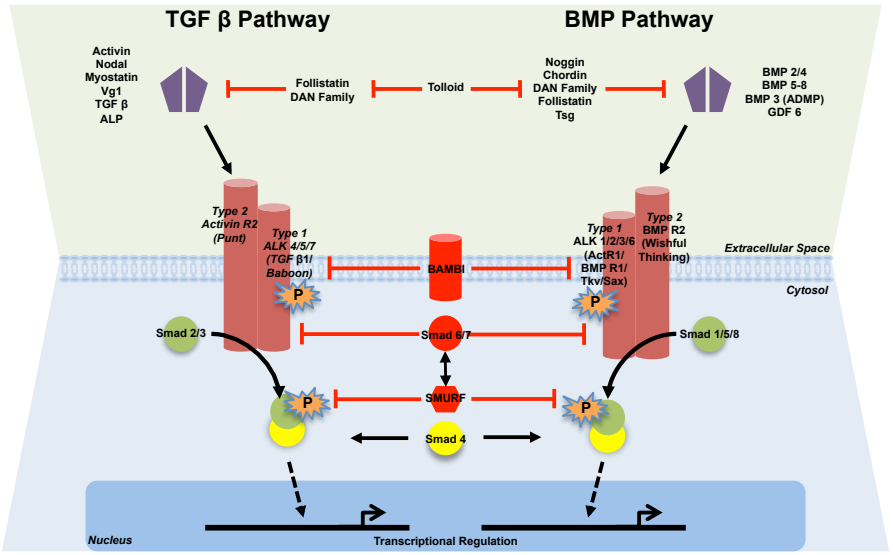


Figure 1.5: *Representation of the canonical signalling pathways for TGF β-like and BMP-like cascades with inhibitors of signalling shown in red and operational signalling shown in black. Only ligands with well-known affinity to one or other signalling pathway listed, with each pathway operating through different combinations of Type 1 and Type 2 receptors, and hence signalling through either Smad 2/3 or Smad 1/5/8 proteins intracellularly. Ligands are regulated extracellularly by a diverse range of inhibitors, which can themselves be cleaved by Tolloid to release ligands and allow signalling to occur. Intracellular regulation of signalling can occur at the receptor level, with BMP and activin membrane-bound inhibitor (BAMBI) recruited in the place of functional Type 1 receptors, or intracellularly, through Smad 6/7 inhibition of signal transduction from receptors, SMURF-mediated degradation of Smad signalling proteins, or a range of further mechanisms not shown here.*

TGF β ligands can be recognised by a characteristic conservation of six cys-

teine residues, which when folded form a structure known as the cysteine knot stabilised by three disulfide bonds (Sun & Davies, 1995). There are 33 distinct genes encoding TGF β ligands in humans, nine in *D. melanogaster*, and five in *C. elegans* (Huminiecki et al., 2009), which can be split into two broad families, the TGF-β/Activin/Nodal subfamily and the BMP (bone morphogenetic protein) families (Yamamoto & Oelgeschlager, 2004). Each of these is recognised by specific receptors, and causes the activation of a particular downstream SMAD signalling cascade.

When TGF β ligands form dimers they bind sequentially to Type 2 and then Type 1 receptors, which form a complex and are phosphorylated, recruiting and activating Smad signalling molecules. These Smads then bind to co-factors and modulate transcription (Moustakas & Heldin, 2009). TGF β ligands can be downregulated extracellularly by ligand traps, such as Chordin, Noggin and the DAN family (Balemans & Van Hul, 2002). Nodal in particular is known to be regulated by Cerl2, a DAN family member (Inacio et al., 2013). These have been catalogued in ecdysozoans and deuterostomes, and it has been observed that protostomes have less diversity in these protein families than vertebrate models (Van der Zee et al., 2008). Tolloid, a zinc metalloprotease, is capable of cleaving Chordin, hence rereleasing the trapped ligand as well as cleaving other potential repressors of TGF β signalling, such as proteoglycans (Scott et al., 1999).

TGF β ligand binding to receptor serine/threonine kinases can also be up- or down - regulated at the cell surface by membrane anchored co-receptors and receptors (Shi & Massagué, 2003). Some co-receptors, such as Cripto and the EGF-CFC family, allow active ligand-receptor complexes to be formed by acting as cofactors (Cheng et al., 2003; Shen & Schier, 2000). Downregulation can be performed by pseudoreceptors such as BMP and membrane bound inhibitor (BAMBI, also known as Nma), which compete with functional Type I receptors for ligand binding (Onichtchouk et al., 1999) or by interference with the Cripto co-receptor by Tomoregulin-1 (Harms & Chang, 2003). The existence of these regulatory mechanisms has been noted in protostomes previously (Van der Zee et al., 2008), but the degree to which these are conserved across the Bilateria is unknown.

Regulation of TGF β signalling continues intracellularly. Upon activation of the Type I receptor within the signalling complex, receptor-regulated Smads (R-Smads) are recruited from the cell membrane with the aid of proteins such as Smad anchor for receptor activation (SARA) (Itoh & ten Dijke, 2007). R-Smads are then phosphorylated and activated by the receptor complex (Massagué et al., 2005). R-Smads can be divided into two families, depending on the ligands to which they respond. Smad 1/5/8 (Mad) responds to BMP signalling, and Smad 2/3 (Smox) to TGF β, Activin, and Nodal signalling. Once activated, R-Smads bind to a co-Smad (Smad 4 (Medea)) to form a complex which mediates transcription in the nucleus, resulting in up- and down-regulation of target genes (Ross & Hill, 2008). Inhibitory Smads (known as Smad 6/7 (Dad)) compete with R-Smads for activation by receptor complexes, thus further regulating the pathway (Ross & Hill, 2008; Heldin & Moustakas, 2012). Activated R-SMAD complexes can also bind to a range of regulatory components within the cell, especially Ski/Sno (Liu et al., 2001; Shi & Massagué, 2003; Itoh & ten Dijke, 2007; Moustakas & Heldin, 2009; Heldin & Moustakas, 2012), although the diversity and actions of these are beyond the scope of this introduction.

Further intracellular regulation of TGF β signalling also occurs. FKBP12 (Huse et al., 1999), Dad/SMAD7 recruited E3 ubiquitin ligases and SMAD ubiquitination regulatory factors (SMURFs) (Di Guglielmo et al., 2003) can up- and down-regulate signalling within the cell (Shi & Massagué, 2003; Itoh & ten Dijke, 2007). These, and other regulatory mechanisms, often also participate in other signalling cascades. The full repertoire of these is beyond the scope of this thesis, and I refer the interested reader to the detailed reviews available on this topic (e.g. Shi & Massagué (2003); Moustakas & Heldin (2009). Nodal, like all TGF β ligands, is therefore extremely tightly regulated.

Nodal is also subject to competition from other members of the TGF β ligand family for binding spots on its Activin-receptor-like kinase (ALK) receptors - ActRI (Baboon) and ActRII (Punt), and in vertebrates is specifically antagonised by 'Lefty'. Lefty interferes with *Nodal* expression by binding competitively to EGF-CFC co-receptors and to Nodal itself (Chen & Shen, 2004), tightly constraining the

region in which *Nodal* is expressed in the left lateral plate mesoderm. Without this tight control, Nodal is expressed ectopically due to auto-regulation, causing a range of downstream defects (Meno et al., 1999; Schier, 2009). Not all TGF β interactions are a hindrance to Nodal signalling, as GDF1/Nodal heterodimers are more effective than Nodal alone at activating the Nodal pathway (Tanaka et al., 2007).

1.4.3 Nodal Targets

Nodal signalling, once activated, has a range of known targets for up-regulation (Dickmeis et al., 2001; Guzman-Ayala et al., 2009), including itself, through the Smad cascade. Perhaps the most crucial Nodal target genes are the transcription factors *FoxH* and *Pitx* (or, more correctly, the paralog *Pitx2* in vertebrates). These genes, along with Nodal and the inhibitor Lefty, are the only components of the pathway yet confirmed to be present in all vertebrates, and *Pitx* is the sole target to yet be identified in this pathway in the Lophotrochozoa. While the core Nodal/Pitx pathway is conserved in the Bilateria (with the exception of the Ecdysozoa), it seems the downstream targets and effectors of these have diverged much over evolutionary time.

FoxH1

FoxH (also known as *Fast1* and *FoxH*) is a member of the Forkhead Box domain containing gene family, which has presently been identified in a range of chordate species, including *C. intestinalis* (Yoshida & Saiga, 2008). FoxH1 acts as the intranuclear modulator of Nodal signalling, as when it is turned on it binds to Smad 2/4 and forms an activin response factor (ARF) complex (Kofron et al., 2004). This complex then binds to enhancer sites and activates transcription of left-sided specific genes in these organisms (Saijoh et al., 2000; Long et al., 2003; Yamamoto et al., 2003). For instance, the Nodal locus contains a FoxH1 enhancer and therefore up regulates its own transcription. Pitx is also turned on by FoxH1 (Yoshida & Saiga, 2008). Pitx then goes on to up-regulate many of the side-specific transcriptional pathways that result in asymmetry (Hamada et al., 2002).

Pitx

Pitx (pituitary homeobox) was first isolated from mouse samples by Lamonerie et al. (1996). It is similar to the RIEG/PITX homeobox genes, containing a homeobox similar to bicoid and orthodenticle/Otx, containing a lysine at residue 50 within the homeodomain, which alters DNA binding specificity slightly in regard to other homeodomain-containing proteins (Gage et al., 1999). Pitx is remarkably conserved between species, with 90-93% similarity within the homeodomain between vertebrate and ecdysozoan examples.

In vertebrates multiple paralogues of the *Pitx* gene exist, each expressed in slightly varied, albeit overlapping, patterns and responsible for different developmental processes. Three examples of Pitx, simply called Pitx1, Pitx2 and Pitx3, are the norm in vertebrate models, with Pitx2 responsible for the downstream pathway from Nodal signalling. In ecdysozoan and lophotrochozoan models thus examined only one Pitx gene has been observed, perhaps indicating that the paralogues seen in vertebrates came about during the serial whole genome duplications that took place in that lineage. Pitx acts as a transcriptional mediator of Nodal signalling, and is responsible for specifying many of the downstream, specific aspects of asymmetric growth, from organ development to craniofacial morphogenesis (King & Brown, 1999; Boorman & Shimeld, 2002). Pitx therefore is perhaps the most distal aspect of this signalling cascade, providing the final cue that results in asymmetric growth.

Other Nodal Targets

Nodal can also have other effects. As well as a role in establishing L/R asymmetry, Nodal has been shown to play a role in maintaining stem cell pluripotency (Vallier et al., 2004) and working directly on the cytoskeleton (Schier, 2009). This work, while not directly involved in the current study, has been useful in elucidating the exact nature of Nodal regulation, and has shown that many other TGF-β ligands, such as PCP genes like *Vg1*, act through the same pathway (Tanaka et al., 2007), and thus might 'fine tune' Nodal expression and action.

1.5 Regulation of Nodal

In addition to the regulatory mechanisms imposed on *Nodal* by virtue of the TGF β signalling pathway, numerous other regulators have been observed to control *Nodal* expression. Space constraints dictate that I am unable to cover the full gamut of *Nodal* regulators here, but some of the most important, and those considered later in this thesis, include:

1.5.1 Pretranscriptional Regulation of Nodal

Cis regulatory regions play a vital role in *Nodal* regulation as binding sites for activators and repressors. In mice it is known that *Nodal* is activated by four intronic enhancers. Two groups of enhancers seem to be responsible for Nodal expression in two waves. An upstream enhancer, the Proximal Epiblast Enhancer (PEE), along with the node-specific enhancer (NDE) and some input from the asymmetric element (ASE), is responsible for node-specific expression in a non-asymmetric fashion (Norris & Robertson, 1999; Adachi et al., 1999).

Once activated, the expression of *Nodal* in the left lateral plate mesoderm is partially regulated by the left side-specific enhancer (LSE, also known as AIE) (Vincent et al., 2004; Saijoh et al., 2005). This enhancer is likely to respond to Nodal in an autoregulatory enhancing fashion (Saijoh et al., 2005). This enhancer region acts with the ASE element to promote left-specific *Nodal* transcription.

These enhancers are generally responsive to the actions of specific signalling cascades. These are detailed further below.

Notch Signalling

Notch signalling is perhaps the best categorised cell-cell communication process, having been described as early as 1917 (Morgan, 1917). The Notch pathway has been shown to be involved in myriad processes, including binary cell fate choice, cell lineage specification and stem cell maintenance. As noted earlier in this work, recently it has been shown to play a part in the specification of L/R asymmetry in vertebrates in two ways.

Notch signalling is involved for the first time in the regulation of *Nodal* when its canonical pathway mediates the expression of *Nodal* in the immediate vicinity of the node. It does this via the node specific 'NDE' Notch-responsive regulatory element found upstream of the *Nodal* gene in mice (Adachi et al., 1999; Norris & Robertson, 1999; Krebs et al., 2003), as detailed in Section 1.3.5 of this thesis.

Notch's second role is on the left of the body, once symmetry is broken. It regulates *Nodal* transcription in the left of the chick, downstream of asymmetric Ca^{2+} signalling (Raya et al., 2004). When BAPTA, a calcium chelator, is applied to chick embryos, both asymmetric *Nodal* and *Dll1* expression is lost. This may act via the same NDE element, although the experiments required to show this have yet to be performed. Asymmetric expression of *Dll1* has been looked for, but not found, in other vertebrates (Kato, 2011). Notch then plays a number of disparate roles in regulating expression of *Nodal* and its downstream genes in the left lateral plate mesoderm (reviewed in Kato (2011)). These roles differ markedly from species to species, but a conserved role for Notch in *Nodal* regulation seems likely.

FoxH1

It has been shown in mice that *Nodal* is regulated by an enhancer within its large conserved first intron, to which the transcription factor FoxH1 binds (Saijoh et al., 2000; Norris et al., 2002). This conserved enhancer sequence, dubbed the ASE (ASymmetric Element), modulates patterns of *Nodal* expression throughout early development. When the ASE is specifically removed, normal L/R asymmetry is disrupted (Norris et al., 2002). It is also regulated by a left side specific enhancer, the LSE (or AIE), which is bound by the FoxH1 containing ARF complex in a similar fashion (Vincent et al., 2004; Saijoh et al., 2005). These elements are responsible for early expression of Nodal, and later asymmetric expression, but not expression directly surrounding the node (Norris & Robertson, 1999). The ASE enhancer also seems to be linked in some way to RA signalling (Uehara et al., 2009).

It is as yet unknown how well conserved across evolutionary time these enhancer elements are, although they are at least Olfactores-wide (Yoshida & Saiga, 2008). To date, FoxH1 has generally been thought to be absent from the Protostoma.

Recent evidence, discussed in Chapter 3 of this work, seems to indicate that it may in fact be present in the *P. vulgata* genome.

Wnt signalling

One of the perinodal enhancer elements, PEE, is has been found to be responsive to Wnt signalling using loss-of-function and gain-of-function tests (Norris & Robertson, 1999; Granier et al., 2011). Of the 1.8 kb PEE, analysis has shown that 125 bp sequence is particularly well conserved, with this fragment containing two Lymphoid enhancer factor/T cell factor protein binding sites, which recruit β-catenin and co-activators (Arce et al., 2006; Granier et al., 2011). Wnt signalling therefore seems to be intricately linked with the formation of the node and the initial stages of Nodal expression, as detailed earlier in Section 1.3.4.

1.5.2 Posttranscriptional Regulation of Nodal

Nodal is regulated by a range of general TGF β signalling-related modulators, as detailed earlier in this thesis. It is, however, worth spending some time examining further aspects of this regulation that relate to Nodal in particular, as well as wider aspects of the posttranscriptional regulation of Nodal.

Ligand-Related Regulation

Nodal has the scope of its interactions with the TGF β signalling cascade limited, not just by classical inhibitors of ligands such as ligand traps, but also by interactions with other ligands themselves. Perhaps the most well-described inhibitor of Nodal signalling in this manner is the ligand Lefty. This ligand, which may well be found only in deuterostomes, constrains the actions of Nodal by competing for receptor binding. Lefty lacks some of the cysteine residues found in most ligands, and thus cannot dimerise, existing natively as a monomer (Babu & Roy, 2013). This allows it to diffuse faster than dimerised Nodal, and restrict Nodal expression markedly.

BMPs, and in particular *BMP 2/4 (Dpp)*, have also been noted as playing a role in

constraining Nodal signalling, although not via competition for receptors, as they signal via different receptor pathways. In the sea urchin, for example, BMP signals specify the left hand side of the body, antagonising Nodal, which specifies the right (Luo & Su, 2012; Molina et al., 2013). However, BMP 2/4 plays an important role in turning on Nodal originally in this species (Bessodes et al., 2012).

miRNA regulation of Nodal Activity

miRNAs have been implicated in the regulation of *Nodal* in several organisms and contexts, although primarily through the modulation of the molecules Nodal interacts with, rather than *Nodal* itself. For example, miR-15 and miR-16 have been noted in *X. laevis* as regulators of the size and location of the organiser by targeting Acvr2a, the receptor for Nodal itself (Martello et al., 2007). miR-430, in zebrafish, has been shown to target the Nodal regulator *lefty* (known as *squint*) , with protection of their mRNAs causing, respectively, enhanced or reduced Nodal signalling (Choi et al., 2007). *Lefty* has also been shown to be regulated by microRNAs in humans - miR-302 in particular (Barroso del Jesus et al., 2011). It is perhaps likely microRNAs have roles in regulating *Nodal* in other organisms, but the relative novelty of the field means that these have not yet been discerned.

A role for miRNAs in regulating asymmetries is by no means restricted to chordates. In *C. elegans* the miRNA *lsy-6* regulates neuronal left/right asymmetry by controlling differential expression of genes in two morphologically bilateral taste receptor neurons (Johnston & Hobert, 2003). *lsy-6* is expressed on the left hand side, but not the right, and knocks down *cog-1*, an Nkx-type homeobox gene, with a range of effects. While Nodal does not exist in *C. elegans* and miRNAs do not have a role in regulating this gene in this species, they could be responsible for a range of similar asymmetries in other species in this manner.

1.6 Comparative Genomics and the Next-Generation Sequencing

The last few years have witnessed a revolution in molecular biology, as research has been able to move beyond the ”big six” model organisms and into new ground, abetted by the development of new technologies for deriving sequence data rapidly and relatively cheaply (Martin & Wang, 2011). Approximately 40 metazoan genome sequences have been published to date. These sequences are, however, heavily skewed towards more traditional model organisms, as can be seen in Fig. 1.6, and as such are of little direct utility to molecular research in the Lophotrochozoa. While the recent publications of genomes from cnidarians (Putnam et al., 2007), placozoans (Srivastava et al., 2008) and lophotrochozoans (Berriman et al., 2009) complements data from more established models, the ability to sequence either the expressed RNA complement or genomic DNA of an organism of interest *de novo* has allowed even smaller labs to answer a raft of biologically interesting questions in any species they desire.

1.6.1 Next Generation Sequencing

Traditional Sanger sequencing is reliable, error-resistant and has a long history of success. Unfortunately, it is also labor intensive and expensive, and increasingly it appears relatively slow compared to more recent techniques. New technologies, collectively known as “next generation sequencing” techniques have begun to supersede Sanger sequencing for a number of purposes and have dramatically transformed our ability to generate sequence data from novel model organisms, specific individual groups or individuals within a population, and even from specific tissues within an organism (Schuster, 2008; Ansorge, 2009).

For work such as that presented in this thesis, prime criterions were cost, ‘assembleability’ and the amount of sequence data obtainable by a given method. A number of next-generation sequencing techniques are commercially available and were considered for use. Others, such as Ion Torrent, Pac Bio and Polonator are also available, but are yet to gain as much traction in the marketplace.

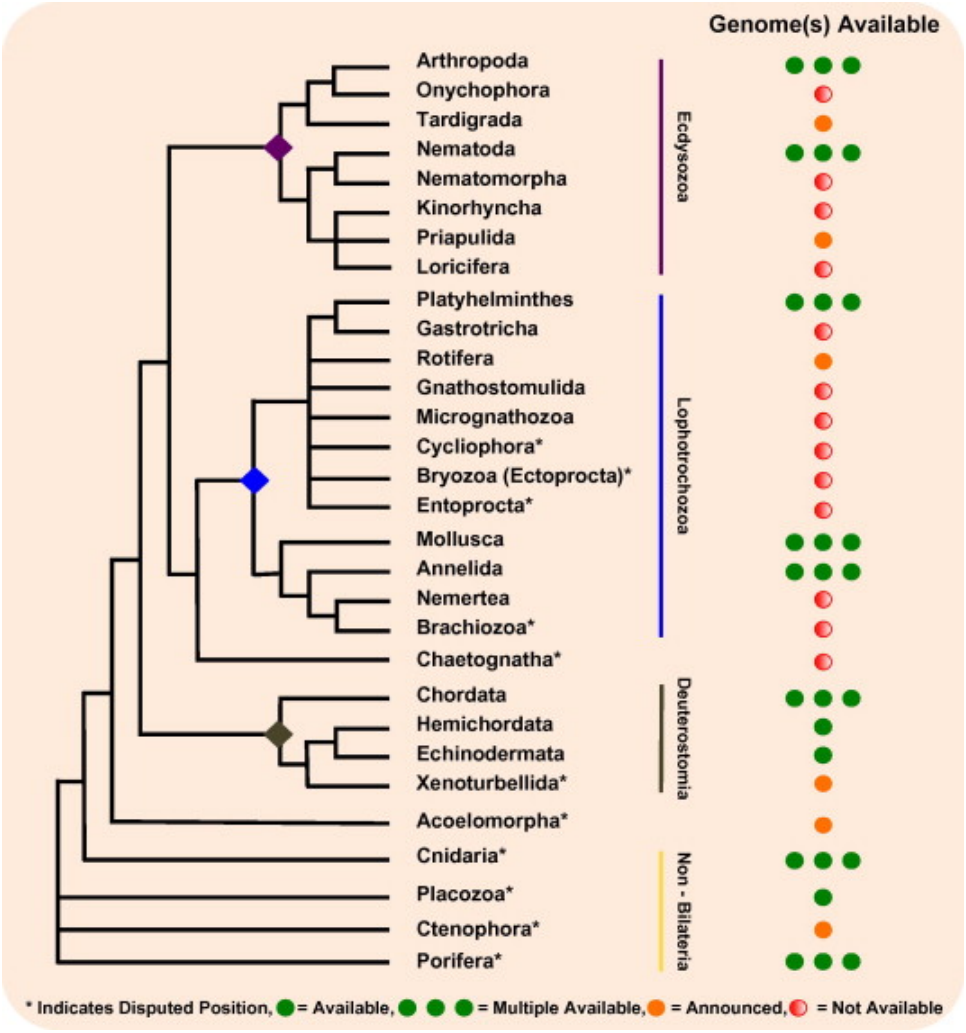


Figure 1.6: *Genome sequence availability across metazoan phylogeny. Based on Dunn et al., 2008, Hejnol et al., 2009, and Pick et al. (2010). Problematica indicated with an asterisk. At right, current public availability of genomic datasets is indicated, according to the key found at the base of the figure. Taken from Kenny et al. (2013)*

Illumina (formerly known as Solexa) sequencing was selected as the next generation platform of choice for the projects presented here. The high volume of information provided by Illumina sequencing, the utility of paired-end sequencing for assembly, and the favourable economics of the availability of Illumina sequencing through the University of Oxford all counted in its favour when making this decision. The compatibility of Illumina reads with a number of well-regarded assembly programs, including ABySS, Velvet and SOAPdenovo also aided this decision.

Illumina sequencing produces a very high number of short reads. To perform Illumina sequencing, DNA is fragmented and ligated to adaptors before being split into single strands. These are then placed in a 'flowcell', the surface of which is covered in reverse complements to adaptor sequences. DNA molecules then bind to these, creating loops (known as a 'bridges') (Shendure et al., 2005; Bentley, 2006).

As the identity of the adaptors is known, these can be used for either single end or paired end sequencing. Firstly large numbers of copies of the DNA molecule are made by standard replication, each of which re-bridges with further adaptors, attached to the surface of the flow cell. A modified version of sequencing by synthesis is then employed, with nucleotides labelled with specifically coloured fluorescent tags and temporarily modified to prevent another nucleotide joining after the first. After each addition of each nucleotide, fluorescence is measured, and the temporary 3' OH group modification is removed, allowing the next nucleotide to bind and be measured in turn.

The mechanism by which Illumina sequencing occurs is well reviewed elsewhere (Bentley, 2006; Morozova et al., 2009; Metzker, 2010). Illumina sequencing is very popular due to the large amount of data it can produce at reasonable read length (up to and beyond 100bp). The paired-end sequencing possible on this platform makes it very amenable to assembly, and a number of assembly algorithms are currently available to work with this data. In recent years a number of studies have shown the suitability of Illumina sequencing for transcriptomics (Rosenkranz et al., 2008; Hegedus et al., 2009; Wang et al., 2010b) and genome sequencing (Li et al., 2010a,b). It was hoped that the paired end reads provided by the present experiments would provide an easily assembled, information-rich basis for future

work.

1.6.2 Assembly Algorithms

In the past genomes have been assembled *de novo* using overlap graphs, aligning reads of sample sequence and gradually "walking" through different fragments to create longer, contiguous sequences, known as 'contigs' which theoretically match part of the DNA content of the original sample (Myers, 1995). A range of methods and algorithms (such as the Atlas, PHRAP, PCAP, TIGR, Phusion, and Celera assemblers) were developed for doing this with large quantities of data. These used the "overlap-layout-consensus" (OLC) approach to systematically compare reads in a pairwise fashion, essentially comparing all sequences against all others. Some used a so called "Greedy algorithm", which calculates the fragments with the largest overlap, merges them, and then begins again (both approaches reviewed in (Pevzner et al., 2001; Nagarajan & Pop, 2013)), In general this worked well, but was computationally intensive, could be affected by the order in which the contigs were assembled by the algorithm, and suffered from a number of problems, especially when dealing with long regions of repeated sequence.

The small read lengths of many next-generation sequencing technologies make traditional alignment methods difficult to perform, and the sheer amount of data produced by a single run on a next-generation platform makes alignment using traditional methods computationally demanding (Pop & Salzberg, 2008). Furthermore, base calling mistakes have a far greater impact in alignments of the results of *de novo* sequencing, due to the shorter read length (Whiteford et al., 2005; Paszkiewicz & Studholme, 2010). To cut down on the number of pairwise comparisons that an assembler needs to make, and to ameliorate some inherent problems, a number of algorithms have been proposed. These separate the vast library of data into a specific data structure to simplify downstream assembly, and tend to be either 'string-based' or 'graph-based'.

Both of these models share some techniques in common - both string based and graph based algorithms, for instance, commonly use paired-end sequence information to simplify assembly and reduce gaps in contigs. Some use a data struc-

ture known as a 'prefix tree', to create smaller subsets from which to build (also known as 'hash searching'). Some iteratively construct assemblies, allowing misassemblies on the first pass but returning to correct for mistakes on subsequent passes (Chevreux et al., 1999).

String based algorithms often use the Greedy extension algorithm referenced earlier, and include the likes of SHARCGS, QSRA and SSAKE (Zhang et al., 2011). In these, reads are selected to form the basis ('seeds') of contig formation. A specific length and read quality cut off is then used to determine which other reads the program will allow to align and 'build' with the growing contig assembly (Pop & Salzberg, 2008). The use of the Greedy algorithm by these assemblers is, however, sub-optimal for assembling reads using long mate pairs, as information from repetitive regions is often lost and subsumed into the assembly with the largest overlap (Nagarajan & Pop, 2013).

Graph based models, after Pevzner et al. (2001), use the concept of a 'sequence graph' or *de Bruijn* graph, which separates reads into nucleotide reads of a certain length, known as *k*-mers, and then connects 'nodes' between *k*-mers. By checking only these nodes for potentially continuing overlap, the computational power required decreased markedly, and these graph based methods form the lion's share of modern next-generation sequencing assembly techniques (Nagarajan & Pop, 2013).

1.6.3 Assembly Software Selection

At the time of writing, approximately 35 *de novo* assembly programs were generally available (Paszkiewicz & Studholme, 2010; Zhang et al., 2011), with no clear consensus yet established as to which is the best performing (Surget-Groba & Montoya-Burgos, 2010; Kumar & Blaxter, 2010). Some assemblers, such as NextGENe and Newbler, are only available under licence, while others remain entirely open source. Some, particularly string-based algorithms, must use particular read lengths, while others can deal with a variety of inputs.

Several attempts have been made to differentiate the 'best' assembly algorithm (Zhang et al., 2011; Miller et al., 2010; Kumar & Blaxter, 2010). As Illumina datasets were used in the present project, the recommendations of investigation by

Zhang et al. (2011) were particularly pertinent, as that paper used simulated Solexa datasets as the basis for its investigation.

Zhang et al. (2011) came up with clear recommendations about the approaches that should be utilised, but not necessarily which programs. For eukaryote-level genomes, or when large amounts of data (≥100,000,000 reads) are involved, extra computational power is a necessity for good assembly, and only *de Bruijn* based methods should be attempted. It is generally advised by Zhang et al. that assembly is performed in the first instance with a number of programs. The choice of final program to optimize for best assembly can then proceed according to personal preference, speed of assembly observed and results gained, although the community surrounding some assemblers (especially Velvet) is now reaching a critical mass, and may push popularity in certain ways in the future (Paszkiewicz & Studholme, 2010; Zhang et al., 2011).

Several *de Bruijn* based methods were therefore trialled in the course of the current project, including ABySS (Birol et al., 2009; Simpson et al., 2009), Trinity (Grabherr et al., 2011) and Velvet/Oases (Zerbino & Birney, 2008; Zerbino et al., 2009; Zerbino, 2010; Schulz et al., 2012). The additive multiple-*k*-mer method of transcriptome assembly was also trialled. This involves sequentially performing assemblies using a range of *k*-mer lengths. It was first described by Surget-Groba and Montoya-Burgos (2010), and removal of redundancy has been used in a number of studies to increase data output, in as widely divergent organisms as horned beetles and bracken fern, as well as others (Choi et al., 2010; Der et al., 2011).

Different *k*-mer lengths tend to be better at optimising for different contig outputs. High *k*-mer length assemblies produce long contigs, but can bias assembly towards highly -expressed genes. Lower *k*-mer lengths will assemble more contigs, including those of lower-expressed (and often more interesting) genes, but tend to be more inaccurate and thus result in assembly going down 'dead ends', resulting in shorter contigs as a whole. By running assemblies at a range of *k*-mer sizes, then removing redundancy by keeping only the best copy of a given contig assembly, the best of both worlds can, in theory, be obtained.

1.7 The Lophotrochozoa

Recent evidence has shown a monophyletic relationship between such apparently diverse organisms as annelids, molluscs and bryozoans (Aguinaldo et al., 1997; Dunn et al., 2008; Halanych, 2004), resulting in the establishment of a new taxonomic grouping - the lophotrochozoan superphyla, on a par with the Ecdysozoa and Deuterostoma, whose taxonomic status is more long standing. There still exists some controversy about the specific position of some taxa, but these three clades are linked by analyses of different molecular datasets such as ribosomal RNAs (Field et al., 1988), Hox gene relationships (de Rosa et al., 1999), and mitochondrial genomes (Stechmann & Schlegel, 1999), and are now generally accepted by the literature and the scientific community.

Lophotrochozoans comprise 14 phyla and are a diverse morphological grouping, as can be seen in the sketches on Fig. 1.7, but are all linked by common characters and processes in development. The Lophotrochozoa contain many organisms that undergo true spiral cleavage, described in more detail in Section 1.7.1, which probably reflects the ancestral form of cleavage within the clade, and are therefore often referred to as the "Spiralia" (Dunn et al., 2008). When going from the 4 to the 8 cell stage, the micromeres that form the animal pole 'twist' and settle in the grooves between the larger macromere cells in a spiral motion (Hejnol, 2010). Each of these cells has a specified 'fate', that is, their descendants will form the same cells in the same area of the adult in every embryo (Shankland & Seaver, 2000), and their division is thus referred to as 'determinate' (Pearson, 2003).

The Lophotrochozoa are primarily made up of two, non-monophyletic groupings, the trochozoans and the lophophorates, whose names refer to the lophophores (feeding structures) and trochophore larvae found in many of the phyla that make up the clade (Dunn et al., 2008). The lophophorates and trochozoans were previously separated on morphological grounds, but have been grouped together on the basis of modern molecular phylogenetic studies, which have shown them to be closely interrelated (Giribet, 2008), although the exact nature of these interrelationships is still subject to some debate.

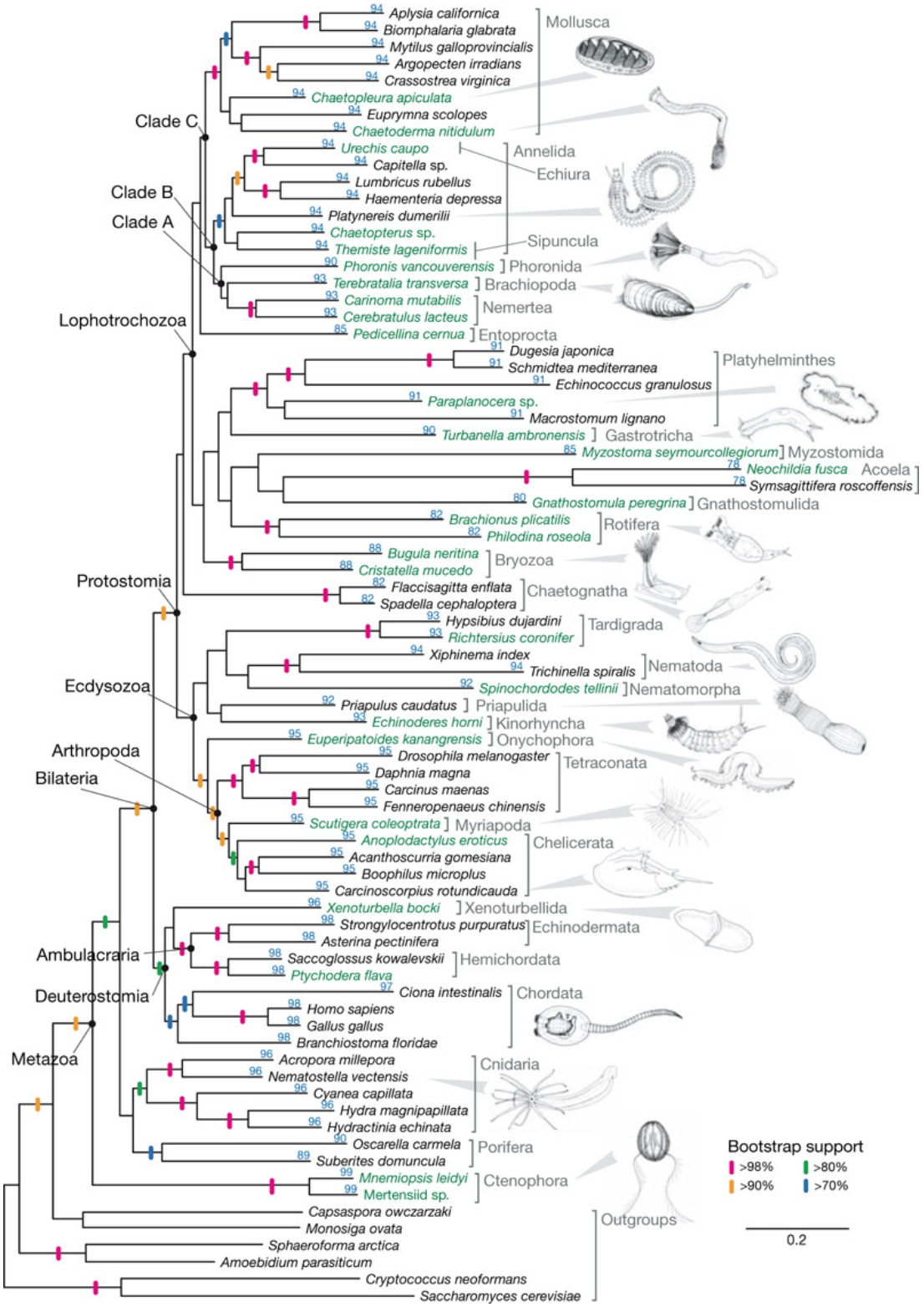


Figure 1.7: *Metazoan phylogenetic relationships, as determined by analysis of 140 genes. Support values are derived from 1000 bootstrap replicates and are shown in blue above each branch. Figure reproduced from Dunn, et al. (2008)*

1.7.1 The Spiralia

The Lophotrochozoa are sometimes referred to as the Spiralia, a reference to the particular pattern by which their cells ancestrally are said to have divided in early development. More uncontroversially, this term is applied to a number of phyla (including molluscs, annelids, platyhelminths and nemerteans) who exhibit canonical spiralian cleavage. The deterministic and constrained nature of spiralian cleavage must be understood before drawing conclusions as to the establishment of L/R asymmetries in these organisms. As the cleavage of embryos is already 'preset', the identity of the lineages which will make up left and right hand sides of the body is often predictable in these species, however, the mechanisms which act to differentiate the two sides of the body are yet to be fully understood.

Spiralian development follows highly conserved lineages of cell division, with a characteristic early cleavage pattern, meaning that the origin of organs can be traced to individual cells. It should be noted that some spiralians can differ from the following, sometimes markedly, but sufficient evidence exists for us to reliably infer the ancestral mode of spiralian development (Hejnol, 2010). Of the Lophotrochozoa, there are a number of phyla, including the molluscs, annelids, platyhelminths and nemerteans, who could be described as possessing true spiralian cleavage (a diagram of which can be seen in Fig. 1.8), while others, such as rotifer, bryozoans and brachiopods, possess a modified version. The Spiralia, while similar in their first divisions, possess a wide range of adult forms. This diversity, when compared to such shared beginnings, make them ideal models for studying how development programs can evolve to adapt and take advantage of novel niches.

At first it was thought that the fate of all cells in spirally cleaving organisms was entirely fixed, rather than regulated as found in those lineages which undergo 'indeterminate cleavage' (Wilson, 1904a,b; Costello, 1945). This phenomenon was known as 'mosaic development', and is widespread among protostomes, but it is particularly prominent in molluscs and annelids (Wilson, 1893). While the specified lineages of many protostome species definitely exist, the original hypothesis of

entirely fixed developmental lineages within the Spiralia has since been proved incorrect, and it seems more likely that the conserved pattern of spirally cleaving cell fates is the result of a conserved framework of inductions (Clement, 1962; van den Biggelaar & Guerrier, 1979; Arnolds et al., 1983; Martindale, 1985, 1986; Boring, 1989; Freeman & Lundelius, 1992; Dictus & Damen, 1997).

Some examples of spiralians developing in a regulative fashion have also been noted (Nielsen, 2010, p. 24-25), although these are poorly understood. The conserved and highly structured nature of spiralian development is therefore likely the result of constraints on the larvae, which must develop quickly to the trochophore stage to be able to feed and respond to the environment. Any changes to the pattern of cleavage are likely to be deleterious in the first instance, rendering the spiralian cleavage pattern resistant to small scale changes, resulting in the degree of conservation we see around us today.

1.7.2 Spiralian Cleavage

> *Spiral cleavage, as all students of zoology discover at some stage in their early career, is not only difficult to see, but even more difficult to grasp and, most difficult of all, to describe.* (Anderson, 1973, Chapter 2, p. 7)

In species outside the Amniota, the site of sperm entry, along with a raft of maternally provided factors placed at specific points in the egg, has already begun to establish a degree of orientation in the embryo (Gerhart et al., 1981; Nusslein-Volhard, 1991; Grunert & Johnston, 1996; Yost, 1998) by the time gastrulation takes place. This provides the first information on the future position of the axes of the body, inferred from the position of the vegetal and animal poles of the egg, by the position and effects of maternally provided mRNA. In contrast, in amniote species, the sperm entry site can help distinguish the parts of the embryo which will go on to form the inner cell mass (and thus the embryo proper), and which parts will form the trophoblast. It does not, however, play a role in the establishment of any axes of the body - and the mechanism which does is still unknown (Mikawa et al., 2004).

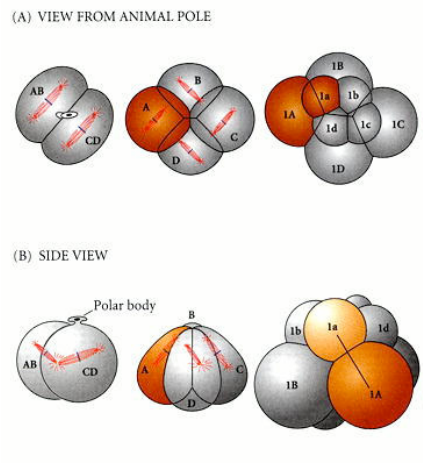


Figure 1.8: *Stereotypical cleavages of a spiralian embryo, in this case the mollusc Trochus, viewed from the animal pole and from the side, showing the angle at which cleavages occur in order to achieve the canonical spiralian pattern.Taken from Gilbert (2000)*

Classical spiralian cleavage at first follows the holoblastic cleavage pattern, with the sperm entry point determining the plane of the first division. This cleavage can be equal or unequal in the volume of daughter cells- if the latter, it is already possible to determine the future posterior end, as it will derive from the larger of the two cells. A 'polar lobe' can also be formed prior to the first cleavage in some species, particularly among the Annelida and Mollusca. This is a portion of cytoplasm which is almost entirely 'budded' off from the side of the embryo, only to be re-absorbed into one of the daughter cells after the first cleavage, and this is described further in Section 1.7.3. As with unequal cleavage, this results in one cell being much larger than the other. This is a clear indication that it will give rise to the lineage at the posterior of the embryo, which will also result in the production of the important '3D' cell, as detailed below - the larger cell is therefore referred to as the CD cell.

The second cleavage then follows at right angles to the first, producing four cells (A, B, C and D), which together are referred to as 'macromeres'. Unequal cleavage or the production of a polar body recurs at this juncture in some species, with one of the four resulting macromeres significantly larger than the other three as a result, and designated the 'D' macromere. It is postulated that the actions of a polar lobe could significantly influence cytoplasmic localisation of maternal determinants in the 'D' lineage, but the details of how this could take place are as yet unknown. Respectively, the macromeres A to D give rise to the left, ventral, right, and dorsal parts of the organism.

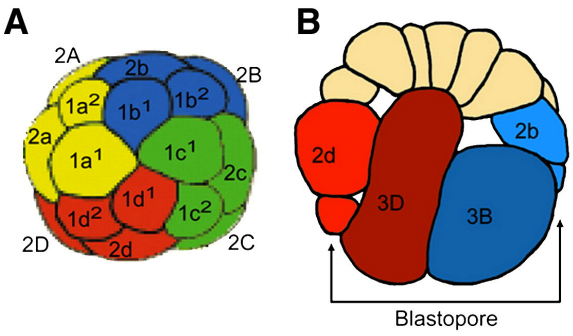


Figure 1.9: *Spiral cleavage and fate map in equally cleaving gastropods based on and adapted from the Patella fate map by Dictus & Damen (1997). A shows the view from the animal pole of a 16 cell embryo, showing the characteristic spiral pattern and nomenclature used to refer to cells. B, shows a lateral view of the process of gastrulation, with the 3D macromere becoming induced to form mesoderm by contact at the blastocoel roof. Figure reproduced from Koop, et al. (2007)*

In the third through to sixth divisions which characterize spiralian cleavage and give rise to its name, cleavage planes are not parallel with or at right angles to the animal/vegetal axis, but rather cleavage occurs at an angle to the animal/vegetal axis. Four micromeres, so named as they are smaller than the macromeres which gave rise to them, result from each division and are referred to as a 'quartet'. The quartet that results from the synchronous third division form a spiral arrangement of blastomeres at the animal pole of the embryo (see Fig. 1.8), but at a slight angle to the true animal/vegetal axis. As a result they nestle into the grooves formed by the four macromeres.

At the end of the third cleavage, an embryo can be divided into four quadrants, initially each made up of two cells, a micromere at the animal pole and a macromere at the vegetal. These divide and develop into stereotypical forms, called the A, B, C and D quadrants, and go on to respectively form the left, ventral, right and dorsal regions (Verdonk & van den Biggelaar, 1983). A general outline of these can be seen in Fig. 1.9. The specification of the D quadrant is a vital point in the development of the spiralian embryo, as it is key to the control of the establishment of the D/V axis and further embryonic development. Sometimes the future D quadrant is already known, as a result of unequal cleavage or the presence of a polar lobe. In many cases, however, the D quadrant is induced as a result of the next two divisions of the embryo.

Firstly, two more synchronous divisions of the macromeres occur, producing two more 'quartets'. The three quartets produced from these synchronous divisions

form ectoderm and some muscle tissue. It is the 6th, non-synchronous division of the macromeres which is, however, perhaps the most interesting.

As gastrulation occurs, a cell is forced to move within the blastocoel of the embryo, generally the largest cell present, which may contain some maternal determinants. It is dubbed the 3D macromere and is induced to divide and form the mesoderm as soon as it reaches the top layer of micromeres (van den Biggelaar & Guerrier, 1979; Arnolds et al., 1983; Boring, 1989), a schematic of which can be seen in Fig. 1.9. This cell division, which is also in many species the first to occur in the 6th round of division, results in the production of the 4D macromere and the 4d cell, which is also known as the mesentoblast or organiser. This cell produces the mesoderm in spiralians, producing bilateral symmetry within the embryo by dividing into two cells, (4d1 and 4d2, also known as M1 and M2, or Ml and Mr) (van den Biggelaar, 1977).

The process of gastrulation thus occurs far earlier in spirally cleaving embryos than in radially cleaving embryos, after far fewer rounds of cell division. This may be due to differing demands in embryogenesis - as some spiralian larvae (trochophores - see Section 1.7.4) are free living from an early point, speedy development may have provided a comparative evolutionary advantage.

1.7.3 Polar Lobes and Cytoplasmic Localisation

As mentioned above, some species within the Spiralia extrude a portion of their cytoplasm, known as a 'polar lobe', before cleaving for the first and second time. An example of this, in the mollusc *Dentalium*, can be seen in Fig. 1.10, and the cell to which this structure is attached goes on to form the D quadrant of the embryo. Up to one third of the total cell volume can be budded into this structure, which is reabsorbed after cleavage has taken place. This process allows a single cell to retain the lion's share of the cytoplasm, and, presumably, maternal determinants within it. Furthermore, when Wilson (1904a) removed the polar lobe at the 2 cell stage, prior to the second division, the abnormal embryos which resulted lacked mesoderm-derived structures. Some species do not possess a polar lobe but instead show 'unequal cleavage', that is, the cell opposite the sperm entry site at the 4 cell

stage possesses a larger amount of cytoplasm and goes on to form the D quadrant, with equivalent results to those species with polar lobes (Guerrier, 1970).

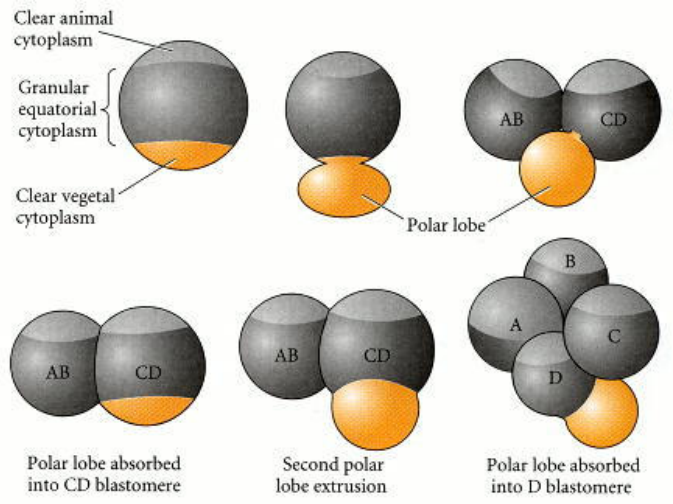


Figure 1.10: *The production and reabsorption of polar lobes in the Mollusc* Dentalium *before the first and second rounds of cleavage, and the specification of the D quadrant that results. Taken from Gilbert (2000)*

Even in species where unequal cleavage or polar lobes are not found, a significant degree of subcellular localisation of maternal determinants takes place, resulting in differential segregation of these into particular cells - generally the presumptive D quadrant, although it is difficult to discern this without the presence of a polar lobe. In the gastropod *Ilyanassa obsoleta*, for example, around 4 % of RNA transcripts are localised to interphase centrosomes, and thence to a specific daughter cell (Kingsley et al., 2007), and in *Crepidula fornicata* the same process has been investigated by RNAseq (Henry et al., 2010b).

The potential influence of maternally provided factors on the development of the Spiralia can therefore not be overstated. Many of the examples considered in this introduction come from more canonical model organisms, which do not share spiralian cleavage or the same maternal determinants as the lophotrochozoan organisms

1.7.4 Trochophores and Later Spiralian Development

Trochophore larvae are found throughout the Spiralia. While early embryos within the Spiralia can be so similar as to be indistiguishable through early cleavage stages,

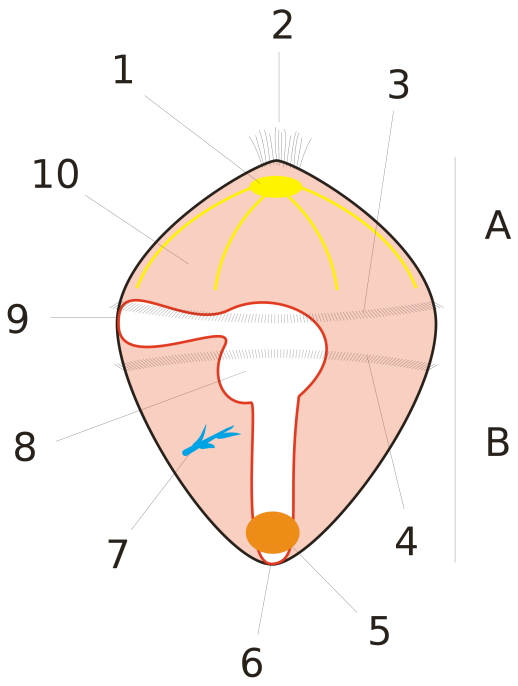


Figure 1.11:
General Trochophore Anatomy:
A - episphere, B - hyposphere
1 - ganglia
2 - apical tuft
3 - prototroch
4 - metatroch
5 - nephridium
6 - anus
7 - protonephridia
8 - gastrointestinal tract
9 - buccal opening
10 - blastocoele
Figure reproduced from Xvazquez, Creative Commons Licence

later larval development can differ widely from phyla to phyla. These larvae are described here as they represent the stage at which L/R asymmetry is most readily first observed in the Lophotrochozoa, and as such they represent the stage at which we assess the majority of our experimentation. The typical trochophore anatomy can be seen in Fig. 1.11.

After the trochophore stage *sensu stricto* it becomes more difficult to directly compare characters across the Lophotrochozoa. Veliger larvae, for instance, which are found in molluscs, are almost certainly a more derived character possessing markedly different features, including an enlargement of ectodermal tissue to form the shell gland, the presence of a phenomenon known as 'torsion', and a 'velum', formed from the prototroch, which aids in locomotion and feeding. Upon the attainment of "competence" - the state at which trochophores settle and move to a more sedentary juvenile stage - all phyla begin to diverge markedly. The trochophore stage is therefore the most suitable stage to compare characters between lophotrochozoan phyla - characters such as L/R asymmetry are already evident, but the latter, more derived elements of development are yet to occur, thus the potential for obfuscation of shared characters is kept to a minimum.

		Genotype	Phenotype
DD ♀ × *dd* ♂	→	*Dd*	All right-coiling
DD ♂ × *dd* ♀	→	*Dd*	All left-coiling
Dd × *Dd*	→	*1DD:2Dd:1dd*	All right-coiling

Figure 1.12: *The genetics of chirality in Limnaea peregra, with parental genotypes at far left and offspring genotypes and phenotypes centrally and at right respectively* after Sturtevant (1923) and Boycott et al. (1930)

1.7.5 Asymmetries in the Spiralia

While asymmetries can be difficult to recognise in many taxa without dissection and internal examination, it has long been noted that members of the Spiralia provide excellent models for research in asymmetry because, unlike other organisms, their asymmetries are clear at first sight. Many snail species, for example, have shells that are either dextrally or sinistrally coiled. Spiralians have already proved indispensable in determining the conservation of D/V axis establishment (Grande, 2010, Table 1), and provided insights into the genetics (Oliverio et al., 2010) and establishment (Grande & Patel, 2009) of L/R asymmetry. Further research in spiralians is bound to reveal many aspects of the true conservation of this pathway across the Bilateria.

The chirality of gastropods is obvious to the naked eye, visualisable as a characteristic coiling of the shell to the left or right, reflecting the directionality of the 3rd division of spiral cleavage. The control of chirality determination attracted attention early in the history of genetics. Sturtevant (1923) and Boycott et al. (1930) determined that in *Limnaea (Lymnaea) peregra*, which can exist in dextral and sinistral forms, chirality is determined by the genotype of the mother, which can differ in phenotype from offspring. This can be seen in Fig. 1.12. The dominant, dextrally spiralling allele *D* and the sinistrally coiling recessive allele *d* are shown, and it is clear that phenotype is maternally inherited. Recently, the locus responsible for alternately chiral forms has been mapped in *Lymnaea stagnalis*, with a region of around 0.4 Mb in size pinpointed for further investigation in this regard (Liu et al., 2013a).

The maternal inheritance of chirality determining factors was further tested by

Freeman & Lundelius (1982), who took cytoplasm from the unfertilised eggs of dextrally coiling snails and injected it into those of sinistrally coiling individuals, recapitulating dextral coiling as a result. The same process was tested in reverse, but sinistral cytoplasm had no effect on dextral eggs. This suggests that some maternal factor controls the dextral orientation of spiral cleavage, perhaps through the orientation of mitotic spindle fibres, in the majority of spiralians. Dextral cleavage seems likely to be the ancestral form within the Mollusca, and most probably through the Spiralia as a whole (Ponder & Lindberg, 1997; Grande & Patel, 2010).

Alongside the most obvious shell-related L/R asymmetries, there are a number of other asymmetries that can be observed in molluscs, from renal organ growth (larger on the right) to gonad position and digestive, muscular and anatomical asymmetry. The pattern of growth of these structures is affected by "torsion", a mollusc specific movement in embryogenesis, where the normal body plan is twisted to one side, possibly as an adaptation to allow the retraction of the body into the shell (Brusca & Brusca, 2002). This can make comparison of adult structures more difficult to accomplish.

Other members of the Spiralia also possess L/R asymmetries, although these tend to be much less noticeable to the naked eye. In annelids, for example, jaw shape (Paxton, 2009), gut morphology (Boyle & Seaver, 2008) and operculum side of origin (Schochet, 1973) all have been shown to possess set L/R asymmetries. These are much less well catalogued than molluscan L/R asymmetry, due to the latter's long standing status as a model for research. Where Molluscs can possess both dextral and sinistral forms, it seems that the Annelida consistently possess only a single form.

A variety of other lophotrochozoan phyla have noticeable and consistent L/R asymmetries, including Platyhelminthes, Nemerteans and Rotifera (Rotifera data Kenny, unpublished data, noted later in this report, all others reviewed in Grande (2010)). Where these have not been catalogued, it is as likely to be the result of insufficient investigation as a lack of L/R asymmetry, as many phyla within the Spiralia remain under researched and under represented in the literature.

1.8 Phyla and Species Studied

1.8.1 Molluscs

The Molluscs are perhaps the most morphologically diverse extant phyla, and are second only to the Arthropoda in terms of categorised species. The total number of living mollusc species is believed to be in excess of 100,000, with the fossil record containing around 25,000 additional catalogued examples.

Molluscs can be found in a remarkable range of habitats, both terrestrial and aquatic, and due to their hard shells, their fossil record is well described, dating their divergence back as far as the Cambrian period (542 to 488 million years ago). Due to their diversity in both morphology and lifestyle, it has proven difficult to reliably reconstruct the true mollusc phylogeny. Molluscs are common subjects for research, both due to their positive economic benefits (food and pearls), and negative aspects (as pests, often introduced, vectors for disease and biofoulers). Despite this, there is still much to learn about this diverse and adaptable phyla.

1.8.2 Molluscan Phylogeny and the Patellogastropoda

Presently extant mollusc species are conventionally divided into eight lineages. The Gastropoda are the largest class of molluscs, and are found in both terrestrial and aquatic environments. Understanding their true phylogeny is important for inferring the features of the common ancestor of all molluscs, in order to be able to draw inferences for the gain and loss of traits across the Spiralia.

The last years have seen a spate of molluscan phylogeny work, for example Vinther et al. (2011); Kocot et al. (2011); Smith et al. (2011). The most robust of these approaches, that of Smith et al. (2011), is shown in Fig 1.13.

True Molluscan phylogenic relationships, while trending towards consensus, are not yet settled. This renders the identification of ancestral morphological features in the mollusc lineage contentious (Henry et al., 2004). While *Patella vulgata* was not used in the recent analyses by Kocot et al. (2011) and Smith et al. (2011), the limpet *Lottia gigantea* was used, and was found to lie basal to all other Gastropoda in both studies, with excellent bootstrap support (see Fig: 1.13). This,

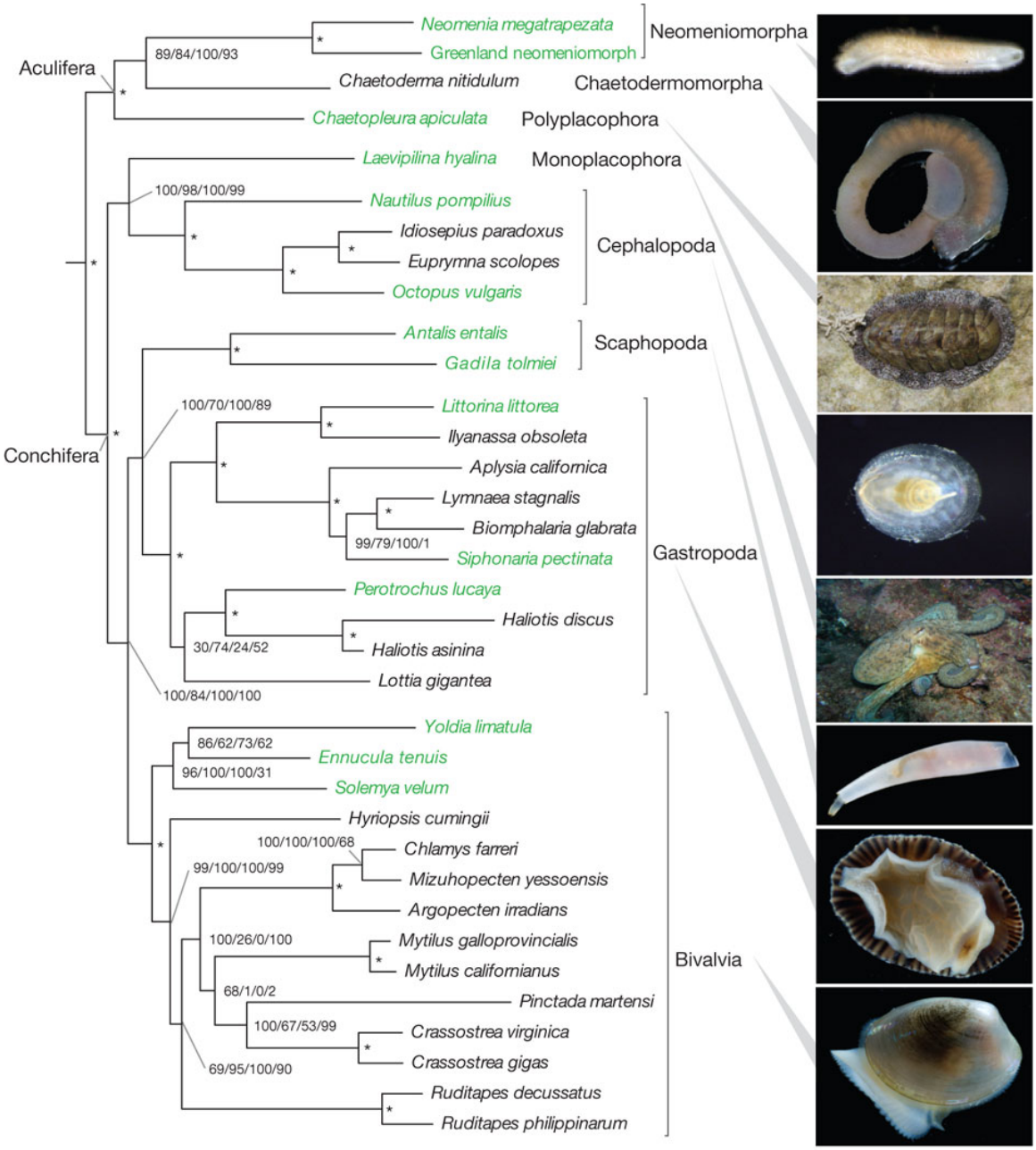


Figure 1.13: *Figure representing intra-Mollusc phylogenetic relationships, as ascertained using a variety of tools. At the base of nodes, asterisks represent 100% support under all tests, while numbers, in the order A/B/C/D represent as follows:. A) RAxML analysis bootstrap support (WAG model) using a large number of genes. B) RAxML analysis bootstrap support (WAG model) using a small number of genes. C) represents MrBayes derived posterior probability (WAG model) using a small number of genes. D) PhyloBayes derived posterior probability (CAT model) using a small number of genes. (Smith et al., 2011)*

when considered alongside the findings of Nakano & Sasaki (2011), who found the Patellidae to lie among the most basal families of limpets, confirms earlier suspicions (Haszprunar, 1988; Ponder & Lindberg, 1997) that limpets represent excellent models for the least derived form of gastropod mollusc.

P. vulgata therefore is a good candidate for determination of ancestral states, at least when considering the Gastropoda, and broader inference can be obtained by comparison with a wider range of model species among the Spiralia. The short branch length of *Lottia* in these studies, while not conclusive when considering suitability, also support the use of limpets as a basal, less derived model for gastropod development.

The Gastropoda

The Gastropoda are by far the largest class of molluscs, and the only one to have successfully adapted to terrestrial environments. Gastropods undergo a process known as "torsion", where the viceral mass twists 180 degrees while at the veliger stage of development, resulting in an anteriorly located mantle cavity, a uniquely shaped nervous system and a gut that exits above the head. This torsion is driven by asymmetric cell proliferation (Kurita & Wada, 2011).

Gastropods generally exhibit a large degree of cephalization, with a large number of sensory structures compared to other mollusc classes. They also possess a range of specialised radulae depending on diet, which can vary markedly. Gastropods are one of the most diverse classes of organisms on the planet, surpassed only by certain insect classes, and to draw wider conclusions as to the ancestral modes of Molluscan development, and those shared with the spiralia and metazoans as a whole, it is crucial to select a representative model. The mollusc phylum, for all its modern diversity, originated as a group of marine organisms, and as noted in Section 1.8.2 above, it is perhaps the Patellogastropod limpets that represent the least derived aspects of gastropod development and morphology. With that in mind, *P. vulgata* was selected as our model organism within the Mollusca, in the hope that it would provide a firm window into the ancestral molluscan form.

Figure 1.14: *Several* P. vulgata*, attached to a rock on the seashore. Image by author.*

1.8.3 *Patella vulgata*

Patella vulgata, the Common European Limpet, is a typical true limpet of the Patellidae family, a univalve gastropod. It is found throughout Europe, as far north as the Arctic circle and as far south as Portugal. It is found attached to firm substrates from the high shore to the edge of the sublittoral zone, although it predominates in areas of wave action. Its shell is conical, up to around 6 cm long, and lacks defined chirality. *P. vulgata* is herbivorous, feeding on algae using a docoglossan radula.

The Patellogastropod order of limpets, to which *P. vulgata* belongs, can be found worldwide, and is well described and widely used as models in studies of ecology, development and evolution (Lindberg, 2008; Nakano & Sasaki, 2011). They are thought to be the least derived members of the Gastropoda, and as such make excellent models for determining ancestral characteristics (Golikov & Starobogatov, 1975; Haszprunar, 1988; Ponder & Lindberg, 1997; Nakano & Sasaki, 2011). Of the Patellogastropod order, the Patellidae family are in a basal position, although the exact nature of this is still under a small amount of debate (Nakano & Sasaki, 2011).

Due to their distribution near historical centres of study, and their phylogenetic position basal to the rest of the order, the Patellidae are perhaps the best understood family of the order. A morphological investigation into the Patellidae by Ridgway et al. (1998) divided the 38 species within this family into four genera, with *P. vulgata* placed as one of 9 species within the genus Patella, where it remains at present.

P. vulgata is believed to be protandrous (Orton et al., 1956), although exceptions to this general rule have been observed. *P. vulgata* reach sexual maturity as males

at around 9 months of age. Growth rates depend on environmental conditions, but the transition to female can occur at any time over one year of age. They are sexually dormant through the early summer and gravid throughout winter months, with the breeding season coinciding with the onset of winter - in more northern positions this can occur as early as August. Eggs (around 160 μm around) are broadcast singly, perhaps stimulated by on-shore wave action and rough conditions (Bowman & Lewis, 1977). It has been suggested that this is to ensure that the shore is damp and larvae are not blown out to sea (Branch, 1981). This spawning is synchronised to an extent, with Ballantine (1965) noting 80 % of reproduction within 2 days. The eggs thus released are then fertilised externally.

Once fertilised, *P. vulgata* develop as free swimming trochophore and veliger larvae, living as pelagic plankton for around 2 weeks. They then settle, generally at around 0.2 mm long. In the lab it has proven difficult to keep Patellidae larvae beyond 10 days post fertilization (Wanninger et al., 1999, 2000), by which point they are approaching competence and preparing to settle. This is a far shorter time than many trochophore larvae, indicating perhaps that *P. vulgata* larvae do not disperse particularly widely compared to other species.

P. vulgata is useful for embryological work due to its typical spiralian development, especially as it lacks the polar lobe found to influence specification of macromeres in other spiralians. The development of *P. vulgata* has been excellently described by Smith (1935), although modern staging of embryonic development has not been performed. In its stead, however, several in depth analyses of cell lineages in this species have been undertaken (Damen & Dictus, 1994, 1996; Dictus & Damen, 1997) and we have some knowledge of the musculature of the larvae (Wanninger et al., 1999).

The presence of a sequenced genome for *Lottia gigantea*, another member of the Patellogastropoda (Simakov et al., 2013), provided an initial resource for degenerate PCR of target genes in this species, before being supplanted by a transcriptome (Werner et al., 2012). *P. vulgata* therefore represented an experimentally tractable model with a history of use in embryological research, a firm framework in which to begin investigations into L/R asymmetry in the Lophotrochozoa. Bivalve and

gastropod molluscs are likely to be well-sampled in the near future, due to their economic importance in the food and pearl trade. The production of a genomic resource in this species was therefore aimed more at the derivation of a resource for comparison of non-coding regions, to allow inference of the degree of conservation in the mechanisms for controlling L/R asymmetry-related gene expression.

1.8.4 *Biomphalaria glabrata* and *Crepidula fornicata*

While not used as embryological models in the present study, these species of gastropod are used for developmental studies in other laboratories and have well described cell lineages (Camey & Verdonk, 1969; Henry et al., 2010a). Both of these species are fairly cosmopolitan, having expanded their home ranges as a result of the aquarium trade and commercial shipping respectively.

B. glabrata lives in fresh water, and belongs to a large genus which is found natively in both the old and new worlds. It is of interest to developmental biology in general and the field of L/R asymmetry in particular as it is a sinistrally coiling species. The genus Biomphalaria (comprising approximately 34 species) and *B. glabrata* itself are best known for their role in the transmission of the parasites which cause schistosomiasis (bilharzia), a disease found in 70 countries and infecting approximately 200 million people worldwide (Feasey et al., 2010). The genetic sequences of parasites which can cause schistosomaiasis have been available for several years, but that of *B. glabrata* itself is still pending despite being identified as a priority target for genomic sequencing as early as 2004 (Raghavan & Knight, 2006). The immune system of this species is of particular interest, as it has been the target of research aimed at determining methods to combat this disease.

C. fornicata is growing in popularity as a developmental model, as it can be harvested in large numbers and kept in good condition in recirculating aquaria far from the sea. It is also amenable to a variety of traditional embryological approaches to experimentation, and may well have its genome sequenced in the near future (J. Henry, personal communication).

Figure 1.15: *Single adult* B. glabrata *(left) and stack of* C. fornicata *(right). Images from Wikimedia courtesy of Biologie, Uni-Hamburg and Steve Trewhella.*

1.8.5 Annelids

The annelids are a large phylum, consisting of around 17,000 species, made up of segmented vermiform organisms, which inhabit many different environments. The annelids have been the subject of substantial phylogenetic study and revision in recent years and possess no one single synapomorphy that separates them from other phyla. They generally share, however, a series of features that demarcate the phyla and were originally used as the basis for taxonomic differentiation. One of these is referred to in their name, which comes from the Latin *annelus*, meaning 'little ring', describing the segments, resembling rings, that can generally clearly be seen making up their bodies.

The annelids are now said to comprise two classes, the Errantia and the Sedentaria, with many traditional classes having been rendered non-monophyletic (Struck et al., 2011). This is controversial, however, and is covered in more depth in Section 1.8.6 below. The Errantia are named for their tendency towards a free-living lifestyle, in contrast with the often immobile (as adults) Sedentaria. More traditionally, annelids were separated into the Polychaeta (bristle worms, names for the hairy chaetae along their sides), Hirudinea (leeches) and Oligochaeta (including earthworms) along with the Pogonophora, Echiura and Sipuncula, which were previously defined as separate phyla, but are now known to belong within the Annelid clade (Struck et al., 2007).

Annelids possess long, segmented bodies, generally divided internally by 'septa' (partitions) which can be noted externally from the ring-like structures called "annuli" which surround them, creating a slight indentation. Segments are

very similar, sharing the gut, nervous system and circulatory system with the rest of the body, and largely mirroring the arrangement of organs found in every other segment (Rouse, 1998). In polychaetes, each segment also possesses a pair of parapodia, which are often used for locomotion. A collagen cuticle surrounds the body itself, formed by secretions from the outer layer of cells.

Most species of annelid reproduce sexually, although asexual reproduction is by no means unusual, and some annelids are hermaphroditic. Annelids are said to be either direct or indirect developers - the former resembling miniature adults from the point of hatching, while the latter go through a trochophore larval stage as described in Section 1.7.4. Indirect development is particularly common in marine annelids, where they represent the primary method of dispersal, travelling as they do in the plankton. Indirect developers probably represent the ancestral state, as the highly conserved trochophore bauplan is found in many spiralian phyla, as detailed earlier in this report.

1.8.6 Annelid Phylogeny and the Serpulidae

Historically, annelids were grouped with arthropods in the supergroup 'Articulata' due to the apparently shared segmentation seen in both phyla. With the advent of molecular phylogeny, however, it has become more apparent that annelids are closely related to the other members of the Lophotrochozoa (Halanych, 2004), a hypothesis strongly supported by embryology, as annelids demonstrate typical spiralian cleavage in the process of development (Shankland & Seaver, 2000).

The traditional view of intra-annelid phylogeny has also been the subject of considerable revision in recent years, due to the application of more modern phylogenetic techniques. Historically, annelids were divided into Polychaeta, Oligochaeta and Hirudinea classes, or in some cases the Hirudinea (leeches) were folded into the Oligochaeta to create the Clitellates. More recent reexaminations have greatly restructured the phylogeny, most recently by Struck et al. in 2011, whose phylogeny can be seen in Fig. 1.16. Most strikingly, instead of being divided into a range of around 20 'families' according to earlier taxonomic standards, only two clades are suggested by recent analysis - the Errantia and Sedentaria, each adapted

to a different way of life, as suggested by their nomenclature.

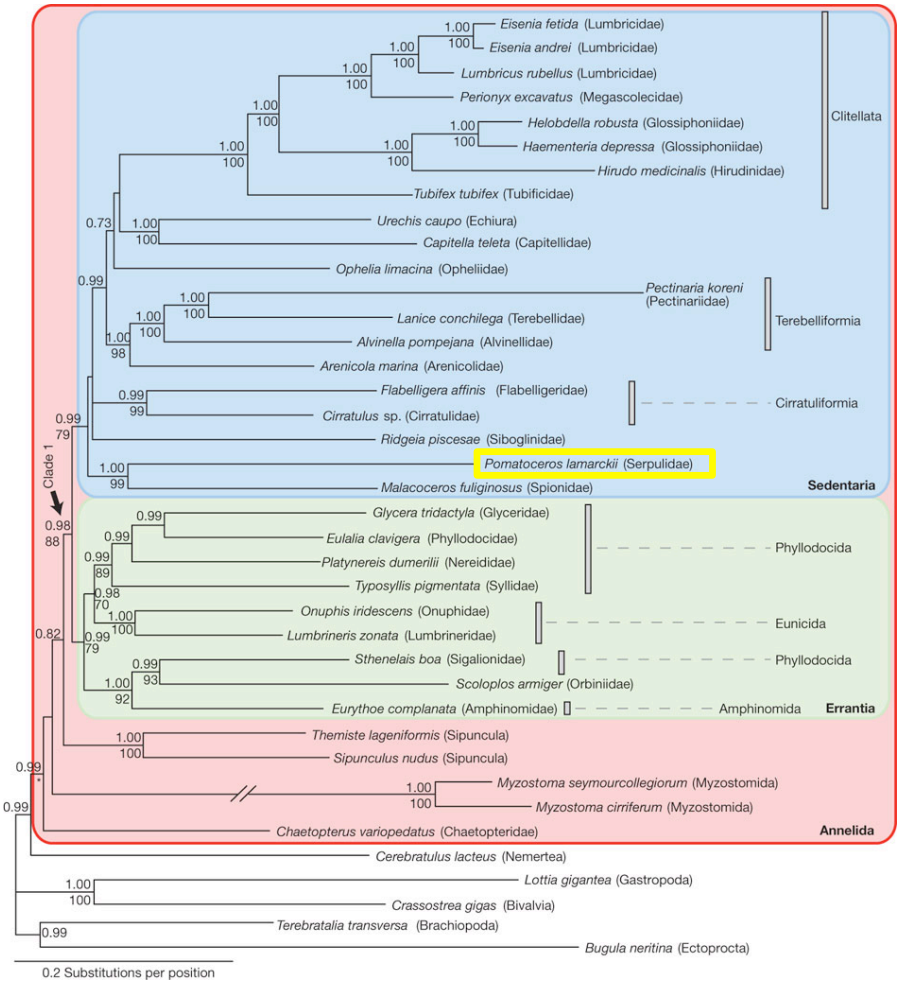


Figure 1.16: *Annelid phylogenetic relationships, as determined by analysis of 231 gene fragments at 47,953 positions. Note position of* P. lamarckii*, highlighted in yellow, basal to the Sedentaria. Figure reproduced from Struck et al. (2011).*

The Serpulidae is a family of polychaete, tube building annelids, to which our model organism, *Pomatoceros lamarckii*, belongs. Its position at the base of the Sedentaria, seen boxed in yellow in Fig. 1.16, make it well suited for determining ancestral characters in comparison with more derived annelid models, a trait that was of much weight when selecting *P. lamarckii* as our subject for research, which contrasts greatly with more established annelid models, such as the leech *Helobdella robusta* and the polychaete *Capitella teleta*.

Figure 1.17: *The Annelid* P. lamarckii*, outside of its tube after dissection (top) and in situ in its tube on a rock on the seashore. Images by author.*

1.8.7 *Pomatoceros lamarckii*

The annelid *Pomatoceros lamarckii* is a tube building serpulid which is widespread in intertidal and sub-littoral zones around the United Kingdom and northern Europe. They are found attached to firm substrates, from rocks, as can be seen in Fig 1.17, to animal shells to man made structures, and often are noted for their detrimental effect on shipping (Hamer et al., 2001; Hamer, 2002). As noted above, the position of *P. lamarckii* in annelid phylogeny makes it a very useful model for developmental studies (Struck et al., 2011).

P. lamarckii is also an incredibly useful model for developmental work, as it provides a readily accessible source of embryonic and larval material (McDougall et al., 2006). It possesses a stereotypical mode of development (Segrove, 1941) that will be vital for interpreting any observed mechanisms of establishing asymmetry in light of that known from other phyla, with an absence of the derived unequal cleavage, polar lobes or direct development seen in other species.

P. lamarckii exhibits what is generally reckoned as the ancestral annelidan reproductive strategy, with two distinct sexes which release gametes into the water column via nephridia (Ruppert et al., 2004). Upon fertilization, the zygote then develops into a typical trochophore larvae, which grows as plankton, as described in 1.7.4 (Rouse, 1998). Once sufficient size is attained, the larvae sink to the seabed.

There, the anterior part of the trochophore (between the apical tuft and prototroch) develops into the head (prostomium), followed immediately posteriorly by the peristomium (around the mouth). The rest of the embryo develops into the "growth zone", which produces segments in the adult, with the exception of the extreme anterior, which develops from the anus of the trochophore into the pygidium (tail) (Ruppert et al., 2004).

Annelids are increasingly well resourced at the genetic level after a long period of scarcity in this regard. The leech *Helobdella robusta* and the polychate *Capitella teleta* both have their genomes published (Simakov et al., 2013), and *Platynereis dumerilii* has a number of genomic resources in production. At the beginning of this project, *P. lamarckii* EST libraries had been published (Takahashi et al., 2009), allowing initial stages of molecular work to proceed, although the production of transcriptomic resources, and more latterly a genome has markedly increased the depth of data to hand. The phylogenetic position, availability, reliability and stereotypical development of *P. lamarckii* therefore made it an excellent choice as a model for investigating how L/R asymmetry is established in the Lophotrochozoa.

1.8.8 The Rotifer *Brachionus plicatilis*

The phylum Rotifera consists of about 2000 described species, which reside in both fresh and salt water, and consists of three classes, the Monogonota, which are perhaps the most typical, with around 1500 species, the Bdelloidea, which are asexual, and of largely freshwater origin, and the Seisonidea, with only two recognised species (Segers, 2002). The most natural candidate for use as a rotifer model for studying the evolution of development is the euryhaline monogont rotifer *Brachionus plicatilis* Muller, (Suatoni et al., 2006) a species distributed worldwide, commonly used in a range of ecological studies, and used commercially for aquaculture.

Rotifer are extremely tractable experimentally, with a short generation time, few care requirements and a history of use as a model for ecology and population dynamics studies (Hagiwara et al., 1995; Wallace, 2002). They can be transfected for RNA interference studies (Timmons & Fire, 1998) through consumption of *E. coli*

expressing double stranded (ds) RNA (Cridge & Dearden, unpublished results), while lipofection and electroporation with short interfering (si) RNA have also been demonstrated in resting and amictic eggs (Shearer & Snell, 2007). They have a variety of structures of developmental interest, including a centrally located eye, a single foot, modified spiral cleavage, and a non segmental body plan, comprised of relatively few cells with a complex organisation (Wallace, 2002). Interestingly for the present thesis, they are asymmetric along the L/R axis, with the ovaries found predominantly on a single side of the adult body.

The evolution of developmental novelties can also be examined in rotifers, as they posses many uncommon features, such as a "corona", or wheel organ, used for feeding, a single eye, and an unpaired "foot" (Wallace, 2002). The rotifer *B. plicatilis* is tolerant of a wide variety of temperatures, salinities and environmental conditions. Due to this flexibility, it is commonly found in estuarine areas and salt lakes around the globe, and is widely used as essential food source in raising marine fish, shrimp and crab larvae due to its tolerance to a wide range of environments (Lubzens, 1987; Dhert & Sorgeloos, 1994).

B. plicatilis is a member of the Monogont class of Rotifera, which is comprised of around 1500 cyclically parthenogenic species. Under most conditions reproduction is via parthenogenesis, but in adverse environments sexual reproduction is induced by accumulating chemical signals produced by the rotifer themselves (Gilbert, 1963; Stelzer & Snell, 2003), and a complete sexual cycle results in the production of "resting eggs", fertilised but dormant embryos which can withstand dessication (Gilbert, 1974). These not only provide a means to withstand seasonal drought, but also represent a means of dispersal (Gomez et al., 2002).

Some work in rotifer has focused on Bdelloids, including genomic sequencing, but, as obligately parthenogenic species these do not express sexual genes, which are often of interest, and possess a degenerate tetraploid genome (Flot et al., 2013) and common horizontal gene transfer that, while interesting, makes comparison of developmental mechanisms very difficult . Unlike Bdelloids, Monogont rotifer are representative of general Rotifera *Bauplan*, developmental processes and genetic composition (Wallace, 2002).

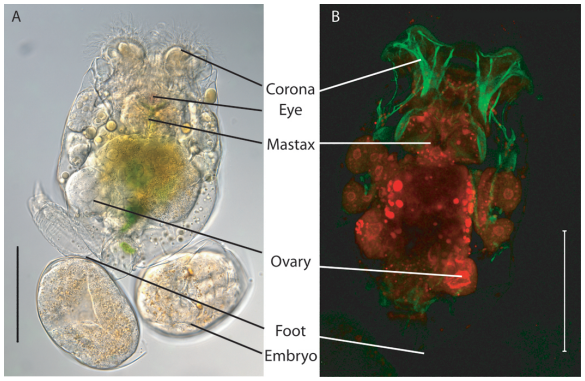


Figure 1.18: *Key features of the rotifer* Brachionus plicatilis. *Images courtesy of Dr. Peter Dearden. Image at right has been stained with propidium iodide (red) for nuclei and phalloidin (green), showing location of actin. Scale bar is 100 μm in length.*

B. plicatilis provides a good representation of the features of the Monogonta in general (see Fig. 1.18. It is a basal member of the Rotifera, its body form, physiology and behaviour is typical of the group (Mark Welch, 2001), and its ubiquity and hardiness make it perfect for use as a laboratory model. It can be easily raised in a variety of media and on a range of diets, and is easily sampled. This has lead to a history of use in studies of population dynamics, ecology and aging, and recommends it as a model for the Lophotrochozoa in general and Rotifera in particular. It is a diploid organism, with a genome of 0.7 pg (around 700 megabases) (Mark Welch & Meselson., 1998), and is well represented in the Rotifera literature, including an online expressed sequence tag (EST) library (Suga et al., 2007).

1.9 Shared Mechanisms of Establishing Asymmetry?

As is the case in many fields of biology, work in the determination of L/R asymmetry was begun in the chordates, specifically in vertebrates such as mice, chicks and *X. laevis*. This work has since been extended to other deuterostomes, such as sea urchins, urochordates and cephalochordates. The next step is to develop our understanding of the true conservation of the mechanisms of breaking, establishing and maintaining L/R asymmetry across the Metazoa.

It is somewhat surprising that the traditional workhorses of molecular biology - ecdysozoans such as *D. melanogaster* and *C. elegans* - have been generally neglected in this regard. Protostomes - both ecdysozoans and lophotrochozoans - do exhibit L/R asymmetries, but they can be slight and difficult to observe, especially in the case of the Ecdysozoa. Examples of consistent ecdysozoan L/R asymme-

try include differential coiling of internal organs, altered neural identity on either side of the L/R axis and (perhaps the most widely used arbiter of asymmetry in *D. melanogaster*), the arrangement of genital organs (Okumura et al., 2008). Some lophotrochozoans, in contrast, exhibit marked L/R asymmetry, the prime example being the coiling of snail shells, which conveys, at a glance, information about asymmetry in these species. In many ways it seems easiest to determine the conservation of L/R asymmetry in these organisms, rather than more longstanding but in this case less experimentally tractable ecdysozoan models.

While it is clear that deuterostomes and lophotrochozoans develop in markedly different ways, Grande & Patel (2009) have shown that at least some aspects of the L/R asymmetry establishment and maintenance cascade are shared between molluscs and vertebrates. Using lophotrochozoan models, I aim to see just how many of the implicated mechanisms listed earlier in this introduction retain a role in lophotrochozoan development, from the first signs of asymmetry though to the latter stages of development, when L/R asymmetry is fully patterned, and in doing so, discern something about the evolution of this vital step in the development of bilaterian organisms.

General Methods

2.1 Laboratory Methods

2.1.1 Enzymes, Chemicals and Solutions

Biological reagents were acquired from Invitrogen and Roche Applied Biosystems unless otherwise specified. All chemicals were sourced from Sigma, with the exception of those noted. Non-S.I. abbreviations used below can be found in the abbreviations list, Appendix A. Solution contents can be found catalogued in Appendix B.

2.1.2 Equipment Used

An Eppendorf BioPhotometer spectrophotometer was used to quantify DNA and RNA samples with UVette cuvettes. A UVITEC UVidoc GelDoc was used to image agarose gels. An Eppendorf Mastercycler Gradient or a Techne TC-512 thermal cycler were used for PCR procedures.

2.1.3 *P. lamarckii* Collection, Maintenance and Spawning

Adult *P. lamarckii* and were gathered from Tinside, Plymouth, UK, along with their substrate, and transported to Oxford for storage. They were kept at 12°C in a recirculating aquarium system until dissection and gamete fertilization.

To remove *P. lamarckii* from their tubes, the rear of the tube was crushed using dental pliers. The *P. lamarckii* were then pushed backwards out of their tube using a metal probe. Upon being placed individually in a cell culture well in filtered sea water (FSW) worms began releasing gametes. Eggs were then transferred to new

cell culture wells, and a droplet of dilute sperm solution (in FSW) added. After ten minutes, eggs were rinsed in fresh FSW, covered, and transferred to an incubator at 12-14°C.

2.1.4 *P. vulgata* Collection, Maintenance and Spawning

Adult *P. vulgata* were gathered from Tinside, Plymouth, UK, or Northney Marina, Hayling Island, Portsmouth, UK, and transported to Oxford for storage. They were kept at 12°C in a recirculating aquarium system until dissection and gamete fertilization.

To obtain gametes, individual adult *P. vulgata* were euthanised, and gametes liberated from the body by dissection of the gonads. Eggs and sperm were kept separate initially, and care was taken not to allow these to mix to ensure near-simultaneous fertilization where possible. Sperm was incubated in FSW at 12 -14 °C, and eggs were washed twice in 12-14°C FSW, to remove detritus and blood. Eggs were then incubated in alkali FSW (5mM NH_4Cl in FSW, adjusted to pH 9.0) for 20 to 60 minutes, until the geminal vesicle disappeared and follicle cells minimised in size. Eggs were then washed three times in fresh FSW, and incubated at 12 -14°C for one hour. Eggs were once more washed in FSW. Sperm was checked for activity using a binocular microscope, and the most active sperm sample chosen for use. 1 drop of sperm per 10 mL of eggs in FSW was added, and the mix left to fertilize for 30 minutes at 12 -14°C. Fertilized embryos were then washed twice in FSW and left at 12 -14°C to develop until fixation.

2.1.5 Ion Channel Blocking Drugs

After fertilization and two washes in filtered sea water (FSW), approximately equal numbers of embryos were placed into 2 mL volumes of FSW in 24 well cell culture dishes and pharmacological treatments applied as listed in Table 2.1. Five biological replicates (combinations of individual male and female *P. vulgata*) were performed for each inhibitor tested.

At 24 hours post the time of first fertilization, cell culture dishes were viewed

Table 2.1: Drug Concentrations Used in Pharmacological Experiments

Drug (Concentration)	Function	Conc. 1	Conc. 2	Conc. 3	Conc. 4	Conc. 5
Omeprazole (μgmL^{-1})	$H^+/Na^+/K^+$-ATPase	0 (DMSO only)	20	40	80	160
Ouabain (μgmL^{-1})	$H^+/Na^+/K^+$-ATPase	0 (DMSO only)	6.25	12.5	31.25	62.5
Lansoprazole (μgmL^{-1})	$H^+/Na^+/K^+$-ATPase	0 (DMSO only)	20	40	80	160
Skelid (μM)	V-ATPase	0	25	50		
Concanamycin A (ng/mL)	V-ATPase	0 (DMSO only)	0.5	1		
Barium Chloride (μgmL^{-1})	K Channel	0	50	100		
9-anthrocene Ca Acid (μgmL^{-1})	Cl Channel	0 (DMSO only)	15	30		

under a dissection microscope, and numbers of embryos in each of the states used in the graphs shown in Chapter 4 (dead, alive but not positively phototactic and negatively geotactic, alive and positively phototactic and negatively geotactic), noted. Motile embryos were transferred to a 1.5 mL eppendorf tube and fixed overnight in 4% w/v paraformaldehyde (TABB) in MOPS buffer, pH 7.5. Samples were then washed incrementally into 100% methanol, and stored at -20 °C.

2.1.6 Embryo Fixation for *In Situ* Hybridization

To fix for *in situ* hybridization, embryos were allowed to sink to the bottom of 1.5 mL eppendorf tubes in FSW. The FSW supernatant was then removed and replaced with 4% (w/v) paraformaldehyde dissolved in MOPS buffer, with the fixative added equating to at least five times the volume of the embryos to be fixed. The embryos in fixative solution were then placed at 4°C overnight.

Fixative solution was removed by washing stepwise into PBS buffer with 0.1 % Tween (PTw), The stepwise wash process proceeds as follows:

- First 25% of the volume of the fixative (the supernatant) was replaced with the same volume of PTw, and embryos were allowed to settle once again to the bottom off the eppendorf tube.
- 50% of the volume of supernatant was then removed and replaced with fresh PTw, embryos were allowed to settle as above.
- 75% of the total volume of supernatant was then removed and replaced with PTw and the embryos were given time to reach the bottom of the tube under the force of gravity.
- The majority of the supernatant was then removed and discarded, and re-

placed with 1 mL of PTw.

- Two full washes, where the supernatant is removed and replaced with PTw, then followed.

The stepwise wash procedure described above was then utilised once more, with methanol taking the place of PTw. Embryos were then kept in methanol at -20 °C until used for RNA *in situ* hybridization, as described in section 2.1.18

2.1.7 Embryo Fixation for Antibody Staining

For antibody staining, embryos were allowed to sink under the force of gravity to the base of 1.5 mL eppendorf tubes in FSW. The majority of the supernatant FSW lying above embryos was then removed and discarded, and replaced with 4% (w/v) paraformaldehyde dissolved in FSW. Embryos were left to fix for 20 minutes at room temperature, then washed stepwise as described in section 2.1.6 into 100% methanol.

2.1.8 RNA Extraction

RNA was extracted using the Invitrogen TRIzol PLUS using standard manufacturer protocols summarized as follows. 100 mg (wet weight) of sample was lysed in 1 mL of Trizol reagent (Invitrogen). This was left at room temperature for 5 minutes, before 0.2 mL of 100% chloroform was added and vortexed thoroughly for 15 seconds. The sample was incubated at room temperature for 2 minutes, and then centrifuged at 13000 g for 15 minutes.

The upper, aqueous phase was pipetted into a new Eppendorf tube, and 1/10th of its volume of 3 M Sodium Acetate solution and three times its volume of 100% ethanol was added. The samples were then incubated at -20°C for 2 hours and centrifuged at 13000 g for 30 minutes. The supernatant was discarded, and the pellet, containing RNA, was washed in 1 mL of 70% ethanol/30% DEPC treated water, with the sample left to incubate for 8 minutes in the ethanol wash solution, followed by centrifugation at 13000g for 5 minutes. The wash process in 1 mL of 70% ethanol/30% DEPC treated water was repeated, and the sample was then left

to air dry at room temperature for 5 minutes, then resuspended in 50 μL of DEPC treated water.

The sample was then quantified using a spectrophotometer as described in section 2.1.2. The sample was then stored at -80°C.

2.1.9 cDNA Production

First strand synthesis was performed using the Invitrogen SuperScript III Reverse Transcriptase Kit. 500 nM Oligo-dT, 3 μg of total RNA, 0.1 mM dNTP was made to 10 μL with RNAase-free water and heated to 65°C for 5 minutes, before placing on ice for 1 minute. This was briefly centrifuged to ensure the solution was at the base of the tube. 4 μL of Invitrogen 5x First Strand buffer, 0.1 μM DTT, 15 U "RNAase OUT RNase inhibitor" and 200 U Superscript III reverse transcriptase were added and gently mixed, before being incubated at 50 °C for 60 minutes. The solution was then heated to 70°C for 15 minutes to inactivate all enzymes.

2.1.10 PCR Protocol

Unless otherwise stated, the PCR protocol used for gene amplification was as follows: Step A: 94°C for 2 Minutes, Step B: 94°C for 30 seconds, Step C: 55°C for 30 seconds, Step D: 72°C for 2 min; repeat B - D 34 times; Step E: 4°C until retrieved. 5 μL of 10x PCR Buffer, 1.5 μL $MgCl_2$, 1 U of Taq polymerase, 0.1 mM of dNTPs were made to 50 μL with DEPC treated H_2O for each sample. Bioline Taq polymerase, 10x PCR Buffer, $MgCl_2$ and 10mM dNTP used, unless otherwise specified.

2.1.11 Agarose Gel Electrophoresis

All agarose gels used were 1% agarose in 1x TAE (w/v). 1x TAE and 0.5 μgmL^{-1} ethidium bromide was used as electrophoresis buffer. Invitrogen 1kb+ ladder was run on at least one lane of each gel to provide a reference for sizing fragments. Gels were visualised on a UVITEC UVidoc GelDoc.

When used for cloning, bands were excised from gels using disposable razor blades, and to purify DNA an Illustra GFX PCR DNA and gel band purification

kit was used according to manufacturers instructions. DNA was eluted in 30 μL of elution buffer for subsequent cloning as described below.

2.1.12 RACE PCR

5′ and 3′ RACE ready DNA was prepared using a Clontech SMART RACE kit via the standard manufacturer protocol. Gene specific primers were designed as in section 2.2.9, with a primer length greater than 23 nucleotides, Tm set to greater than 65°C and a minimum GC content of 50% imposed. The following protocol was used for both first and second round PCR, when a second internal round was required:

Step A: 94°C for 4 Minutes, Step B: 94°C for 30 seconds, 72°C for 3 minutes, repeat step B 5 times. Step C: 94°C for 30 seconds, 70°C for 30 seconds, 72°C for 3 minutes, repeat step C 5 times. Step D: 94°C for 30 seconds, 68°C for 30 seconds, 72°C for 3 minutes, repeat step C 25 times. Step E: 4°C until retrieved. DNA was then run on an agarose gel and purified as described previously in section 2.1.11.

2.1.13 DNA Cloning and Transformation

Each gene was then cloned into Invitrogen pCRII vector using the Roche Rapid DNA Ligation Kit the following protocol. 3 μL of poly-A tailed DNA (sourced from gel purified PCR reactions as described above) was added to 5 μL of T4 DNA Ligation Buffer (2x conc.) and 2 μL of DNA Dilution Buffer, 5x conc. 0.5 μL of Invitrogen pCRII vector and 1 U of Roche T4 DNA ligase were then added sequentially. Reactions were mixed gently by pipetting, and left at room temperature for not less than one hour to ligate.

For each transformation, 25 μL of Bioline α-Select Competent Cells (Bronze Efficiency, Genotype F- deoR endA1 recA1 relA1 gyrA96 hsdR17(rk-, mk+) supE44 thi-1 phoA Δ(lacZYA-argF)U169 Θ80lacZΔM15 λ) 5-α *Escherichia coli*, were thawed on ice. 6.5 μL of the ligation mixture described in the previous paragraph was added to each tube, and the bacteria and ligation mix were incubated on ice for 5 minutes. They were then heat shocked for 30 seconds at 42°C, before being returned to ice

for 2 minutes. 80 μL of SOC medium was added to each tube, and the tubes were incubated at 37°C, for 1 hour. The transformed bacteria from each tube were then plated onto prewarmed 37°C LB agar ampicillin plates,which had previously been treated with 0.1 mM IPTG and 40 μgmL^{-1} of X-gal to allow blue/white selection. These were incubated at 37°C overnight, before white colonies were checked for inserts via colony PCR.

2.1.14 DNA Preparation from Transformants

Colony PCR was carried out on five white colonies from each plate, in order to confirm that inserts were of the expected size. These colony PCRs were carried out using the cycle described in section 2.1.10 above, with 0.4 μM of each M13 forward and M13 reverse primer, 2 μL of 10x PCR Buffer, 1 U of Taq polymerase and 0.1 mM dNTPs made up to 20 μL total volume with milliQ filtered H_2O.

A new petri dish containing LB agar and ampicillin was divided into a grid, denoting a specific position for each of the samples that were to be subjected to colony PCR. For each of the samples tested by colony PCR, a pipette tip was used to collect a sample of bacteria from a single white colony, which was first touched to its specific point in the grid of the new petri dish, then placed in the PCR mix. After resting in the colony PCR mix, the pipette tip was shaken, removed from the mix and discarded. Tubes were capped, PCR was performed and the samples were run on an agarose gel as described above. While PCR was performed, the petri dish was placed at 37°C to allow colony growth.

Those colonies found to contain an insert of the correct size were collected from their location on the petri dish using a pipette tip and placed into 3mL of LB broth containing 0.5 μgL^{-1} ampicillin, and grown overnight, shaken at 120 RPM at 37°C.

The overnight samples were spun at 4000 g for 5 minutes at 4°C. The media supernatant was then removed, autoclaved and discarded. The plasmid DNA was then extracted from the bacterial pellet using a Qiagen Qiaprep Miniprep kit according to standard manufacturer protocols (`www.qiagen.com/literature/render.aspx?id=370`).

Following extraction, the samples were subjected to a diagnostic digest to con-

firm they contained fragment DNA of similar size to that originally purified from PCR. 2 μL of plasmid DNA, 2 μL of New England Biolabs EcoR1 digestion buffer, 1 μL New England Biolabs EcoR1 restriction enzyme and 15 μL of milliQ filtered water were placed at at 37 °C to digest for a minimum of 2 hours. The resulting digested DNA was then run on a gel as described in section 2.1.11. Where DNA bands were approximately equal in size to those gained by PCR and excised in the first instance, an aliquot of the miniprepped plasmid DNA was sent for sequencing.

2.1.15 Sequencing

For each desired sequencing reaction, 4 μL of plasmid DNA, 2 μL milliQ filtered water, 2 μL of Applied Biosystems 5x sequencing buffer, 1 μL of Applied Biosystems BigDye Terminator Ready Reaction Premix and 1 μL of the desired primer at a concentration of 3.2 pmol/ μL were placed in a 200 μL eppendorf tube and gently mixed by pipetting up and down. The following PCR sequencing reaction was then performed:

Step A: 95°C for 30 seconds, Step B: 50°C for 20 seconds, Step C: 60°C for 4 minutes repeat steps A-C 34 times. Step D: 4°C until retrieved.

DNA was then precipitated using 1/10th of its volume (1 μL) of 3M sodium acetate and 3 times its volume of -20°C, 100% ethanol (30 μL) and incubated for a minimum of 2 hours at -20°C. The sample was then centrifuged at 13000 g for 30 min. The supernatant was discarded, and the pellet was washed by gentle pipetting in 0.5 mL of 70% ethanol/30% DEPC treated water, followed by centrifugation at maximum speed for 5 minutes. The wash solution supernatant was discarded, and the pellet air dried at room temperature in the dark for 10 minutes.

Sanger sequencing was then performed by the Sequencing Unit, Zoology Dept, University of Oxford. The NCBI Blastx tool (NCBI, 2011) was used to confirm the identity of these sequences, and also to determine direction of insertion of the gene within the pCRII vector relative to the primer used for sequencing, and thus determine the appropriate RNA polymerases for probe synthesis via the protocol described in section 2.1.16 .

2.1.16 DIG Labelled RNA Probe Preparation

A modified version of the protocol established by Dearden et al. (2002) was used for probe manufacture. For each of our genes of interest, linear templates for RNA probe synthesis were generated by digestion of plasmids containing the cloned gene fragments. Unless internal restriction sites were found within cloned gene fragments, pCRII template containing gene of interest were cut with New England Biolabs sourced BamH1 and Xho1 restriction enzyme, with an appropriate substitute used when internal restriction sites were in fact present . 5 μL of plasmid, 5 μL of the appropriate digestion buffer, 2 μL of restriction enzyme and, when necessary, 0.5 μL of 100 x BSA solution, made up to 50 μL with DEPC treated water. This was left in a 37°C water bath for digestion overnight.

To ensure complete digestion and a lack of internal restriction sites, 2 μL of the resultant digest was run on a gel. When only a single band of the appropriate length was present, the DNA was purified by phenol/chloroform extraction and ethanol precipitation as follows. 52 μL of DEPC water was added to the digest for a total volume of 100 μL. 100 μL of 1:1 phenol (pH 7.9)/chloroform was then added, and vortexed thoroughly. The phenol/chloroform/digest mix was then spun at maximum speed in a centrifuge for 5 minutes. The upper, aqueous layer was transferred to a fresh eppendorf tube and the interphase and lower organic layer discarded. 100 μL of chloroform was then added, the sample vortexed, and spun again at maximum speed for 5 minutes. The upper layer was then transferred to a fresh eppendorf tube, and 1/10th of its volume (10 μL) of 3M sodium acetate and 3 times its volume of -20°C, 100% ethanol (300 μL) were added, and incubated for a minimum of 2 hours at -20°C. The sample was then centrifuged at 13000 g for 30 min. The supernatant was discarded, and the pellet was washed by gentle pipetting in 0.5 mL of 70% ethanol/30% DEPC treated water, followed by centrifugation at maximum speed for 5 minutes. The wash solution was removed, the 70% ethanol/30% DEPC final wash process was repeated, and after all wash solution was removed the sample was then left to air dry at room temperature for 5 minutes, then resuspended in 11 μL of DEPC treated water.

The sample was then taken and 2 μL of 10x transcription buffer, 2 μL of DIG RNA labeling mix, 0.1 μM of DTT and 15 U of RNase inhibitor (all Roche Molecular Diagnostics) were added. This mix was then combined with 40 U of the appropriate polymerase, either SP6 or T7, according to the direction that the DNA fragment had transformed into the pCRII vector. The mix was incubated at 37°C for 2 hours and treated with 15 U DNase I for 30 minutes at 37°C to remove template.

RNA was then precipitated in 1/10x volume of 3M sodium acetate (2 μL) and 2.5x volume of -20°C, 100% ethanol (50 μL), and incubated for a minimum of 2 hours at -20°C. The sample was then centrifuged at 13000 g for 30 minutes. The supernatant was discarded, and the pellet, containing DNA, was washed in 0.5 mL of 70% ethanol/30% DEPC treated water, with the sample left to incubate for 8 minutes in the ethanol wash solution, followed by centrifugation at 4000g for 5 minutes. The wash process was repeated, and the sample was then left to air dry at room temperature for 5 minutes. 20 μL of DEPC water was then added to resuspend the RNA probes, 1 μL of which was then run on a gel to check the efficiency of transcription. If the transcribed probe was approximately the known length of the cloned product, 40 μL of hybridization buffer was added, and the DIG labelled probes were stored at -80°C until use.

2.1.17 Immunohistochemistry Protocol

P. lamarckii and *P. vulgata* embryos were collected, fixed and stored as described in section 2.1.7. In 1.5 mL eppendorf tubes, embryos were washed stepwise from methanol into PTw (see section 2.1.6 for stepwise washing instructions). They were then washed 5 times in PTw to remove any trace of methanol, then washed twice for 15 minutes each in 80% PTw/20% heat treated sheep serum (HTSS) at 4°C. Samples were then incubated for 30 min in 2% w/v Roche Blocking Reagent in PTw at 4°C. Samples were then incubated overnight at 4°C in fresh 2% w/v Roche Blocking Reagent in PTw containing the appropriate dilution of primary antibody. For Rabbit anti-serotonin antibody (Abcam 10385), this was found experimentally to best be a 1:10,000 dilution.

The next day samples were washed thrice in 80% PTw/20% HTSS at 4°C for 5

minutes per wash, to remove excess unbound antibody. To ensure all unbound primary antibody was removed, samples were then washed thrice in 80% PTw/20% HTSS at 4°C for at least 30 minutes per wash. Samples were then incubated for 30 min in 2% w/v Roche Blocking Reagent in PTw at 4°C. Samples were then incubated overnight at 4°C in fresh 2% w/v Roche Blocking Reagent in PTw containing the appropriate dilution of secondary antibody. For Donkey anti-Rabbit antibody (Abcam 97061 - anti-DIG conjugated), this was found experimentally to work most accurately at a 1:1000 dilution.

Samples were then washed at least three times in 80% PTw/20% HTSS at 4°C for 5 minutes per wash, to remove excess unbound antibody. To ensure no unbound secondary antibody was present, samples were then washed thrice in 80% PTw/20% HTSS at 4°C for at least 30 minutes per wash. Samples were then washed twice in Mg-free AP buffer (see Appendix C) and twice in AP buffer, then into staining solution. Samples were left to stain until satisfactory signal had developed, then washed twice in Mg-free AP buffer, then into PTw three times, then into 70% Glycerol/30% PTw for imaging.

2.1.18 *P. vulgata* RNA *In Situ* Hybridization Protocol

Embryos, fixed and stored as described in section 2.1.6, were washed stepwise out of methanol into PTw (see section 2.1.6 for stepwise washing instructions). They were then incubated in 1.5 mL eppendorf tubes in 200 μL PTw containing 4 μgmL^{-1} proteinase K for 5 minutes, then washed twice in excess PTw. Embryos were then postfixed in 4% paraformaldehyde in MOPS for in excess of 30 minutes, then washed thrice in PTw. As much PTw was then removed as possible, and the embryos washed twice in hybridisation solution.

1 μL of each DIG labeled probe to be used for experimentation was preincubated with 500 μL of fresh hybridization buffer at 60 °C. Hybridization buffer was removed from samples, and the probe and hybridization buffer mix was added. The samples were then incubated at 60°C overnight.

The probe and hybridization buffer mixes were then removed from each sample, 1 mL of *Ciona* wash buffer was added, and the sample was incubated at 60°C

for 5 minutes. Such wash steps were repeated four times, each with 60 minutes of incubation time at 60°C. Samples were then washed three times in 1 mL PTw at room temperature, then incubated in 1 mL of block solution, consisting of 80% PTw/20% HTSS for a minimum of 2 hours on ice. Fresh 80% PTw/20% HTSS was then prepared to make antibody solution, containing 1:3000 v/v anti DIG-AP antibody (Roche). This was left to preabsorb for a minimum of two hours, alongside the samples in block solution, on ice. Block solution was then removed from samples, and antibody solution was then added. Samples were then incubated at 4°C overnight.

The next day samples were washed five times for at least 5 minutes in 1 mL PTw at room temperature, and then twice in Mg free AP buffer, made as per Appendix C. Samples were then washed twice in 1 mL AP buffer for 5 minutes at room temperature. Each sample was then transferred to a cell culture dish, allowed to settle, and excess AP buffer was removed. AP staining buffer was then made and added to the samples. The samples were placed in a dark environment, and checked every 10 minutes to monitor staining.

Once samples were stained they were washed once in Mg free AP buffer, then three times in 1 mL PTw to remove staining buffer. They were then destained by stepwise washing (as previously described) into methanol. Samples were left in methanol overnight to destain. Stepwise washing into PTw was then carried out, then samples were washed three times in 1 mL PTw, and finally placed in 1 mL of 70% glycerol/30% PTw for storage and visualisation.

2.1.19 DAPI Nucleic Acid Staining Protocol

Samples were placed in 500 μL of of 70% glycerol/30% PTw solution, and 1 μL of 1mg/mL DAPI was added. This was left at room temperature in a dark environment for 2 hours before photography.

2.1.20 Sub-cloning Protocol

10 μL CiCrys-EGFP vector Shimeld et al. (2005) or 10 μL of cloned *P. vulgata* genomic DNA sequence, 2 μL each of Sal1 and BamH1 restriction enzymes, 10 μL of NEB restriction buffer 3.1 and 76 μL of milliQ filtered water were left at 37°C overnight. The mixture was purified using an Illustra GFX PCR DNA purification kit according to manufacturers instructions, with DNA eluted in 20 μL of buffer. The DNA eluate was then run and gel purified on a 1% agarose gel as described in section 2.1.11, with appropriately sized fragment excised and purified.

To ligate vector and DNA, 5 μL of Roche T4 DNA Ligation Buffer (2x conc.), 2 μL of Roche DNA Dilution Buffer, 5x conc., 1 μL of linearised EGFP vector, 1 μL of the appropriate DNA band and 1 U of Roche T4 DNA ligase were mixed and left at 4°C overnight. The ligated product was then transformed into bacteria and mini prepped as described previously (section 2.1.13).

2.1.21 Microscopy

Photographs of samples were acquired using a Axioskop 2 plus microscope with FluoArc HBO100 fluorescent and HAL100 bright field illumination and an AxioCam HRc camera. All samples were visualised on Labserv 25mm x 75mm x 1mm slides, covered by Labserv '1' 18mm x 18mm coverslips.

2.1.22 RNA Preparation for Next-Generation Sequencing

The *P. lamarckii* embryonic RNA sample used in this analysis was from the same stock as used in an earlier EST screen (Takahashi et al., 2009), however, only RNA from individuals 24 hpf - 72 hpf was utilized. The quality of the total RNA was checked using an Agilent 2100 Bioanalyzer, with the RNA 6000 nano kit used to prepare the sample for analysis. Samples were prepared using an Illumina mRNA-Seq kit incorporating poly-(A) selection and sequenced on a single lane by the High-Throughput Genomics unit at the Wellcome Trust Centre for Human Genetics, Oxford, on an Illumina GAIIx platform.

2.1.23 Genomic DNA Preparation for Next-Generation Sequencing

Gonads were dissected from a single male *P. vulgata*, and left in a petri dish in FSW to allow sperm to disassociate. Large fragments of somatic tissue were then removed from the petri dish, and the solution transferred to a 15 mL falcon tube. The solution was then spun at 4000 RPM at 4°C for 5 minutes. The supernatant was then removed, and washed thrice in 3 times its volume of 1x PHB (for components see Appendix 3), and spun at 4000 RPM at 4°C to pellet following each wash step. 1 mL PHB containing 3 μL of 5 M NaCl and 60 μL of 10 mg/μL Proteinase K was then added to the pellet, which was gently pipetted. This was then left overnight at 50°.

After digestion, the solution was phenol/chloroform extracted as described in Appendix 2.1.16, a process which was repeated three times more, until no identifiable protein layer was observed. The DNA pellet was then ethanol precipitated as described in Appendix 2.1.16. The washed pellet was then left to air dry at room temperature, and resuspended in 100 μL milliQ filtered water, and the concentration determined as described in Section 2.1.2.

2.2 Bioinformatic and Genomic Sequencing Related Methods

2.2.1 DNA Sequencing and Quality Control

Transcriptome RNA was prepared with a fragment size of 300 bp, and genomic DNA was prepared with fragment sizes of 200 bp and 500 bp for sequencing by the High-Throughput Genomics Group, at the Wellcome Trust Centre for Human Genetics, Oxford. A single lane of Illumina GAIIx plate was used for *P. lamarckii* RNA sequencing. A single lane of Illumina HiSeq2000 was used for genomic sequencing of *P. vulgata*, at 100 bp paired end read length. The resulting reads were made available to us in fastq format on an external server, and were downloaded for local analysis. Initial assessments of the quality of the transcriptome and genomic data were performed using FastQC (Andrews, 2011) with excellent quality

(median Phred quality higher than 30) observed through to the last base in both datasets.

2.2.2 *P. lamarckii* Transcriptome Assembly

Transcriptomes were assembled using Velvet (version 1.1.04) (Zerbino & Birney, 2008; Zerbino, 2010), Oases (version 0.1.8) (Schulz et al., 2012), SOAPdenovo (V1.05) (Li et al., 2010b), ABySS (version 1.2.7) (Simpson et al., 2009) and Trinity (release 2011-08-20) (Grabherr et al., 2011) on a local Linux X86 server and on the ORAC (Trinity, Oases) and HAL (ABySS) servers at the Oxford University Supercomputer Facility. Various *k*-mer lengths were used to assess performance, with 27-mers best representing the results of the various assemblers, with the exception of Trinity, which is limited to a *k*-mer size of 25. Oases assemblies of *k*-mer lengths of 21, 23, 25, 27 and 29 were used for the final additive multiple-*k* analysis. Transcriptome assembly metrics were ascertained using a Perl script, available upon request.

2.2.3 Removal of Redundancy from Multiple-*k* Build

Redundancy was removed from concatenated Oases transcriptome builds using Vmatch (version 2.1.6) (Kurtz, 2011) software with a search length of 100 bp and a 100% similarity cut off. The TGICL clustering tool (Pertea et al., 2003) was used to assess for further redundancy or overlap within my dataset, using the CAP3 program (Huang & Madan, 1999), with 100% similarity cut offs for assessing redundant contigs.

2.2.4 *P. vulgata* Genome Assembly

P. vulgata genomic DNA was assembled using the ABySS and SOAPdenovo assemblers (Simpson et al., 2009; Li et al., 2010b), with *k*-mer lengths between 31 and 61. The final assembly of the *P. vulgata* dataset used in the present manuscript was created using AbYSS, with a *k*-mer length of 57, using the abyss-pe driver script and all default settings. A final minimum contig length of 300 bp was imposed for

further analysis.

2.2.5 *B. plicatilis* Genome Assembly and Redundancy Removal

After initial assembly trials using Velvet (version 1.2.07) (Zerbino & Birney, 2008; Zerbino, 2010), SOAPdenovo (version 2.04) (Li et al., 2010b), and ABySS (version 1.3.3) (Simpson et al., 2009), Velvet was found to be the best performing assembler in general across a range of *k*-mer lengths. Assemblies were carried out using Velvet (version 1.2.07) (Zerbino & Birney, 2008; Zerbino, 2010) at *k*-mer lengths of 21, 31, 41, 51, 61, 71, 81 and 91, a *k*-mer cutoff limit of 3 for assembly and all other defaults. Vmatch (version 2.2.0) (Kurtz, 2011) software was run to remove redundancy with a search length of 100 bp and a 100% similarity cut off for contigs 100-499 bp in size, and a search length of 500 bp and a 100% similarity cut off for contigs 500 bp in length and over.

2.2.6 *B. glabrata* and *C. fornicata* Transcriptome Assembly

Trinity (release 2011-08-20) (Grabherr et al., 2011) on a local Linux X86 server was used to assemble both transcriptomes, with all default settings and a minimum contig length of 200 bp imposed.

2.2.7 Statistical and Phylogenetic Analysis

Statistical analyses (using the χ^2 test) were carried out using Excel using algorithms bundled within that program. For calculation and presentation of graphs, R64 v 1.43 was used where stated.

2.2.8 Gene Identification

Unless otherwise stated, genes were identified within extant *P. lamarckii* and *P. vulgata* genetic datasets by finding a homologue of the desired gene of known identity within the NCBI database (`http://www.ncbi.nlm.nih.gov/`). The protein sequence of these known homologues were used to conduct a tBlastn search (Altschul et al., 1997) as part of the Bioedit program (Hall, 1999). Putatively identified contigs

within datasets were then subjected to reciprocal BlastX analysis against the NCBI nr database to verify their identity.

For targeted searches for transcription factors and miRNA, these were identified by local tblastn search of data, using protein or nucleotide sequences of known identity downloaded from NCBI, miRBase (Kozomara & Griffiths-Jones, 2011) and HomeoDB (Zhong et al., 2008; Zhong & Holland, 2011). These were then reciprocally blasted against the nr or miRBase database to confirm identity, using the blastx (protein) or blastn (nucleotide) function. Where diagnostic sequence features exist, these were used as a further confirmation of identity.

2.2.9 Primer Design

Sequences identified by BLAST search and confirmed through phylogenetic analysis were entered into Primer3Plus software `http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi`. Prospective primer pairs were then entered into Amplify3 (University of Wisconsin, 2005) and chosen on the basis of lowest predicted primer dimerisation potential. These primers were then ordered from Thermo Electron GmbH (`www.thermo.com/biopolymers`).

2.2.10 Functional Annotation and KEGG Pathway Assignment

The dataset was automatically searched for homologues and annotated according to gene ontology (GO - http://www.geneontology.org) terms using Blast2GO 2.5.0 web start against the nr database (Conesa et al., 2005; Gotz et al., 2008). GO term distribution within the *D. melanogaster* genome was downloaded from B2GO-FAR (Gotz et al., 2011) and quantified (as with transcriptome data) using the Combined Graph function of Blast2GO. For assignment to KEGG pathways I uploaded my final dataset to the KAAS-KEGG Automatic Annotation Server (`http://www.genome.jp/tools/kaas/`) for processing using the single-directional best-hit (SBH) option, the default (60) blast bit score threshold and the hsa, mmu, xla, dre, cin, spu, dme, ame, cel, smm, nve and tad datasets.

2.2.11 Phylogenetics

To confirm the identity of sequences putatively identified by Blast (NCBI, 2011), phylogenetic analysis was performed, using sequences derived from our own transcriptomic and genomic sequencing efforts along with those of known identity derived from the NCBI protein database (`http://www.ncbi.nlm.nih.gov/protein/`) and other sources, as noted when used. Where necessary, protein sequence was derived from nucleotide data using Expasy (Gasteiger et al., 2003).

Multiple alignments of protein sequence were carried out using MAFFT with the L-INS-i setting or as noted (Katoh et al., 2002; Katoh & Toh, 2008). Where noted alignments were tidied using the Gblocks program, with all less stringent settings imposed, (Castresana, 2000) available at `http://www.phylogeny.fr/version2_cgi/one_task.cgi?task_type=gblocks`.

In general, trees shown are maximum likelihood amino acid trees made by the MEGA5 program, with 1000 bootstrap replicates and all default priors (Tamura et al., 2011) with the model used stated in figure legends, with model selected where necessary using ModelGenerator (Keane et al., 2006). Mr Bayes v.3.1.2 software (Huelsenbeck & Ronquist, 2001; Ronquist & Huelsenbeck, 2003) was used with amino acid substitution model, determined on a gene by gene basis and noted alongside trees. Putative phylogeny was reconstructed via Monte Carlo Markov Chain search, run over at least 1,000,000 generations, with trees sampled after every 1000 generations. The first 25% of trees were discarded as "burn in", and trees are displayed with posterior probability values at the root of each branch. Trees were annotated in MEGA5 or Dendroscope (Huson et al., 2007) and coloured labels and boxes added in Powerpoint for presentation.

Establishing Lophotrochozoan Sequence Resources

3.1 Establishing Lophotrochozoan Sequence Resources

As of 2010 few comprehensive sequence resources were publicly available within the Lophotrochozoa outside the Schistosoma, although some data was available ahead of publication with the understanding that no papers would be published using this data until the authors had established primacy. While these datasets were of general use to us, they neither represented species available to us nor allowed us full scope for experimentation.

To allow the most straightforward approach to identifying gene sequences within my species, and following painstaking, albeit successful, attempts to clone genes via degenerate PCR based approaches, I began sequencing a range of datasets, beginning with a mixed embryonic transcriptome of *P. lamarckii*. This was followed by the sequencing of genomic DNA from a single male *P. vulgata*, and thereafter by a range of analysis and collaborative work on a number of different lophotrochozoan datasets.

From this work I have been able to establish two draft genome sequences and three transcriptomic resources (along with some work on a previously existing transcriptome) which have formed the basis for a range of investigations, and are in the process of being prepared for publication in a range of guises. In this chapter I have summarised these findings, presenting the basic metrics for each resource and highlighting some of the analyses I have performed. These include

- The sequencing of an embryonic *P. lamarckii* transcriptome, its analysis and publication.
- Production of genomic data for *P. vulgata*, and investigations using this resource.
- Assembly of a *B. plicatilis* genome, and its use to derive basic information on the Nodal pathway in this species.
- Construction and analysis of transcriptomic resources for the molluscs *C. fornicata* and *B. glabrata*.

These resources are the basis of the work detailed in Chapters 4 and 5 of this thesis, and will be of use to a wider range of research, both in my laboratory and others around the globe.

3.2 *P. lamarckii* Transcriptomic Results

For my initial *P. lamarckii* transcriptomic investigations, an additive multiple-*k* assembly of Oases-derived assembly data was performed after initial comparison of assembler performance, combining the power of Oases in retaining splice form data at a given *k*-mer size with the benefits of multiple-*k* assembly. Multiple-*k* mer assemblies use assemblies derived from a variety of *k*-mer sizes, which are then combined together and any redundant sequence removed, allowing the recovery of more data from reads than single *k*-mer length assemblies.

As well as targeting members of the *Nodal* pathway, as summarised later in this thesis, for the purposes of initial publication I examined the coverage in this transcriptome of members of the homeodomain-containing superclass and forkhead box (Fox), Sry-related HMG box (Sox) and T-box (Tbx) classes. These groups of transcription factor are well described in the Ecdysozoa and deuterostomes (Zhong et al., 2008; Zhong & Holland, 2011; Kaestner et al., 2000; Jager et al., 2006; Shimeld et al., 2010b; Paps et al., 2012), thus facilitating categorisation, and perform a variety of roles in patterning and differentiating tissue in developing embryos (Carlsson & Mahlapuu, 2002; Papaioannou & Silver, 1998; Bowles et al., 2000). Together they

represent a major proportion of the larval transcription factor complement, and as such allowed us to estimate the utility of this method for developmental gene discovery, as well as providing a resource for future experimentation, as few of these genes have been described functionally in the Lophotrochozoa.

The majority of the work completed in regards to my *P. lamarckii* embryonic transcriptome investigations can be seen in Kenny & Shimeld (2012), which can be found beginning on Page 275 of this thesis. However, several other analyses were performed to optimise assembly and ascertain key facts relating to L/R asymmetry, and these are summarised here.

3.2.1 Sequencing Results and Quality Control

Basic statistics for Illumina reads used as the basis for this study can be seen in Table 3.1. The mean GC content of the reads (43.33%) is similar to an earlier value from *P. lamarckii* derived from EST data (42.42 % (Takahashi et al., 2009)).

Table 3.1: *Summary statistics pertaining to reads obtained on the Illumina GAIIx platform used in* P. lamarckii *transcriptome assembly.*

Platform	Illumina GAIIx
Paired-end Fragments Sequenced	72,916,154
Read Length (bp)	51
Insert Size (bp)	300
Amount of Data (Gigabytes)	3.719
GC content (%)	43.33

The quality of sequencing was assessed via the quality score metric provided with reads using FastQC, which found the median per-base sequence quality to be excellent, above the 39 mark from the start of reads through to the 51st base. A number of sequences were overrepresented in the first 11 positions of reads, These are likely to be a result of bias in Illumina hexamer binding as reported previously (Hansen et al., 2010). As no quantitative assessment of the data was attempted this did not affect my analysis in a measurable fashion.

3.2.2 Transcriptome Assembly Programs, Relative Performance

A variety of assembly programs were tested, and metrics relating to their comparative performance can be seen in Table 3.2. Assemblers were assessed using a

number of criteria in order to determine which would be best suited for performing multiple-*k* analysis. Oases was chosen for further use, due to its robust performance in almost all metrics, particularly in the number of bases incorporated into the assembly and in the high mean contig length.

Table 3.2: *Summary of the metrics of assemblies gained from each of five assembly programs trialled. All results presented at a* k*-mer size of 27 with the exception of Trinity (*k*-mer = 25).*

Assembler:	ABySS	Oases	SOAPdenovo	Trinity	Velvet
Number of contigs	33,989	91,542	118,207	153,548	116,928
Max contig length (bp)	7,741	8,309	4,154	8,346	3,583
Mean contig length (bp)	475.89	374.54	194.7	227.22	191.42
Median contig length (bp)	288	174	144	143	141
N50 contig length (bp)	794	757	199	240	196
# contigs in N50	5,596	11,679	32,035	30,050	31,724
# contigs 1kb $\leq$	3,969	8,178	816	3,829	678
# bases, total	16,175,159	34,286,127	23,014,749	34,888,902	22,382,612
# bases in contigs 1kb $\leq$	6,640,794	14,094,850	1,110,845	6,359,792	900,520
GC Content %	38.17	42.18	41.97	41.98	42.12

There was a fivefold difference in the number of contigs produced by assemblers, although the total number is expected to decrease when contig length rises as shorter contigs are joined together. The distribution of these can be seen in Fig. 3.1, and raw data figures seen in Table 3.2. A preponderance of short contigs relative to other assemblers output may suggest that an assembler is not making the best possible use of data. Shorter contigs are also less likely to span conserved domains and thus be recognized by blastx for annotation.

Fig. 3.1 shows the distribution of the numbers of contigs by length for each assembler trialled in the present study, along with the distribution of contig length in the final, additive multiple-*k* assembly. Of the assemblers, it can be seen that ABySS produces fewer contigs of short read length while the vast majority of the contigs produced by SOAPdenovo, Velvet and Trinity are less than 200 bp long. ABySS produced the smallest quantity of contigs by a considerable margin, while Velvet, Trinity and SOAPdenovo produce the largest number, each in excess of 100,000. Oases also produces a large number of small contigs, but unlike SOAPdenovo and Velvet, recovers a large number of contigs larger than 1000 bp in length. It therefore seems Oases is the best-performing assembler for recovering the maximum possible data from reads, efficiently assembling both short and long contigs, and perhaps recognising the potential diversity of splice variants and polymorphism in

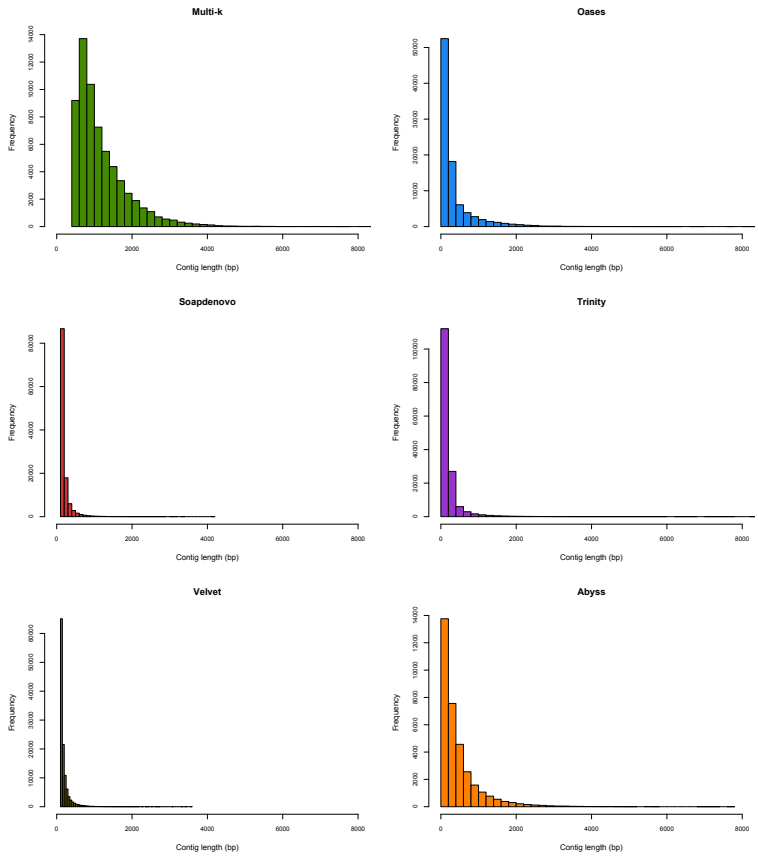


Figure 3.1: *Comparison of Multiple Assembly Contig Build Statistics. Histograms showing distribution of contig numbers according to length for the five assembly programs tested, as well as that of the final additive multiple-*k *build. Statistics for individual assemblers shown taken from a* k*-mer size of 27 (Trinity* k*-mer = 25).*

this dataset.

The number of longer contigs (1kb $\leq$) assembled provides another important metric for assessing build quality. Oases, with 8,178, clearly outperforms other assemblers in this regard, with ABySS and Trinity the next best performing, providing 3,969 and 3,829 contigs of this length respectively. N50 metrics also provide an understanding of the assembly of long contigs, and can be seen in Table 3.2.

Oases combines larger contigs with a contribution from large numbers of smaller contigs, as previously seen in Fig. 3.1 while the ABySS assembly is skewed towards long contigs only, which suggests it is not best able to assemble shorter contigs. Trinity, Velvet and SOAPdenovo assemblies are almost entirely made up of short contigs - the N50 for these assemblers is below 250 bp. Oases again seems to be the best suited candidate for an additive multiple-*k* approach, reporting a far greater number of long contigs than other assemblers.

3.2.3 Construction of Additive Multiple-*k* Dataset

Oases was run sequentially for *k*-mer sizes from 21 to 29 bp. For statistics on the results from each iterative assembly, please refer to Kenny & Shimeld (2012), which can be found beginning on Page 275 of this thesis. In general, lower *k*-mer sizes resulted in longer contig assemblies, a somewhat surprising finding, although I anticipated the rise in contig number as lower *k*-mer size is correlated with increased efficiency in recovering lowly expressed transcripts (Zerbino & Birney, 2008; Surget-Groba & Montoya-Burgos, 2010). This resulted in a total of 486,589 contigs before the removal of redundancy, as can be seen in Table 3.3.

After removal of redundancy with V-match, the additive multiple-*k* assembly was trimmed to consist of contigs 500 bp and longer for further analysis, as the large number (242,641) of contigs longer than 100 bp proved unworkable for Blast2GO analysis and functional annotation. Table 3.3 shows the metrics of the additive multiple-*k* dataset after TGICL clustering and trimming to 500 bp.

Table 3.3: *Multi* k-*mer Oases assembly metrics. Summary of metrics relating to the Oases assemblies, after merger of all assembles k-mer size 21-29, before and after the removal of redundancy and imposition of a 500 bp minimum length.*

Multi *k*-mer build:	Before Redundancy Removal	After Removal of Redundancy
Number of contigs	486,589	50,151
Max contig length (bp)	9,325	9,325
Mean contig length (bp)	403.23	1221.54
Median contig length (bp)	185	985
N50 contig length (bp)	822	1,419
# contigs in N50	63,154	14,219
# contigs 1kb $\leq$	48,848	24,562
# bases, total	196,206,027	61,261,605
# bases in contigs 1kb $\leq$	85,137,886	42,926,898
GC Content %	42.01	43.87

3.2.4 Determining the Utility of the Additive Multiple-*k* Approach

The GC content of the final assembled transcriptome closely mirrors that of reads, at 43.87% and 43.33% respectively. This agrees well with the results of earlier EST screening in *P. lamarckii* (Takahashi et al., 2009), which suggested a GC content of 42.42%. Statistics on final transcriptome assembly can be seen in Table 3.3 above.

BlastX was used to compare the contig set against the human proteome. 27,693 contigs had hits to the human proteome with e $\leq$ 1.0e-9, however this value fell to 6,540 hits within the human proteome when multiple best hits from different *P. lamarckii* contigs to the same human target were counted only the first time they occurred. This suggests remaining redundancy in the *P. lamarckii* contig set, likely due to polymorphism present within my dataset, and splice variation reported by the Oases assembly. This could also reflect assembly errors, however genuine redundancy is expected. We specifically selected the Oases assembler as it should maintain multiple splice variants for single genes, and single errors in reads should be excluded from the dataset as a result of the coverage cutoff function, which discards *k*-mers which occur less than a set threshold number of times before performing contig assembly, discarding likely sources of sequencing error.

There was also further variation at the level of single nucleotides (data not shown). These too were expected as the original mRNA sample was derived from multiple individuals from an outbred population, and they likely reflect allelic variants. Sedentary or sessile broadcast marine spawners with planktonic larvae may have very large effective population sizes and maintain high levels of genetic poly-

morphism, as seen for example with *Branchiostoma floridae* and *Ciona savignyi* (Small et al., 2007; Putnam et al., 2008). The multiple variants in my dataset represent true variation in the population rather than individual sequencing errors, as the coverage cutoff utilized (threefold recovery of a k-mer sequence) and the expected error rate (less than 1 in 10,000) suggests that the likelihood of the same error being found concurrently at a single site in three k-mers and incorporated into contig assembly is very low. My dataset thus will provide a valuable resource for population genetic studies of *P. lamarckii*.

One caveat must be applied to the multiple-*k* approach. It can quickly and efficiently produce large amount of contig sequence. Analysis tools, have, however, lagged behind my ability to produce data, and I was forced to limit subsequent annotation analysis to reads greater than 500 bp in length, resulting in nearly 80% of contigs being excluded from Blast2GO analysis. For the transcription factor dataset discussed below, those contigs 100 bp - 500 bp were included in analysis, although only a handful of genes were recovered from these. It is believed that trimming to 500 bp and above, while necessary for Blast2GO analysis and functional annotation, removed from consideration a number of low-expressed genes whose transcripts had not assembled to large contig sizes. These 100 bp - 500 bp contigs were therefore incorporated into the gene discovery portion of this work.

The additive multiple-*k* approach therefore, while an excellent source of sequence data, forces decisions to be made as to the limits of analytical capability. It is particularly good at ensuring that the largest possible contig builds are identified and retained, and for identifying isoforms of contigs, whether SNPs, splice variants or allelic forms.

3.2.5 Functional Annotation and Analysis

Blast2GO v. 2.5.0 (Conesa et al., 2005; Gotz et al., 2008) was used to perform functional analysis, firstly by performing blastx against the NCBI nr protein database with all default prior settings, before subsequent gene ontology (GO)-mapping and annotation. Of 50,151 contigs, 34,846 (69.4%) returned a blastx hit above the cut off score (1e-3). Top hit (blastx) distribution by species can be seen in Fig. 3.2. These

were distributed among a range of species, with invertebrate deuterostomes being the most common top hits, which may reflect a shared tendency to less-derived gene complements in these species, which are also noted as slow-evolving (Putnam et al., 2008; Arendt et al., 2008). This result may also reflect the present paucity of Lophotrochozoan sequences within the nr database.

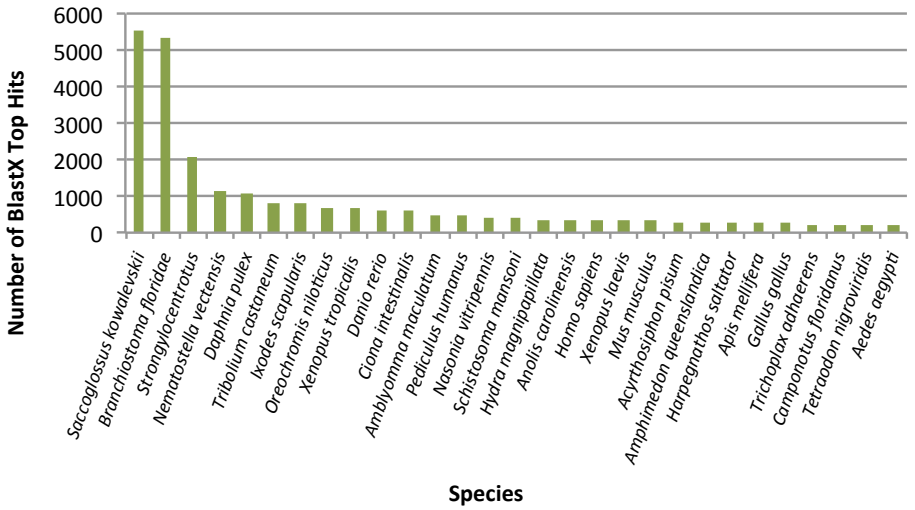


Figure 3.2: *By species distribution of top blastx hits from transcriptome to the NCBI nr database. Distribution of the species source of best hits found for each contig by blastx against the NCBI nr database using Blast2GO (Conesa et al. 2005).*

Of 50,151 contigs assessed, 17,967 (35.8%) were assigned into a wide variety of GO groups at all levels. The distribution of sequence GO assignations into second-level functional categories, divided between the three top level divisions (biological process, cellular component and molecular function) of GO, can be seen in Fig. 3.3. These are presented alongside the results for the proteome of *Drosophila melanogaster*, as downloaded from B2GO-FAR (Gotz et al., 2011), used as a very well annotated protostome dataset for comparison as no Lophotrochozoan equivalent is yet available.

Differences between GO annotation of a transcriptome and a complete predicted proteome are to be expected, and could derive from species differences, from differences in expression associated with different developmental stages, and from the transcriptome assembly process itself. In many GO categories, contigs were slightly underrepresented compared to the *D. melanogaster* sample, which is per-

haps the result of the limited temporal sampling (3 larval stages in this transcriptome, as opposed to the complete proteome) represented by the present dataset. Significant underrepresentation (assessed by Fishers Exact Test with multiple testing correction of FDR (Benjamini and Hochberg, 1995)) in the GO categories reproduction, multi-organism process and locomotion (p values of 0, 1.40e-154 and 3.00e-208 respectively) were unsurprising, but underrepresentation in the terms such as growth, developmental process, cellular component organization and cell proliferation (p values 8.10e-223, 0, 0 and 7.80e-21) were not expected, and may be a result of genes assigned to these areas being weakly transcribed or when assembled being shorter than 500 bp, and as such not being represented in my annotated dataset.

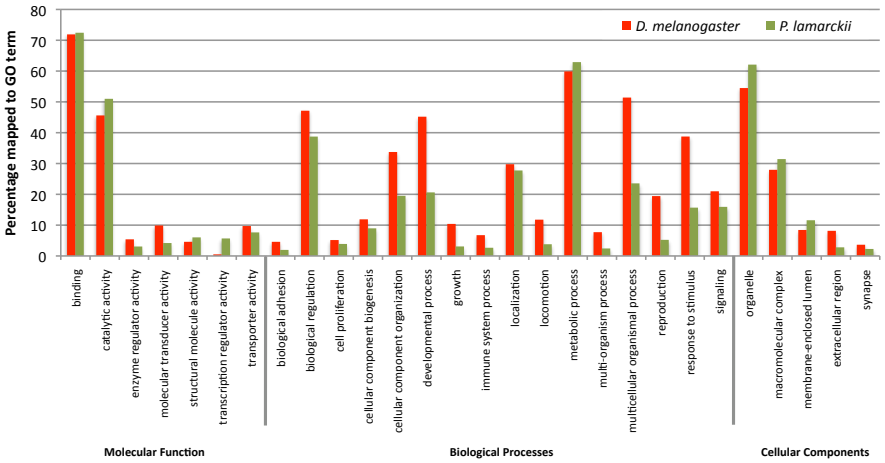


Figure 3.3: *GO Assignation Distribution* P. lamarckii *transcriptome (green) and* D. melanogaster *(red) assignation distribution at the second level of GO mapping is presented as the percentage of each dataset that falls within a given term. The* P. lamarckii *dataset was made up of 17967 contigs assigned GO terms, and the* D. melanogaster *dataset was downloaded from B2GO-FAR (Gotz et al. 2011).*

The most notable overrepresentation in the *P. lamarckii* dataset is in the proportion of genes with the GO term transcription regulator activity (p value 3.20e-04). This assignation is eleven times more common in this dataset than in the *D. melanogaster* proteome as a whole, significantly over-represented, and may indicate a large amount of differentially activated transcriptional activity is occurring in larvae at these stages of their development, a finding that would concur with what is understood of their biology. Other second-level GO terms also overrepresented

in this dataset relative to the *D. melanogaster* proteome include the terms structural molecule activity and catalytic activity (p values 9.50e-7 and 5.30e-05), perhaps an indication of the processes of differentiation occurring within the developing larvae. Alternatively, the Oases multi-*k* approach could be recovering a multitude of splice forms for transcription factors, which, when compared to a proteome would appear to be over-represented. This was not observed empirically, but has not been rigorously tested.

3.2.6 *Nodal* Pathway Recovery

Initial investigations from this transcriptomic data focussed on *Nodal* pathway components for cloning and RNA *in situ* hybridisation work, and seemed to indicate that, while many core signalling components are present, many regulatory mechanisms commonly found in deuterostomes seemed to be lacking from the *P. lamarckii* transcriptomic data as seen in Fig 3.4. The conservation of these components across the Metazoa is discussed in more detail in Chapter 5 of this report, but the presence of so many Nodal signalling components was of much utility for ongoing molecular investigations.

The relative absence of regulatory componentry is possibly a result of the time points sampled in my dataset, the earliest of which was 24 hours post fertilisation. At this point L/R asymmetry is relatively well established, and regulatory proteins (such as the Dan family) may already have performed their roles. It is also possible that these proteins are expressed only at a very low level or in a small subset of cells - *in situ* hybridisation or quantitative PCR experiments could provide the answer to this if desired.

3.2.7 Homeodomain-Containing Contigs

From the transcriptome I was able to identify twenty-eight contigs whose sequence entirely spanned the homeodomain, and a further nine whose sequence was recovered sufficiently for satisfactory phylogenetic analysis. The homeodomain superclass of genes is well described, and performs a variety of crucial roles in develop-

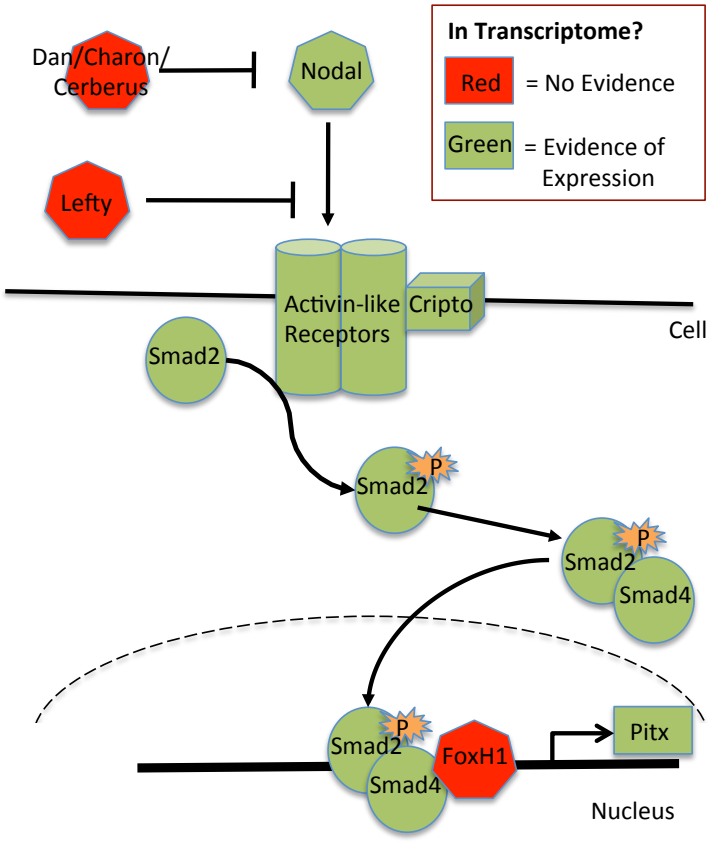


Figure 3.4: *Evidence for elements of the core* Nodal *pathway in the* P. lamarckii *transcriptome. Green represents evidence of presence, Red no evidence of presence found. For evidence of presence of Nodal, Activin-like Receptors, Smads, Cripto and Pitx please refer to Figs. 5.4, 5.9, 5.10, 5.11 and Fig 4.10 respectively.*

mental patterning across the Metazoa (Zhong et al., 2008; Zhong & Holland, 2011). The homeodomain genes identified from the transcriptome dataset can be seen in Table 3.4, as noted from phylogenetic analysis and examination of signature residue sites. I also note the presence of a number of other homeodomain-containing proteins whose sequence did not overlap sufficiently for firm annotation.

I note an apparent paucity of HoxL class genes within this dataset, and suspect that this is a result of the timings of samples taken to create the RNA pool, a finding confirmed by the expression patterns of these genes in other lophotrochozoans (Zhang et al., 2012a). Other subclasses, particularly the NKL and PRD classes, are well represented.

Two homeodomain genes heretofore undescribed in lophotrochozoan species were found in the course of analysis, homologues of the *Dmbx* and *Hnf (Hmbox)* transcription factors. For further confirmation of the identity of these sequences, please refer to Kenny & Shimeld (2012), which can be found beginning on Page 275 of this thesis. In deuterostomes it is believed that *Dmbx* plays an important and pos-

Whole Homeodomain Recovered:				**Partial Homeodomain:**	
ANTP/HoxL		**Prd**		**Prd**	
Cad		Dmbx		Msx	
Mnx		Gsc		Nk 6 b	
Evx		Hbn		Abox-like	
ANTP/NKL		Otp		**Cut**	
Bsx		Otx		Cmp HD1	
Dlx a		Paired-like		**Tale**	
Hhex		Pitx		Pknox	
NK 2.1 a		Vsx		Tgif	
Nk 2.1 b		**Lim**		**Pou**	
Nk 2.2		Lhx 1/5		Pou2	
Cut		Lhx 2/9		Pou 3	
Cmp HD2		Lhx 6/8		**Lim**	
Cux		Ist/Tup		Lmx	
Onecut		**Six**			
Tale		Six 3/6			
Irx		**Hnf**			
Pbx		Hmbox-like			
Column colour:					
Colour:		Identity confirmed by diagnostic conserved domain or specific amino acid signature			
		Identity unable to be confirmed as transcriptome sequence does not cover area			
		Diagnostic conserved domain or specific amino acid signature not present			

Table 3.4: *A portion of the* P. lamarckii *homeodomain-containing gene complement, as inferred from the present transcriptome study. Coloured boxes note further confirmation of identity, when possible from signature amino acid residues (data not shown, refer to Kenny & Shimeld (2012)).*

sibly deuterostome-specific role in establishing the midbrain/hindbrain boundary, and the potential conservation of this in the Annelida would be a good candidate for future experimentation (Takahashi & Holland, 2004).

The transcriptional repressor *Hmbox* was traditionally believed to be chordate-specific (Takatori et al., 2008) and has only recently been noted in an ecdysozoan (Lesch & Bargmann, 2010). Hmbox identity was confirmed here by examination of clearly conserved features of this class, including an extended homeodomain courtesy of additional amino acids between the h2 and h3 domains (data not shown), underlining the conservation of this gene in the Bilateria. The existence of *Cmp* homologue sequence supports the hypothesis that this gene is shared ancestrally within the Bilateria, with *Satb* (which is found only in vertebrates, but appears similar to *Cmp*) likely representing a divergent vertebrate-specific form of this gene (Burglin & Cassata, 2002).

Some features of homeobox genes generally observed in ecdysozoan and deuterostome datasets do not seem as prevalent in the sequences obtained by this transcriptomic analysis. Tinman (Tn) domains, which are generally observed in NKL family homeobox genes, were not observed in a homologous position in many of the *P. lamarckii* NKL genes identified. These may lie in regions as yet unsequenced, but their presence could not be used as a diagnostic feature to con-

firm the identity of a portion of my data. This is also the case for the octapeptide domains generally found in PRD-class homeobox genes, although the latter have other characteristic regions used to confirm identity.

The recovery of sequence for *Dlxa* and *Dlxb*, identical to that found in McDougall et al. (2011) gives confidence in the validity of the multi-*k* approach utilised. While the sequence recovered for *Dlxb* does not span the entirety of the homeodomain, the sequence recovered matches that known for this gene in *P. lamarckii*. I also recovered a sequence which appears similar to the Pax β sequences first identified in the leech *H. robusta* (Schmerer et al., 2009).

3.2.8 Fox, Sox and T-box Containing Transcripts Recovered

Fox proteins are members of the helix-turn-helix class of proteins, with an 80 to 100 amino acid "Forkhead Box" motif that defines the class. They are separated into 23 families (FoxA - FoxS) and perform a variety of roles in specification of tissues during embryonic development, as well as regulating a number of metabolic functions. Sox genes belong to the HMG box superclass, and regulate many diverse elements of development, growth and differentiation. They are found throughout the Metazoa and their phylogeny is generally well understood. The T-box genes are less numerous than those families detailed above, but nonetheless are vital for a range of developmental roles They are defined by the presence of a roughly 200 amino acid-long region known as the T-box domain, first identified in Brachyury. The names of the *P.lamarckii* genes firmly identified within the dataset can be seen in Table 3.5.

Transcription Factors:		
Fox Family:	**Fox Family (Ctd):**	**T-box Family:**
FoxAa	*FoxL1*	*T-Brain*
FoxAb	*FoxN1/4*	*Brachyury*
FoxC	*FoxN2/3*	*Tbx 2/3*
FoxG	*FoxO*	**Sox Family:**
FoxI	*FoxP*	*Sox B1*
FoxJ1	*FoxQ2a*	*SoxB2*
FoxJ2/3	*FoxQ2b*	*Sox C*
FoxK	*Fox (Unknown Family)*	*Sox D*

Table 3.5: *The portion of the* P. lamarckii *Fox, Sox and T-box gene complement derived from the present transcriptome study.*

The Fox gene complement is particularly well represented in the transcriptome dataset, and can be seen in Fig. 3.5, alongside the Forkhead genes of *H. sapiens, D. melanogaster* and those known from *C. teleta* (Shimeld et al., 2010a). The presence of two *Fox A* and *Fox Q2* sequences may represent recent duplications within these families in this species, while *Fox B, D, H* and *M* genes are absent, and also absent from annelid datasets as represented on Genbank.

My search for Sox transcription factors uncovered four of the six known invertebrate Sox families, as can be seen in Fig. 3.6, using the well-understood *Takifugu rubripes* Sox genes (Koopman et al., 2004), and those of *P. dumerilii* (Kerner et al., 2009) and the purple sea urchin, *S. purpuratus* (Sodergren et al., 2006) to create a phylogeny. The chordate-specific families (Groups A, G and H) were not present, and this transcriptome has therefore uncovered two thirds of the expected Sox complement, with only Groups E and F absent in the dataset. A monophyletic SoxB1 group could not be recovered. Confirming my assignment, however, the SoxB1 sequence can be recognised as the SoxB1 orthologue by blast, and possesses the characteristic arginine amino acid residue at position 2 and threonine at position 78, while the SoxB2 sequence possesses the diagnostic proline at this latter position.

Three T-box family genes were uncovered, of the approximately seven/eight broad subfamilies of T-box genes identified by Papaioannou & Silver (1998), and can be seen in Fig. 3.7, alongside the *D. melanogaster* and *M. musculus* complements, and those identified in a number of Cnidarian and Poriferan species. The presence of *Brachyury* was expected, given its conserved role in the Bilateria (Lartillot et al., 2002a). More surprising, however, was the presence of a clear and well-supported homologue of *T-brain*, which has been described extensively only in deuterostomes, and is represented outside this superphylum in the NCBI database by one unpublished sequence (*Hydroides elegans*, Accession ACA48210). The expression of this gene in the hemichordate apical organ and in the vertebrate forebrain makes it an interesting target for future consideration with regards to conservation of expression and role (Tagawa et al., 2000).

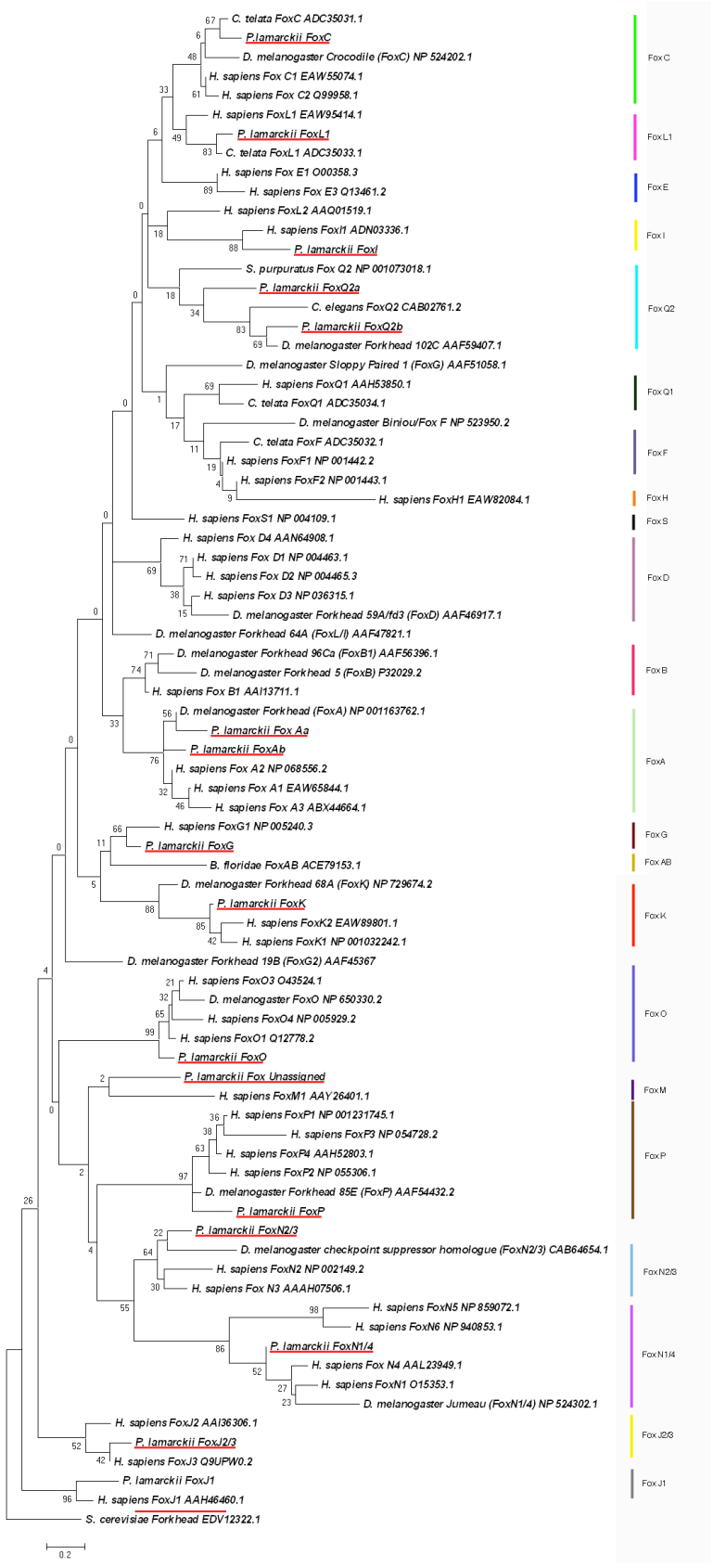


Figure 3.5: *Maximum likelihood Fox gene phylogeny computed from aligned Forkhead Box domains (45 aa long after gap removal) using Mega5 under the JTT model with all other default settings. Sequences from* P. lamarckii *underlined in red. Bootstrap proportions (from 500 replicates) are found at the foot of each node. Tree rooted with* Saccharomyces cerevisiae *Forkhead (EDV12322.1). The scale bar represents 0.2 substitutions per amino acid at the given distance along branches on tree. Previously described Fox families, as defined by (Shimeld et al. 2010b) are shown on the right of the figure. Alignment can be seen in Appendix C, Fig. 1*

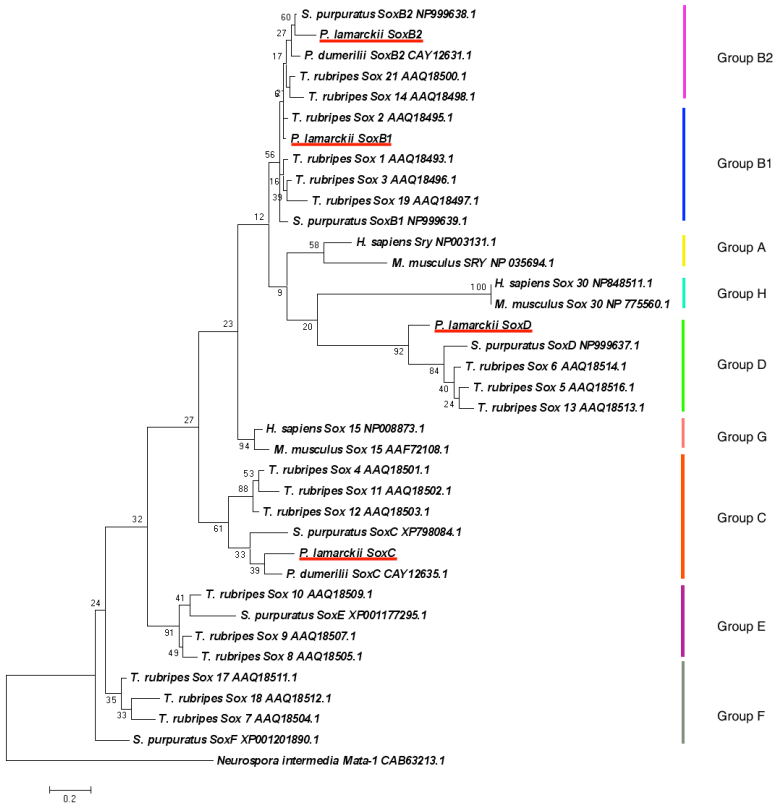


Figure 3.6: *Phylogeny of* P. lamarckii, T. rubripes, P. dumerilii *and* S. purpuratus *Sox genes, computed from area surrounding the HMG domain (73 amino acid alignment). All annotated* P. lamarckii *Sox genes are found in well-supported monophyletic groups with genes of known annotation, with the exception of the* SoxB1 *homologue. This maximum likelihood tree was constructed utilizing Mega5 with the JTT model and all other default priors, along with 500 bootstrap replicates, the proportions of which can be seen at the base of each node.* P. lamarckii *Sox genes are underlined in red. Tree is rooted with* Neurospora intermedia *Mata-1 (CAB63213.1). Scale bar represents 0.2 substitutions per amino acid at the given branch length. Previously defined Sox gene families (Jager et al., 2006) are shown on the right of the figure. Alignment can be seen in Appendix C, Fig. 2*

3.2.9 Estimation of Coverage

No firm conclusion can be drawn as to the coverage provided by this transcriptome until the advent of a complete *P. lamarckii* genome, as it can be difficult to estimate the coverage provided by transcriptomic studies due to variation in gene expression levels and temporal expression profiles. However, by determining the proportion of genes found in comparison with known housekeeping datasets I can gain an understanding of how representative my transcriptome is. KEGG mapping (Moriya et al., 2007) for example, has revealed approximately 70% coverage across a range of key conserved metabolic pathways, with coverage in some cases being

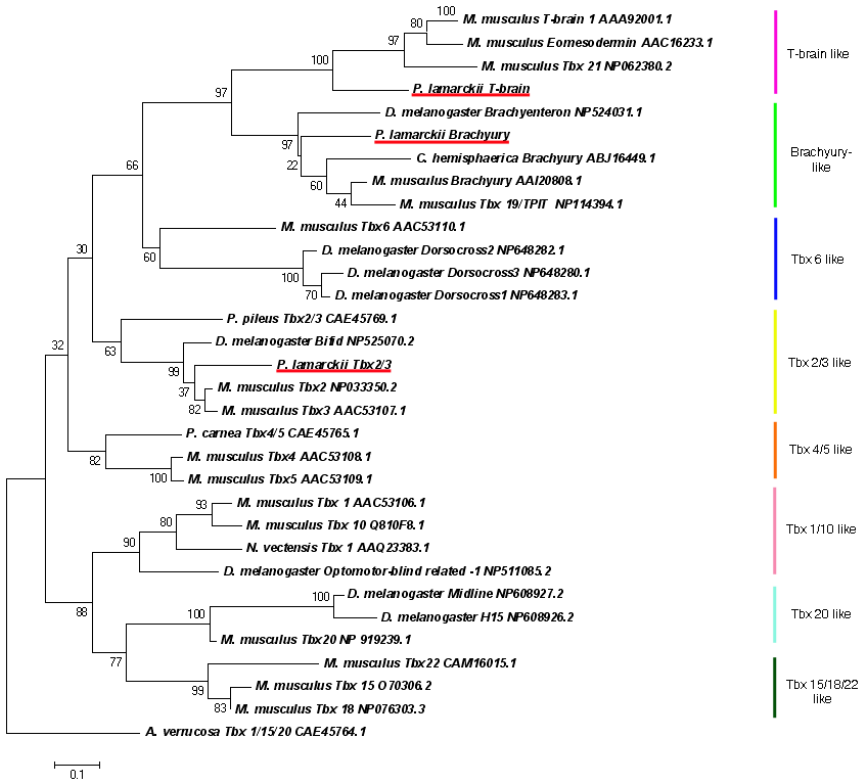


Figure 3.7: *Maximum likelihood phylogeny of T-box genes from* M. musculus, D. melanogaster, P. lamarckii *and a number of Cnidarian and Poriferan species derived from T-domain alignment.* P. lamarckii *sequences are underlined in red. Phylogeny computed in Mega5 using the JTT model and all other default priors. Tree rooted with* Axinella verrucosa *Tbx1/15/20 protein (CAE45764.1). The scale bar represents 0.1 substitutions per amino acid at the given distance along branches. Previously defined T-box gene families (Papaioannou & Silver, 1998) are shown on the right of the figure. Alignment can be seen in Appendix C, Fig. 3*

even higher, particularly as many KEGG pathways contain genes found only in a subset of species,

Of transcription factor families examined, approximately 60% of the expected complement can be reliably identified within my dataset based on comparison to DNA binding domains, while more ambiguous assignments can be made for those transcription factors whose sequence was recovered away from these.

The assignation by Blast2GO of my dataset into 34,846 blastx hits, split into 17,967 GO categories at all GO levels also suggests I have uncovered a substantial proportion of the expressed transcriptome. Together with the large number of unique blastx hits recovered by my assemblies I suspect that at least 60-70% of the 24 72 hpf trochophore transcript complement is represented in my additive

multiple-*k* assembly.

Using the Oases assembler in building an additive multiple-*k* assembly uncovered the sequence of a range of developmentally important genes. Other assembly programs also represent potentially useful tools for transcriptome assembly, with their own advantages and disadvantages. The additive multiple *k* assembly technique also holds much promise, with a variety of programs capable of removing redundancy (including TGICL, CD-HIT EST and Vmatch) now available. It increased the number of identifiable contigs, as well as their length, relative to individual *k*-mer size builds, although the amount of data produced proved more difficult to analyse using Blast2GO and some other tools, necessitating a higher size cut-off value (500 bp).

The large number of transcription factors discovered presents many opportunities for future studies. The identification of annelid homologues of *Dmbx* and *T-Brain*, as well as a possible *Hnf* representative, raise questions about the conservation of these genes in the Lophotrochozoa, and provide opportunities to test their ancestral role. The phylogenetic trees reconstructed in this study, along with the blastx top hit results found through Blast2GO analysis, reinforce the hypothesis that polychaete annelids are slow-evolving compared to other taxa.

3.2.10 Transcriptome: Conclusions

In terms of my broader research, the construction of this transcriptome provided a raft of sequence data that has already been utilised in a range of contexts. The sequence of the *P. lamarckii* homologue of *Nodal*, as well as a number of other genes of interest for the examination of the establishment of L/R asymmetry, have all been derived from this dataset, and cloned for use in RNA *in situ* hybridisation and other work. The entirety of the cloned *Nodal* componentry from this species has now been handed on to another student to perform, and it is hoped that this dataset will be of wider utility well into the future.

3.3 *P. vulgata* Genomic Work

The sequencing of a *P. vulgata* genome has allowed us to explore a range of fascinating questions relating to the establishment of L/R asymmetry, as well as opening the doors to a range of investigations in other areas. A sperm sample from a single male *P. vulgata* was used for a lane of Illumina sequencing at two insert lengths, a strategy that worked well for assembly to a draft standard. While my research on the analysis of the *P. vulgata* genome seems to indicate that coverage of *P. vulgata* coding sequence is fairly complete, much work remains to be done in scaffolding these to a draft genome standard if this data was to be used for assessing syntenic relationships. However, this resource has already proven its worth for my ongoing investigations, as detailed later in this thesis.

3.3.1 Sequencing Results and Quality Control

Basic statistics for the Illumina reads used as the basis of this investigation can be seen in Table 3.6. Two DNA fragment libraries, 200 bp and 500 bp in length, were constructed for sequencing on an Illumina HiSeq2000 platform . The mean GC content of the reads (39 and 38% respectively) is similar to that gained from an RNA sample sequenced previously by the Shimeld lab, at 39% (Werner et al., 2012).

Table 3.6: *Summary statistics pertaining to reads used in assembly of the* P. vulgata *genome.*

Library Insert Size:	200 bp	500 bp
Paired-end Fragments Sequenced	118,329,948	69,107,905
Read Length (bp)	100	100
Amount of Data (Gb)	65.3	38.1
GC content (%)	39	38

FastQC analysis revealed excellent data quality, with per base sequence phred score averaging 38 for both datasets, and the median by position at the 'excellent' level (30≤) through to the 100th nucleotide in each case. While initial trials at read cleaning were conducted (using Sickle and Musket software (Liu et al., 2013b), data not shown), no additional benefit was derived from such efforts by any standard assembly metric, and the results these efforts were not used for further analysis.

3.3.2 *P. vulgata* Genome Assembly

Assembly statistics for the best assembly trials for the three assemblers used can be seen in Table 3.7. All assemblers performed consistently better at relatively long *k*-mer sizes, reaching a maximum recovery of bases and longest average N50 at around 61 *k* (SOAPdenovo, 63). Velvet performed poorly in comparison to its rival assemblers. In general, SOAPdenovo performed more adequately in scaffolding contigs to longer lengths when compared to ABySS, but while it assembled the longest contigs well, shorter contigs were not as well-assembled, contained a high proportion of unresolved runs of 'N', and were of shorter size when measured by statistics relating to N50 and mean/median contig length. For consideration of synteny at a given locus, SOAPdenovo may be the best initial option, but more consistent recovery of longer contigs could be gained from ABySS.

For genome assembly, particularly for the purposes intended in the present project, gene discovery and the recovery of non-coding sequence surrounding specific genes was of more importance than recovering the absolute maximum amount of sequence data. The best possible ABySS build for a) gene discovery and b) scaffold length was optimized by alteration of parameters and repeated trial, with slight improvements to results, and is shown here in column one of Table 3.7. Considerable variability in the GC content of assemblies was observed, with none of the assemblies matching the GC% observed of reads, noted earlier in Table 3.6. The cause of this is unknown, but may relate to the differential ability of the assemblers used in coping with repetitive sequence.

Table 3.7: *Summary of assembly metrics from the best* P. vulgata *genome assemblies performed with Velvet, SOAPdenovo and ABySS, at* k*-mer lengths 61, 63 and 61 respectively. A minimum scaffold length of 300 bp was imposed before the generation of these statistics.*

Assembler (*k*-mer):	ABySS (61)	Velvet (61)	SOAPdenovo (63)
Number of scaffolds	310,316	480,207	527,149
Min scaffold length (bp)	300	300	300
Max contig length (bp)	49,465	38,259	83,238
Mean contig length (bp)	1957.30	1136.34	1274.73
Median contig length (bp)	1,215	737	703
N50 contig length (bp)	3,181	1,636	1,911
# contigs in N50	53,289	94,198	75,683
# contigs 1kb $\leq$	178,057	177,983	177,191
# bases, total	607,381,932	545,676,923	671,970,408
# bases in contigs 1kb $\leq$	531,182,537	379,789,914	473,205,776
GC Content %	35.39	35.76	32.09

From this 61-*k* ABySS assembly, coverage was assayed by examination of the Hox, Fox, Sox and T-box complements, along with comparison with extant *P. vulgata* resources. Preliminary results indicate that coverage of coding regions of the genome is exceptional, however, assembly into larger scaffolds is poor as can be noted from the results above, perhaps due to incomplete assembly across repetitive regions of the genome. This lack of resolution is probably due to the short (200 and 500 bp) paired end read fragment sizes which were used for assembly, which would struggle to span regions of low information in the *P. vulgata* genome. The creation of larger mate pair libraries for sequencing or the use of technologies with long read lengths could solve this problem if desired in the future.

3.3.3 Coverage

The genome size of *P. vulgata* has yet to be determined, so direct comparison of my dataset to an external measure is not currently possible. According to the Animal Genome Size Database (`http://www.genomesize.com/`) the C-value (in pg) of the ten species of Patellogastropoda thus examined varies between 0.43 (*L. gigantea*) and 0.94 (*Acmaea mitra*). Some gastropod molluscs possess C-values in excess of three times the latter value, although these appear to be outliers, with the general mollusc complement tending to be a C-value of around 1.

We can also estimate the genome size of *P. vulgata* using my reads, leveraging the fact that, theoretically, every *k*-mer is equally likely to be represented in my data. At any given *k*-mer size, by dividing the total number of reads within my sequencing data by the mean coverage of this *k*-mer size in my genome, then multiplying by the *k*-mer size, I can estimate the haploid genome size to be between 960 Mbp and 1.1 Gbp. This is in line with the general mollusc genome size, and would mean that my genome builds are recovering a little over half of the complete genome sequence. If highly repetitive areas, which can take up the sizeable majority of a genome, are discounted, this may mean that I have recovered almost all non-repetitive sequence in my dataset. The veracity of this, of course, will require a full genomic sequence to test.

As a proxy measure for coverage, I used a mantle-derived *P. vulgata* transcrip-

tome, which has recently been published by members of the Shimeld laboratory (Werner et al., 2012), as the basis for comparison. The full transcript complement of an Oases build of this data was compared to my genome by blastn, with a expected cut off value of 10^{-6} imposed. Of 127,055 transcripts in the Oases transcriptome, 124,768 hit a scaffold in my genome with an e value of less than 10^{-6}, 53459 with 100% similarity and 58388 with an e value of zero. This represents, at best, a 98.2% recovery rate of transcriptome data.

I am therefore confident that my sequencing and assembly efforts have recovered the lion's share of coding sequence within the *P. vulgata* genome, although repetitive intergenic sequence is poorly assembled. We can therefore expect this dataset to contain the majority of genes useful for my investigations into L/R asymmetry, and it to reconstruct gene families, although true absence of any particular gene from the *P. vulgata* genome is still difficult to assert at this stage.

3.3.4 *Nodal* Pathway Recovery

For the purposes of investigations detailed later in this thesis, the recovery of Nodal pathway genes was of paramount importance. The genomic resource allowed us to access the sequence of a number of key genes (particularly regulatory components such as *Cripto*) which were absent from previous sequencing efforts in this organism. The componentry found and firmly identified in the *P. vulgata* genome can be seen in Fig. 3.8. We recovered the majority of the pathway as described in deuterostome models, with some interesting exceptions.

In Fig. 3.8 it is noted that I was unable to identify a *Lefty*-like gene within my dataset. This absence, in concert with mounting evidence for the complete absence of Lefty-like ligands from other Lophotrochozoan genomes, suggests that the regulation of Nodal protein activity by competition with the faster-diffusing Lefty ligand in early embryos of some deuterostomes is a novelty in that clade. How this regulation is accomplished, and if it is in fact necessary in Lophotrochozoans, is yet to be established.

Most tantalising, perhaps, was the discovery of a potential *FoxH* ligand in the *P. vulgata* genome dataset. Along with evidence from *L. gigantia*, this would suggest

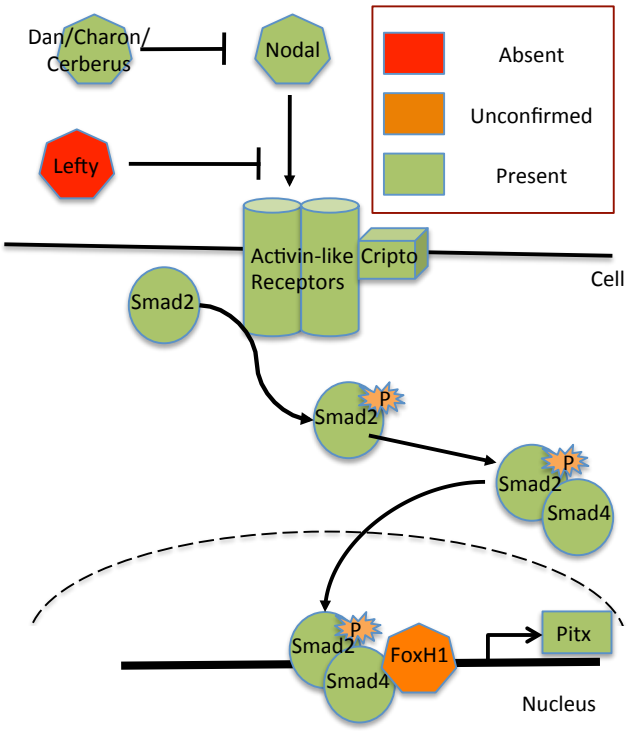


Figure 3.8: *Evidence for elements of the core* Nodal *pathway in the* P. vulgata *genome. Green represents evidence of presence, Orange conflicted evidence requiring focussed investigation, Red no evidence of presence found. For evidence of presence of Nodal, Activin-like Receptors, Smads, Cripto, Dans and Pitx please refer to Figs. 5.4, 5.9, 5.10, 5.11, 5.12 and Fig 4.10 respectively.*

that FoxH-mediated activation of Nodal target genes was present as early as the Protostome/Deuterostome common ancestor. However, other Lophotrochozoan genomes from outside the Gastropoda do not possess a clear *FoxH* sequence, casting doubt on the assignation of this sequence as a firm ortholog of the FoxH family, and further investigation will be necessary if identity is to be confirmed.

3.3.5 *Nodal* Locus Recovery

One of the stated goals of sequencing the *P. vulgata* genome was the recovery of the genomic region surrounding the *Nodal* locus, allowing examination of the regulation of this gene in this species, and by comparison the inference of ancestral modes of *Nodal* regulation amongst the Metazoa. To my surprise, not one but two Nodal loci were discovered.

While there is no guarantee that sequencing efforts will recover any particular locus within a genome, portions of the two *Nodal* loci were successfully found within all builds thus attempted. The consensus build of these, when combined with data derived from more conventional PCR and RACE PCR based methods can be seen in Fig. 3.9. We satisfactorily discerned the sequence of an approxi-

mately 15kb fragment of one *P. vulgata Nodal* locus and approximately 5kb of the other, which should be of great utility in examining the regulation of this gene. For further discussion regarding these Nodal loci, please see Chapter 5 of this report.

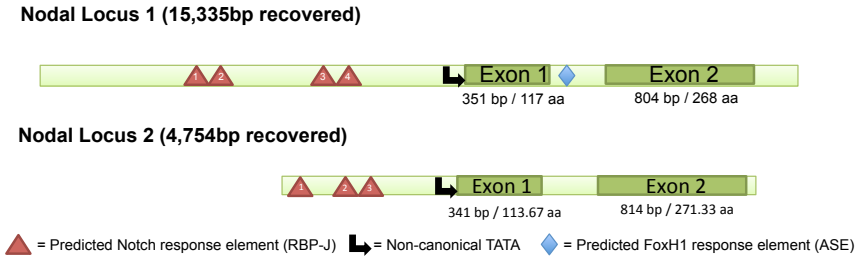


Figure 3.9: P. vulgata Nodal *Loci Recovery (not to scale), derived from PCR-based methods and genomic sequencing. Also shown are RBP-J protein (Notch mediator) and FoxH predicted binding sites*

3.3.6 The *P. vulgata* Homeodomain Gene Complement

As a further proxy for coverage and to provide an insight into molluscan developmental pathways, a number of interesting families of genes have been briefly investigated. The *P. vulgata* homeodomain-containing gene complement, as ascertained from the current best build of the genome, seems a quite complete list, containing most of those expected of the metazoan complement, as can be seen in Table 3.8. This provides further support for the assertion that the majority of the coding sequence of the genome has been recovered in my build. Unfortunately, however, scaffolding has not been successful enough to recover a homeobox cluster.

Several apparent losses, highlighted in yellow in Table 3.8, could be the result of my build not containing sequence data, rather than actual loss. As a test of this, the *L. gigantia* genome, available at `http://www.jgi.doe.gov/`, was searched for representatives of those genes apparently missing from my dataset, as loss in *L. gigantia* could suggest Patellagastropoda, gastropod or mollusc-wide loss, rather than a simple absence from my dataset. Of the 14 apparent absences from my dataset, losses were also observed in *L. gigantia* for all Hnf-class genes, *Bari*, *Pou1* and *Vax*. *Pax* β also appears absent from *L. gigantia*, and may therefore be an annelid-specific, rather than lophotrochozoan-specific paired box gene. The loss of *Pou1* may indicate that this gene has been lost throughout the Protostoma, as it is present in

Cnidaria and Deuterostoma but absent from all protostome species thus sequenced.

This leaves a final total of up to eight homeodomain genes absent from my dataset, either as a result of poor assembly, or through real loss. Out of a total of nearly 100 homeodomain gene families, this represents exceptional recovery from a single lane of sequencing.

ANTP							
HOXL subclass							
Cdx	*Evx*	*Gbx*	*Gsx*	*Hox1*	*Hox2*	*Hox3*	***others in class:***
Hox4	*Hox5*	*Hox7/Antp*	*Lox2*	*Lox4*	*Lox5*	*Lox18*	*Lox gene conservation unknown*
Meox	*Mnx*	*Post1*	*Post2*	*Xlox*			
NKL subclass							
Abox	*Barh*	*Bari*	*Bsx*	*Dbx*	*Dlx*	*Emx*	*Ankx - amphioxus only, Barx - poss deuterostome specific, related to Bar*
En	*Hhex*	*Hlx*	*Lbx*	*Msx*	*Msxlx*	*Nedx*	*Hx - amphioxus only, Lcx - amphioxus only, Nanog - mouse and human?*
Nk1	*Nk2.1*	*Nk2.2*	*Nk3*	*Nk4*	*Nk5/Hmx*	*Nk6*	
Nk7	*Noto*	*Ro*	*Tlx*	*Vax*	*Demox*	*Ventx*	
PRD							
Alx	*Arx*	*Dmbx*	*Drgx*	*Gsc*	*Hbn*	*Otp*	*AprdA-E - Amphioxus only, Argfx -vert only, CG11294 - insect only? Dprx - placental*
Otx	*Pax2/5/8*	*Pax3/7*	*Pax4/6*	*Pax9*	*PaxBeta*	*Phox*	*Dux - not expected in invertebrates, Esx - vert only, Hesx, Hopx, Isx - deut only*
Pitx	*Prop*	*Prrx*	*Rax*	*Repo*	*Shox*	*Uncx*	*Leutx - primate specific, Mix, Nobox, Obox, Rhox, Sebox, Tprx - all vert only*
Vsx	*Other/unclassified (4)*						*PaxBeta could be annelid specific*
LIM							
Lhx1/5	*Lhx2/9*	*Lhx3/4*	*Lhx6/8*	*Lmx*	*Isl*		
POU							
Pou1	*Pou2*	*Pou3*	*Pou4*	*Pou6*			Pou1 - in cnidaria and human, Pou5 - vert only, Hdx, Human only
SINE							
Six1/2	*Six3/6*	*Six4/5*					
TALE							
Irx	*Meis*	*Mkx*	*Pbx*	*Pknox*	*Tgif*	*Unknown*	Atale - amphioxus only
CUT							
Cmp	*Cux*	*Onecut*					Acut -amphioxus only, Satb - vertebrate only
PROS							
Prox							
ZF							
Zfhx	*Zeb*	*Tshz*					Adnp - only verts, Azfh - amphioxus only,Zhx/Homez - no dros, but could be present
CERS							
Cers							
HNF							
Hmbox	*Hnf1*						Ahnfx -amphioxus only

Table 3.8: *The* P. vulgata *homeodomain complement, as inferred from ABySS genome builds. Gene identity inferred from Blastx analysis of sequences in comparison to NCBI nr database. Some genes are present in multiple copies (data not shown). Genes absent from dataset indicated in yellow, and particularly surprising genes indicated in light green. ANTP gene family information provided by J. Hui*

Two particularly intriguing gene sequences are also apparently present in my dataset - *Demox* and *Ventx*, highlighted in green in Table 3.8. These genes are not expected in lophotrochozoan complements, being thought to be poriferan and deuterostome specific respectively. If these sequences are in fact members of this class, their presence could result from contamination either in sample preparation or sequencing itself, although this is unlikely, as blast identity does not suggest that these gene sequences belong to known (and regularly sequenced) model organisms, or from close relatives to these. Tree building efforts suggest that these could in fact be divergent copies of Paired-class genes, which group with these genes through weak long branch attraction, rather than true orthologues of these genes. Further investigation is necessary to ascertain whether the presence of these genes in *P. vulgata* is real or artifactual, and their likely role if their presence is con-

firmed, but these sequences provide a clear target for future research to investigate.

3.3.7 The *P. vulgata* Fox, Sox and T-Box Complements

My dataset was further examined to uncover the *P. vulgata* Fox, Sox and T-Box complements. Again, these were found to be remarkably complete. Of the 23 generally accepted families of Fox gene, only one (*FoxQ1*) which was expected to be found was not recovered. This may be a case of genuine loss, miscategorisation of hits into families by blast identity or absence from my dataset.

Fox Genes:			**Sox Genes:**	**T -box Genes:**
FoxA (1)	FoxH1 (1)	FoxM (1)	Sox B1 (1)	Brachyury (5)
FoxAB (1)	FoxI-like (1)	FoxN1/4 (1)	Sox B2 (1)	T-brain (2)
FoxB (1)	FoxJ like	FoxN2/3 (1)	Sox C (1)	Tbx 1/10 (1)
FoxC (1)	FoxJ1 (4)	FoxO (2)	Sox D (1)	Tbx 2/3 (11)
FoxC-like (1)	FoxJ1-like (1)	FoxP (1)	Sox E (1)	Tbx 4/5 (0)
FoxD (1)	FoxJ2/3 (2)	FoxQ1 (0)	Sox F (1)	Tbx 6 (1)
FoxE (1)	FoxK (2)	FoxQ2 (3)	Sox H (2)	Tbx 15/18/22 (2)
FoxF (1)	FoxL1 (1)	FoxQ2-like (1)		Tbx 20 (3)
FoxG (2)	FoxL2 (1)	Fox Unknown (3)		Tbx Unknown (1)
Not expected (Vertebrate Specific):		Fox R, S	Sox A, Sox G	

Table 3.9: *The* P. vulgata *Fox, Sox and T-Box gene complement, as inferred from ABySS genome builds. Gene identity inferred from blastx analysis of sequences in comparison to NCBI nr database. Number of homologues within each catalogue are indicated in brackets. Genes absent from dataset indicated in yellow. Fox gene family information provided by S. Shimeld*

The entire expected Sox complement was recovered (including *SoxH*, until recently firmly thought to be chordate-specific), while the *T-box 4/5* gene family was the only member of the T-box complement to be absent from my dataset. As can be seen in Figure 3.7 the *T-box 4/5* family is also absent from *D. melanogaster*, and further investigation could not find a homologue in the present *L. gigantia* and *C. teleta* genome assemblies. The *T-box 4/5* family may therefore have been lost in the protostome lineage, as it is present in the Porifera and Deuterostoma.

As noted in section 3.3.5, of particular interest for my studies into the establishment of L/R asymmetry was the recovery of a *FoxH1* orthologue, the first to be described outside the Deuterostoma. This gene plays a variety of roles in the regulation of *Nodal* and establishment of L/R asymmetry, so its recovery would raise further questions regarding the degree of conservation of this process across

the Metazoa. Using the *P. vulgata* sequence, blast searches through the *L. gigantia* genome also recover a sequence with intriguing homology to this gene, and investigations are continuing in the Shimeld laboratory with the aim of confirming this assignation.

3.3.8 *P. vulgata* Genome: Conclusions and Future Work

The sequencing of the *P. vulgata* genome has provided us with a resource of great utility for a range of investigations, not just those directly related to this thesis. The Lophotrochozoa in general and the Mollusca in particular remain undersampled in terms of genomic resources, and the additional dataset described here will provide another reference point for future studies. The high apparent recovery of coding regions will facilitate this, and while syntenic data is perhaps lacking from this resource, it will nonetheless be of utility to a range of investigations in a number of fields in the near future.

3.4 Rotifer (*Brachionus plicatilis*) Genome Assembly

In collaboration with the Dearden laboratory at the University of Otago, I have been performing a range of analyses on a genome for the monogont rotifer *B. plicatilis*. This has complemented previous work, performed by myself and others, on a mixed transcriptomic sample of this species, which is now being prepared for publication alongside a range of expression data. The enigmatic phylogenetic position of rotifer makes it an interesting candidate for genomic comparison, as insights gleaned from this resource will allow a range of inference into the biology of this and other species, with Chapter 5 providing a good example of what can be learnt from such a study.

3.4.1 Sequencing results and Quality Control

DNA from a mixed age sample of *B. plicatilis* was extracted using a standard PHB-based protocol by members of the Dearden laboratory. This was sequenced on a

single lane of an Illumina HiSeq 2000 platform by New Zealand Genomics Limited, and the data made available to me for quality assessment and assembly.

The raw data for the genomic sequencing results can be seen in Table 3.10. The GC content of the genomic reads (28 %) is quite low, and is lower than the GC content of a 454-based transcriptome previously sequenced in the Dearden laboratory (33.7 %).

Table 3.10: *Summary statistics pertaining to reads used in assembly of the* B. plicatilis *genome, and for* B. plicatilis *sequence read cleaning trials*

Metric	Original Reads	Reads After Cleaning
Library Insert Size:	300 bp	300 bp
Paired-end Sequences	168,368,886	164,111,408
Single-end Records Kept	n/a	2,094,740
Read Length (bp)	101	20-101
Amount of Data (Gb)	41.08	39.81
GC content (%)	28	28

FastQC (`http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc`) was used to assess the quality of read data. Unfortunately, while sequencing quality was more than adequate (the median per-base sequence quality was excellent, above the 36 phred mark from the start of reads through to the 101st base), a number of sequences were found to be over-represented in my dataset.

To test if these were the results of bias in base calling, and to attempt to correct errors if present, Sickle and Musket were run on my dataset to trim and correct erroneous sequences. The results of such filtering can be seen in the final column of Table 3.10. Significant numbers of paired-end reads had one read entirely discarded as a result of this process (2,043,459 from one end, 51,281 from the other), and empirical testing of assemblies before and after read cleaning suggested that this process greatly assisted assembly of contigs into longer scaffolds.

3.4.2 Coverage

The cleaned and corrected reads resulting from sickle and musket were used as the basis for assembly trials using a range of assembly software. Velvet was noted at an early stage to outperform other assemblers used by most metrics, and was chosen as the basis for a number of more targeted trials.

Early on in my assembly efforts, it was noted that two markedly different dis-

tributions of *k*-mer coverage appeared to be represented in my samples. These can be seen in Fig. 3.10, with distributions of *k*-mer coverage shown at a *k*-mer size of 51 (this data comes from reads prior to cleaning, but similar results can also be seen after this). Two clear peaks of coverage can be seen, one with a peak at around 28-fold coverage, and one at 67-fold coverage.

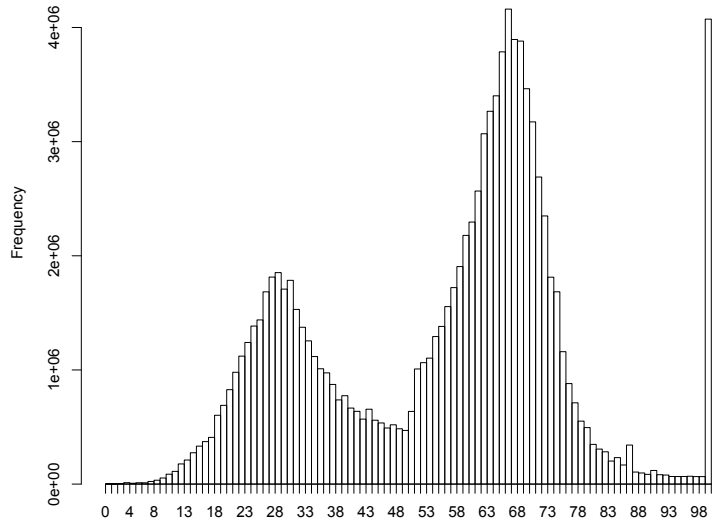


Figure 3.10: *Weighted histogram of coverage at a k-mer size of 51. Note two distinct peaks in coverage, one at approximately 28 fold coverage, and one at 68 fold coverage.*

From a single lane of HiSeq, I would expect approximately 40-fold coverage of the estimated diploid genome size of *B. plicatilis*, which is around 0.7 pg (Mark Welch & Meselson., 1998). Instead, the two peaks indicate an effective genome coverage of 134 and 56 (at k-mer coverages of 67 and 28 respectively). This would point to 123 and 294Mb genomes, given the approximately 168 million reads I began with. While there is no guarantee that the published estimated size is correct, my data is well wide of this figure.

A number of hypotheses for what caused this were postulated. Contamination is a possibility, although considerable amounts of this would be needed to produce such clear peaks, and would be easily detected. This distribution may also indicate an ancient whole genome duplication in this species, or at least duplication

of large portions of the genome, which has been noted before in the Rotifera (Flot et al., 2013). Wide variation in coverage has also been noted before in the Honeybee genome (The Honeybee Genome Sequencing Consortium (2006), P Dearden, personal communication).

In order to test whether the difference in distributions of coverage represented contamination from another source, Velvet was used to assemble the data with an assembly cut offs of above and below a *k*-mer coverage of 49. While some bacterial contamination (and entirely recovered bacterial genomes) was detected in the larger-than-49 coverage dataset, the coverage of these contigs was in excess of 100-fold on average, and does not contribute substantially to the peak seen at 67-fold coverage. Only one complete mitochondrial genome was recovered in this dataset, and metazoan contamination seems unlikely.

When the two datasets are compared directly, GC content varies widely, as does the number of annotatable genes. The GC % of the greater-than-49-fold coverage set is 23.73, whereas that of the less-than-49-fold coverage dataset is 36.52. A small number of genes of interest can be found in the high cutoff dataset (along with many genes of clear bacterial origin), but by far the largest proportion of interesting genes can be discovered in the low cutoff dataset. A survey of homeodomain-containing genes, for instance, finds them almost entirely in the low cut off set (less than 49) when searches are made of these assemblies.

It seems therefore that the two peaks of coverage correspond to regions of gene density, with high coverage (and possibly repetitive) regions of the genome containing little in the way of genes. Why these regions are found in such high number compared to other regions of the genome is at present a mystery. There is no 'smoking gun' evidence for any form of whole genome duplication that I have uncovered in my investigations, and why these regions would be retained while gene-containing regions are lost is difficult to explain. The answer to this could perhaps be gained from a larger sequencing effort, with a variety of long mate pair libraries, which would facilitate scaffolding. Unfortunately, due to funding constraints, this is unlikely to proceed any time in the near future.

	Velvet 21	Velvet 31	Velvet 41	Velvet 51	Velvet 61
Min contig length	100	100	100	101	121
Max contig length	9,202	14,554	37,484	39,718	226,365
Mean contig length	243.2	300.92	372.21	530.52	963.47
Standard deviation of contig length	274.97	388.77	503.22	856.68	3283.11
Median contig length	151	160	192	243	248
N50 contig length	303	471	613	1,049	7,336
Number of contigs	485,226	527,324	412,382	287,487	144,244
Number of contigs >=1kb	12,044	24,862	30,541	**35,383**	16,467
Number of contigs in N50	**93,366**	83,074	60,692	33,401	4,343
Number of bases in all contigs	118,005,179	**158,684,367**	153,491,510	152,518,619	138,974,342
Number of bases in contigs >=1kb	18,466,350	40,320,153	53,327,208	78,288,712	101,610,873
GC Content of contigs	**44.21%**	38.22%	35.89%	33.16%	30.61%
	Velvet 71	**Velvet 81**	**Velvet 91**	**All Contigs**	**After Vmatch**
Min contig length	141	161	181	100	100
Max contig length	226,670	**356,246**	154,866	356,246	356,246
Mean contig length	1498.26	**2095.09**	1983.34	518.26	435.52
Standard deviation of contig length	6106.52	**7398.92**	3776.77	2132.39	1829.92
Median contig length	251	277	**586**	192	204
N50 contig length	18,464	**19,652**	5,868	1,311	769
Number of contigs	83,536	56,529	**47,477**	2,044,205	1,036,057
Number of contigs >=1kb	10,197	10,451	18,021	157,966	68,577
Number of contigs in N50	1,721	1,606	4,170	111,119	98,045
Number of bases in all contigs	125,158,924	118,433,287	94,163,033	1,059,429,261	451,219,455
Number of bases in contigs >=1kb	104,023,211	**104,276,066**	82,782,711	583,095,284	199,853,355
GC Content of contigs	27.82%	26.02%	24.21%	32.98%	34.67

Table 3.11: *Assemblies of rotifer (*B. plicatilis*) genome data, and final multi k statistics. Highlighted in bold font are the statistics representing the 'best' recovery of each category, as conventionally measured.*

3.4.3 Summary of Assembly Results

In order to gain the best possible assembly of my dataset, it was decided to undertake a additive multiple-*k* assembly of the rotifer dataset, allowing the longest build of each sequenced contig to be retained. In order to allow for the long repeat regions found in multiple parts of genomic builds, which could obfuscate removal of redundancy (as Vmatch would be unable to tell a repeat sequence from one portion of the genome from the same sequence elsewhere in the genome) I set two sequence lengths for comparison and redundancy removal. Contigs 100 bp - 499 bp in length were compared with a window size of 100 bp, with all sequences 100 % identical over that length assessed to see whether they were redundant copies, with the shortest sequence discarded in those cases. Sequences longer than 499 bp were compared with a window size of 500 bp, with similar removal of redundancy at this window size.

The results of my final, additive multiple-*k* assembly, as well as the constituent assemblies which make it up, can be seen in Fig. 3.11. This assembly will no doubt still contain some redundancy as I have been conservative with my cut-offs for

comparison using Vmatch, so as not to discard potentially useful data. However, the recovery of 450 Mb would still represent a considerable fraction of the estimated haploid genome of *B. plicatilis*, even if some redundancy is incorporated into this figure. The recovery of less than the full expected genome size is however commonplace in de novo assembly efforts, as seen earlier in this chapter, as repetitive sequences are poorly resolved and recovered in such assemblies.

The longest build shown in the final Vmatch assembly, with a length of 356,246 bp, is in fact a bacterial genome, and several other long bacterial fragments are part of this dataset, which skews the N50 slightly upwards. These have been left in the interim in my assembly, as other rotifer have been shown to incorporate bacterial genes into their genomes (Flot et al., 2013). In case of publication, the longest fragments will be removed, but the data will not entirely be scrubbed, as it is possible that these represent recent horizontal transmission. A strong caveat will, however be appended to my discussion in this regard.

From the perspective provided by the two *k*-mer coverage restricted builds discussed in the previous section, the GC % content found in the individual builds is of interest. A very high GC % (44.21) is recovered at the lowest *k*-mer size (21), suggesting that this build recovers a good proportion of the coding portion of the genome of this species. In contrast, the highest *k*-mer size used (91) recovers a dataset with only 24.21 % GC bases. The final assembly after removal of redundancy contains 34.67 % GC proportionality, which may be an overestimation of the amount in the rotifer genome (repetitive sequence is often AT rich) but is much more in line with the expected metazoan percentage.

3.4.4 Rotifer TGF β Complement

The genome resource presented here has already been used to identify and extract a range of developmentally interesting genes for cloning and analysis of expression. For the purposes of this thesis, the most interesting of these are the TGF β complements, and in particular the Nodal pathway of this species.

The *B. plicatilis* TGF β signalling complement has proven markedly reduced when compared to the more usual metazoan cassette. From my genome I was able

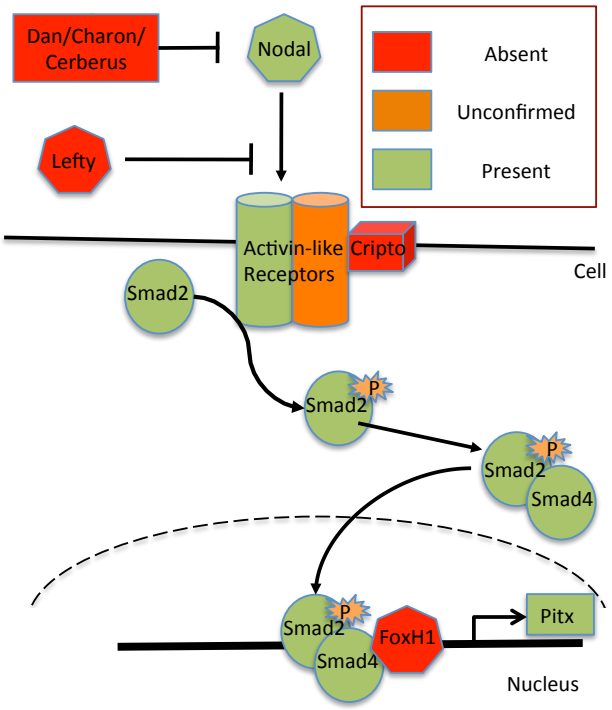


Figure 3.11: *Presence and absence of the core Nodal signalling cascade in the rotifer B. plicatilis, as revealed by genomic and transcriptomic sequencing. A variety of regulatory proteins are absent, while the identity and role of the ALK receptors in this species remains unknown. For evidence of presence of Nodal, Activin-like Receptors, Smads, Cripto and Pitx please refer to Figs. 5.4, 5.9, 5.10, 5.11 and Fig 4.10 respectively, with rotifer Pitx sequence studied in detail in Grande et al (2014). For more details refer to Chapter 5*

to identify a range of components of the core Nodal signalling cascade, but with some notable exceptions, particularly in the regulatory complement of this species. While proving absence from an incomplete genome is difficult, it should be noted that these components are also missing from an independent deep transcriptomic build of this species, while the genes present in the genome are also found, at least in fragmentary form, in this resource.

This genome resource therefore provides reasonable recovery of conserved signalling pathways as a whole. While I cannot make firm inference as to absence of individual genes, the broad schema provided by this data is helpful in investigating the evolution of this cascade in the Lophotrochozoa as a whole. This is described in more detail in Chapter 5 of this report.

3.4.5 Comparison with a Published Rotifer Dataset

The genomic sequence of the bdelloid rotifer *A. gaga* has been published (Flot et al., 2013) and provides an excellent benchmark for my results. The genome of this bdelloid species, a clade noted for an absence of males, hermaphrodites and meiosis, is notable for its oddities, the result of generation upon generation of mitotic

reproduction.

The total genome size of *A. gaga* is around 244 Mb, smaller than estimates in *B. plicatilis* derived from fluorometry and corroborated by my data. It has, by some analyses, undergone two rounds of whole genome duplication, with extensive loss occurring both before and after this event. While my genome build is not of sufficient quality to allow large scale comparison of haplotypes in the same way as that of Flot et al. (2013), I find no evidence for tetralogy in the genes I have examined in detail.

In the few gene families I have examined in *B. plicatilis* I find a much more representative sampling of common metazoan genes. The homeobox genes, for example, appear to be present in single copy number, although posterior homeobox genes are entirely absent. Transcriptomic analysis (data not shown) also mirrors this. Prior investigations into the neurotransmitter receptor complement of *B. plicatilis* by myself and others have revealed no evidence of large-scale duplication in a range of well-described gene families.

Up to 8% of the genome of *A. gaga* is estimated to be derived from non-metazoan species as a result of horizontal gene transfer (HGT). While I am loathe to use my genomic data to estimate whether HGT is present in the genome of *B. plicatilis*, due to contamination by bacterial DNA, prior transcriptomic analysis (performed before the work described in this thesis) contained no obvious evidence of HGT, and to my knowledge no evidence for this has been presented in Monogont rotifer.

3.4.6 Summary, Rotifer Genomics

The genome of *B. plicatilis* will represent a sample of interest to a wide range of evolutionary and developmental biologists. While I am cognisant of the limitations of my dataset with regards to inference of synteny and as a result of microbial contamination, for 'gene mining' it has already shown its worth. In combination with transcriptomic analysis and in situ analyseis of expression of important developmental genes (performed by others) this is in preparation for publication, and will represent an excellent out group to the very derived data presented in the bdelloid

rotifer *A. gaga* (Flot et al., 2013).

3.5 *Biomphalaria glabrata* and *Crepidula fornicata* Transcriptomic Study

During a collaborative visit to the Grande laboratory in the Universidad Autonoma de Madrid I became involved in a range of work on the transcriptomes of the non-model gastropod species *B. glabrata* and *C. fornicata*. This is primarily aimed at the generation of resources for gene identification and cloning. These sequences have, however, also proven invaluable when used to confirm the presence of genes of the TGF β signalling cascade in the wider Mollusca, and have been useful for identifying full length sequence in the *P. vulgata* datasets used in Chapter 5 of this thesis.

Table 3.12: *Transcriptome assembly metrics for* B. glabrata *and* C. fornicata. *Transcriptomes assembled using Trinity using RNA samples prepared in the Grande laboratory.*

Organism:	*Biomphalaria glabrata*	*Crepidula fornicata*
Number of contigs	250,605	326,118
Max contig length (bp)	28,281	28,997
Mean contig length (bp)	1320.50	563.91
Median contig length (bp)	556	350
N50 contig length (bp)	2,852	717
# contigs in N50	33,393	61,551
# contigs 1kb $\leq$	90,315	37,979
# bases, total	330,923,222	183,902,140
# bases in contigs 1kb $\leq$	265,182,009	72,136,873
GC Content %	37.74	42.29

It is hoped that the *B. glabrata* dataset will be published under the title "Deep Transcriptome of the Schistosomiasis Vector *Biomphalaria glabrata* (Gastropoda: Planorbidae) Illuminates Molluscan Disease-Related Pathways" in the near future, with a focus on the innate immune components and other disease resistance pathways found in this species. However, as much of the gene mining in this regard has been performed with and by others this work is not presented here. The *C. fornicata* dataset is less likely to be published in its current form, but may be incorporated into analysis on a genomic resource in the future.

3.5.1 *B. glabrata*

As a sinistrally coiling mollusc, *B. glabrata* is a vital outgroup for developmental study, most of which takes place in dextrally coiling species. Much work has already been conducted into the genetic and molecular responses made by *B. glabrata* after infection with schistosomiasis, in the hope of identifying potential targets for treating and mitigating the effects of this disease (Deleury et al., 2012), providing a limited set of transcriptomic resources.

This prior investigation has revealed the sequence of several molecular families within *B. glabrata*, but has been hamstrung by the limitations of EST- and specific target gene-based approaches. Presently existing public sequence resources for *B. glabrata* are limited, despite a range of prior efforts, with EST based (Lockyer et al., 2007, 2008), BAC (Adema et al., 2006) and transcriptomic datasets (Deleury et al., 2012) identifying, at best, around 700 annotatable contig sequences.

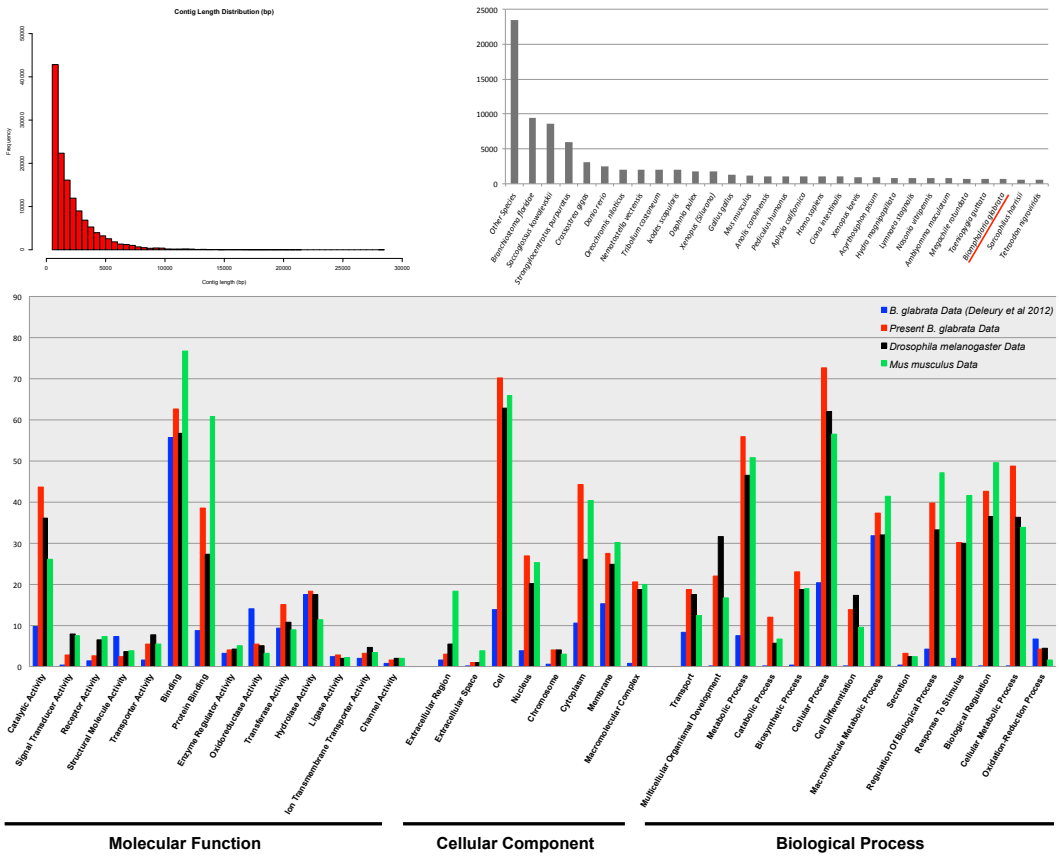


Figure 3.12: *Analysis of the composition of the transcriptome of* B. glabrata. *Contig length distribution graphed using R. Blast2GO used to compare the distribution of Blastx hits vs the nr database by best blast hit by species (top right) and by GO category (below).*

My data, as shown in the first column of Table 3.12, provides a much deeper resource to draw from than heretofore available in this species. With 90,315 contigs greater than 1 kb in length, a large number of identifiable sequences were obtained, although a substantial amount of polymorphism is reported by Trinity and must be taken into account in this figure. KEGG pathway mapping (data not shown) suggests I have near-100% coverage of all major signalling and metabolic pathways, with well conserved processes such as gluconeogenesis and the citrate cycle demonstrating complete coverage of all their constituent steps.

Fig. 3.12 shows a variety of metrics relating to this data. Most interesting is the assessment of my results using Blast2GO. Firstly, despite the recent accessioning of the oyster *C. gigas* genomic data onto the nr database, the closest hits gained using BlastX are more heavily weighted towards deuterostome species. These species have previously been noted as having a slow rate of molecular evolution, and it may be this which causes this result. The small number of *B. glabrata* top hits (underlined in red, top right of Fig 3.12) also testifies to the presently depauperate state of *B. glabrata* representation in the nr database.

Comparison of the results of Blast2GO analysis between my data, previously extant data for this species, and the distributions of the proteomes of the well-described model organisms *D. melanogaster* and *M. musculus* confirm that my dataset is more representative than that available previously. While some differences are to be expected in the distribution of GO identities from species to species, my results, in red, mirror those of the fully sequenced species much more closely than those of previous *B. glabrata* datasets shown in blue. In concert with my KEGG mapping results, I am therefore confident that this dataset contains the sizeable majority of transcribed RNA in this species, although it is likely a small number of RNAs of low copy number and restricted temporal expression are not present.

For my studies, the sequences resulting from this transcriptome have already proven valuable both for confirming the presence of key genes in the Mollusca and as the basis for molecular cloning. A deep transcriptomic resource will also allow a range of biomedical investigations to take place, provide a firm underpinning to assertions regarding the *B. glabrata* genetic cassette, and further allow research

into invertebrate immune systems, an area where our knowledge is still nascent at best. The dataset presented here will therefore stand as a vital resource for the assessment of patterns of evolution within the Mollusca, and for human health, in efforts investigating the progression of schistosomiasis within *B. glabrata*.

3.5.2 *C. fornicata*

My *C. fornicata* dataset is not as well-assembled by many metrics as that of *B. glabrata*, as can be seen in column 2 of Table 3.12, but still represents a sizeable addition to our knowledge of sequences in this organism. The markedly lower assembly quality by many metrics is not clearly the result of differences in RNA quantity or quality, as these were of an approximately similar standard to the *B. glabrata* sample used. However, fewer reads were returned to us from the lane used for sequencing this data, and these were of marginally poorer standard, with a median PHRED score of less than 30 at the last bases (data not shown).

Perhaps as a result of the comparative paucity of the reads used for assembly, lower number of bases make up the *C. fornicata* assembly than that of *B. glabrata* (183,902,140 vs 330,923,222). Other metrics suffer similarly, perhaps a combination both of the lesser read quality and the higher diversity you would expect to find in a highly polymorphic sample (the *C. fornicata* sample was wild-caught, while the *B. glabrata* sample came from a captive, albeit not inbred, strain). However, the dataset summarised in column 2 of Table 3.12 is still represents a significant advance in our knowledge of *C. fornicata* sequence, which has hitherto been limited to the results of a single study (Henry et al., 2010b).

The results of a broad analysis of the *C. fornicata* dataset can be seen in Fig. 3.13. My dataset was heavily biased towards small contigs, as can be seen in the graph at top left. Like many of my other lophotrochozoan datasets, basal deuterostomes were overly represented in my Blastx best hits. However, *C. gigas* was the most common species found in this comparison. While the *C. gigas* dataset was not submitted to NCBI before the analysis shown for *P. lamarckii* was performed, it was online at the time the *B. glabrata* dataset was tested. This may suggest that *C. fornicata* and *C. gigas* are closer in molecular sequence identity than *B. glabrata* to either

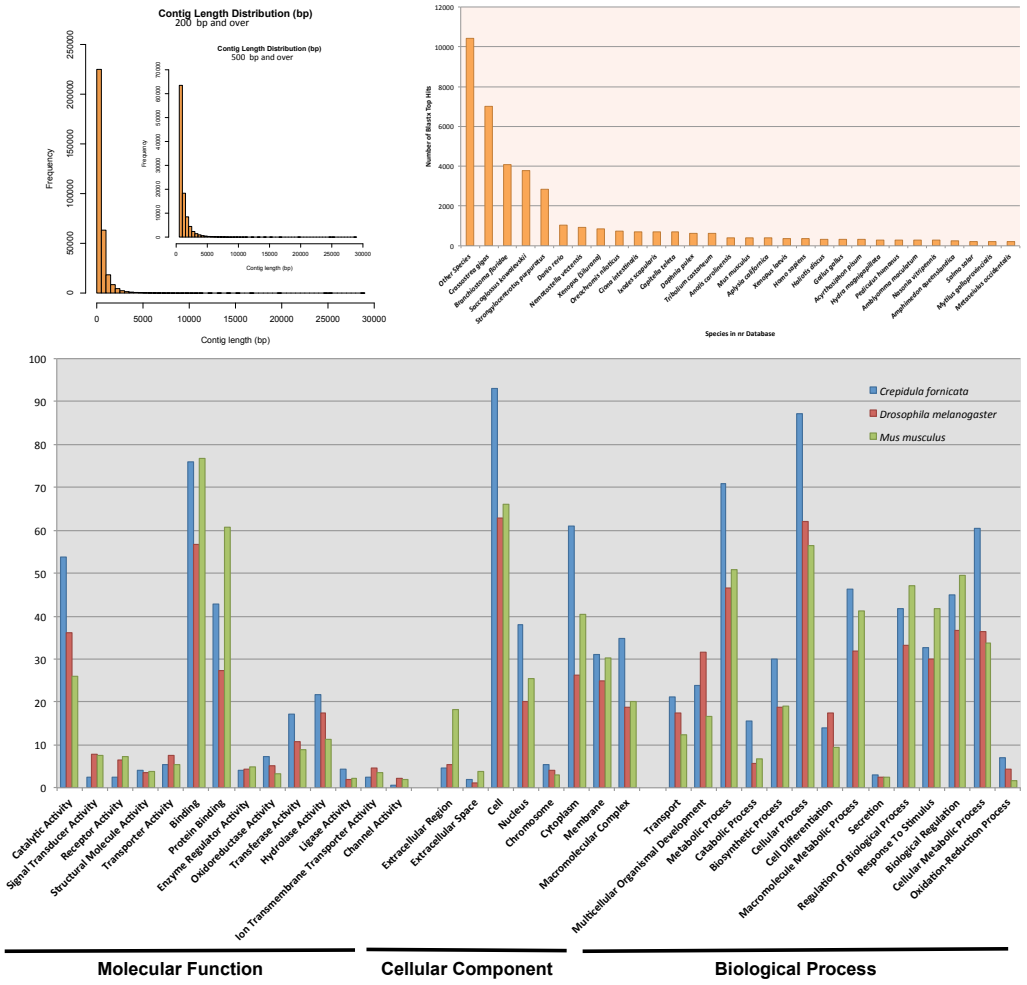


Figure 3.13: *Analysis of the composition of the transcriptome of* C. fornicata. *Contig length distribution graphed using R. Blast2GO used to compare the distribution of Blastx hits vs the nr database by best blast hit by species (top right) and by GO category (below).*

species, a surprising finding given the evolutionary distance between gastropods and bivalves.

Comparison of the distribution of the *C. fornicata* dataset by proportional GO identity to well annotated ecdysozoan (*D. melanogaster*) and deuterostome (*M. musculus*) data sets indicates that my data is perhaps weighted towards structural and housekeeping genes. We find over-representation of classes such as 'Metabolic Process' genes, while more developmentally interesting datasets such as 'Regulation of Biological Process' and 'Biological Regulation' are more typical in their proportional recovery. This may be a result of the mixed embryonic and larval stages used to construct the transcriptome, which will be expressing the full complement of

structural and regulatory genes as they grow, perhaps at the expense of more specialised classes of gene involved in specific pathways, which would be expected to be found in discrete adult tissues. KEGG mapping results (data not shown) corroborate this, with major metabolic and cellular pathways well recovered in my dataset.

This *C. fornicata* resource has already been used to classify and clone a variety of genes noted as related to L/R asymmetry in other laboratories. While there are no extant plans to publish this resource, it may be used for a variety of further investigations if the mooted *C. fornicata* genome project is begun. While the *C. fornicata* transcriptomic dataset presented here is not as deep and well-assembled as some others noted here, it nonetheless represents a substantive contribution to the knowledge of the molecular repertoire of a still under sampled clade, and one of use to a range of investigations worldwide in this developing model species.

3.6 Conclusions

This section of my thesis has been very successful, establishing resources which will be of my utility for my studies, that of the wider research group and the lophotrochozoan community for many years to come. While initial work on this sphere of my work - cloning using limited genetic data and degenerate PCR methods - was painstaking and often fruitless, I now have access to a number of databases containing the sequence of all the genes required for my studies, which form the basis of the work described later in this report.

Breaking Initial Lophotrochozoan Symmetry

4.1 Breaking Initial Symmetry

As detailed earlier in this work, a number of mechanisms have been implicated in the initial break of symmetry in vertebrate model systems. These include cytoskeletal cues, ion movement, cilial motion, and retinoic acid, Notch and Shh signaling (Vandenberg & Levin, 2010). To date my knowledge of a role for these is largely limited to model species within the Chordata, with occasional evidence from the wider Deuterostomia, and many mechanisms seem peculiar to a single species or small clade.

Lophotrochozoans represent an excellent out-group for the derivation of the truly ancestral mechanisms for establishing L/R asymmetry from currently contradictory evidence. The retention of *Nodal* in this clade suggests that the ancestral mechanisms for activating this gene and establishing L/R asymmetry may also be conserved, unlike in ecdysozoan species yet examined. With that in mind, I set out to test a number of factors which have been shown to be involved in the break of initial symmetry in deuterostomes in my model species.

This chapter is subdivided into several sections, each with a number of key findings. These are

- The identification and determination of the wild-type expression of Na^+/K^+ ATPases in *P. vulgata* embryos, along with the distribution of serotonin in early cleavage stages.

- The classification of the wild-type mRNA expression patterns of the key genes *Nodal* and *Pitx* as markers of the correct establishment of asymmetry.
- The functional testing for a role of ion channels in the establishment of asymmetry in *P. vulgata*, using a range of inhibitors.
- Speculation as to a potential role for Na^+/K^+ ATPases in the specification of the D quadrant, given preliminary evidence from omeprazole treated samples.

This work, while unpublished to date, may form the basis of further investigations into both asymmetry establishment and D lineage specification in the future.

4.2 Putative Symmetry Breakers

While there are several mooted mechanisms for the breaking of symmetry in early embryos, time pressures dictated that a subset of these were selected for investigation. I therefore concentrated on the mooted role of ion channels in providing the initial disparity in charge across embryos, due to the functional tests available in this regard using a range of pharmacological inhibitors. For detailed explanations of the role of these mechanisms in the establishment of vertebrate L/R asymmetry, please see the introduction to this thesis. I was also able to investigate the localisation of mooted morphogens said to act downstream of the initial break, as also described later in this section.

While within the Deuterostomia it has been clear for some years that ion flow plays a conserved role in establishing L/R asymmetry (Hibino et al., 2006; Shimeld & Levin, 2006), to date the conservation of this within the wider Bilateria has not been tested. A variety of ion channels have been implicated in this process, with Na^+/K^+ and H^+/K^+ ATPases the most robustly evidenced to play a functional role in establishing asymmetry.

'H^+/K^+' ATPases have been shown to act in the patterning of the head of adults in the Platyhelminth *Dugesia japonica* (Nogi et al., 2005). The role of these in other members of the Lophotrochozoa, or in the establishment of asymmetry at early em-

bryonic stages, is as yet unknown. *H^+/K^+ ATPase* mRNA distribution represents one of the earliest observable signs of L/R asymmetry in *X. laevis* (Levin et al., 2002). If this functional role was also observed in lophotrochozoans it would represent potent evidence of an ancestral role for these ion pumps in the establishment of L/R asymmetry.

4.2.1 ATPase Phylogeny

Investigations into the presence of *H^+/K^+ ATPases* in my model species did not uncover these genes within my datasets. Instead, as can be seen in the maximum likelihood tree shown in Fig. 4.1, within the P-type cation transport ATPases, vertebrate *H^+/K^+ ATPase* genes appear to have emerged as paralogs of *Na^+/K^+ ATPases* within the vertebrate common ancestor, perhaps as a consequence of known whole genome duplication events.

Lophotrochozoans, and invertebrates in general, do not possess the subunits which make up H^+/K^+ ATPase proteins. This can further be seen in Fig. 4.2. Invertebrates do, however, possess Na^+/K^+ ATPases, which have been implicated directly in the establishment of asymmetry in Zebrafish (Ellertsdottir et al., 2006). The identity of the cation may be unimportant if the role they play remains constant - establishing a gradient of charge to enable polar molecules to differentially sort along an axis.

Ciona intestinalis Na^+/K^+ ATPase α subunits (two of which are present in the genome of this species, although this clade is collapsed and one is shown here) fall outside the monophyletic protostome and vertebrate clades, perhaps drawn to the outgroup by long branch attraction. However, only one copy of a Na^+/K^+ ATPase can be found in the genomes of other Deuterostome phyla (*B. floridae*, *S. purpuratus* and *S. kowalevskii*), and these cluster more conventionally with Na^+/K^+ ATPases (data not shown).

My investigations suggests that the *D. japonica* H^+/K^+ ATPase α subunit referenced by previous work (Nogi et al., 2005) may represent an independent duplication within this species, as the two representatives described in this species group with fair bootstrap support with each other, rather than separately in H^+/K^+ and

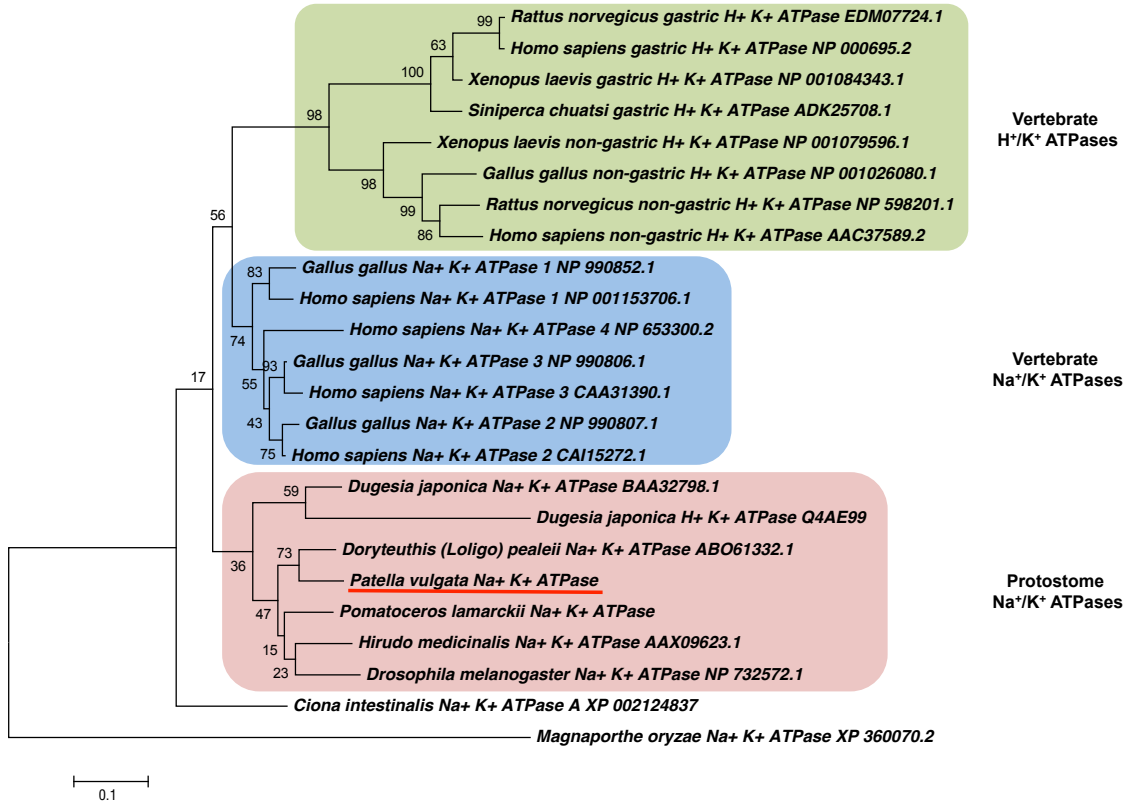


Figure 4.1: *Maximum likelihood ATPase α subunit phylogeny, reconstructed using MEGA5 with 1000 bootstrap replicates under the WAG+4G+i model with all other default prior settings, rooted with* Magnaporthe oryzae *Na^+/K^+ ATPase α subunit sequence. Sequences taken from NCBI with accession numbers as given, or from the datasets described in Chapter 3. Scale bar represents amino acid substitution rate per site. Alignment can be seen in Appendix C, Fig. 4*

Na^+/K^+ clades. Whether these genes pump sodium or hydrogen ions is presently unknown, and will require functional testing to discern. However, given that ancestrally these pumps are sodium channels, this seems more likely.

Analysis of ATPase β subunit phylogeny (Fig. 4.2) also suggests that the H^+/K^+ ATPases are vertebrate innovations, although a monophyletic grouping of Na^+/K^+ ATPases is not seen in my tree, which is unrooted. It is clear from Fig. 4.2 that duplication of β subunits is ancestrally shared by Patellogastropod molluscs.

There are therefore clear homologues of Na^+/K^+ ATPases in the genomes of my lophotrochozoan models. Before going on to functional testing, I examined their expression, particularly to see whether the early asymmetric distribution seen in *X. laevis* was discovered in these organisms.

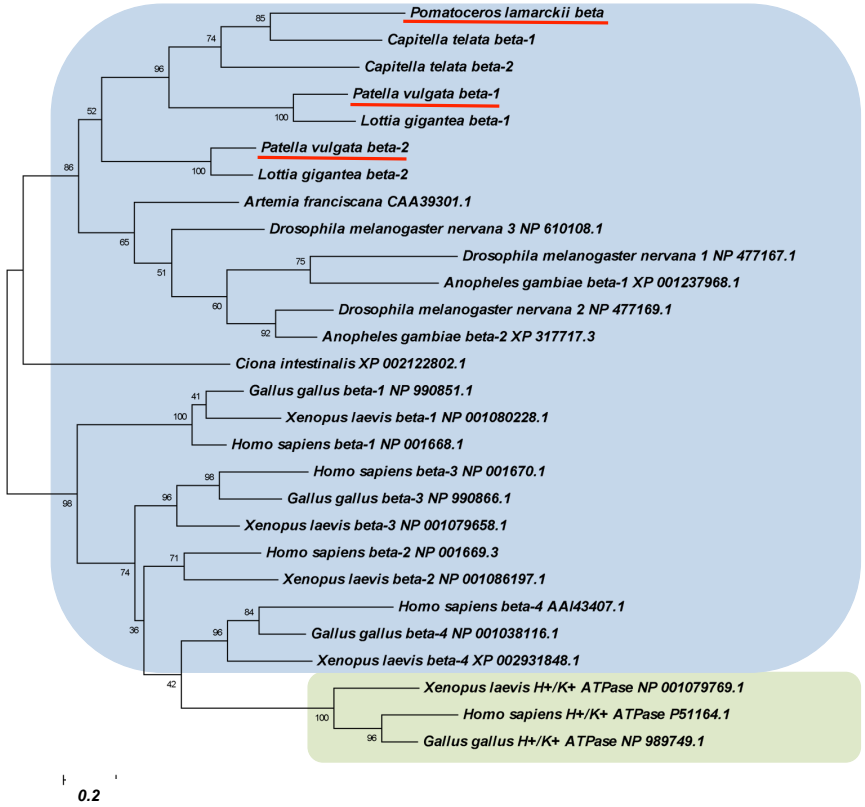


Figure 4.2: *Maximum likelihood ATPase β subunit phylogeny, reconstructed using MEGA5 with 1000 bootstrap replicates under the JTT+4G model, unrooted.* P. vulgata *sequence is underlined for clarity. Scale bar represents substitutions per site. Alignment can be seen in Appendix C, Fig. 6*

4.2.2 (H/Na)$^+$/K$^+$ ATPase Expression

Both *P. vulgata* and *P. lamarckii* ATPase genes have been cloned and probes prepared for RNA *in situ* hybridisation, although those for the latter species have been handed over to another student. Both *P. vulgata* ATPase β subunits have also been subjected to RNA *in situ* hybridisation. Most interesting for the present investigation is the expression of ATPase α subunits, which can be seen in Fig. 4.3.

Fig. 4.3 suggests that *P. vulgata (H/Na)$^+$/K$^+$ ATPase* α mRNA is either maternally provisioned or very strongly expressed from early on in development, the latter being very unlikely. This is confirmed by the result shown in Fig. 4.4, which show that the expression pattern seen mirrors the position of the *P.vulgata* nucleus. This perinuclear expression is also observed at the two cell stage, and in a proportion of four cell embryos. Some four cell stage embryos, such as the one shown in

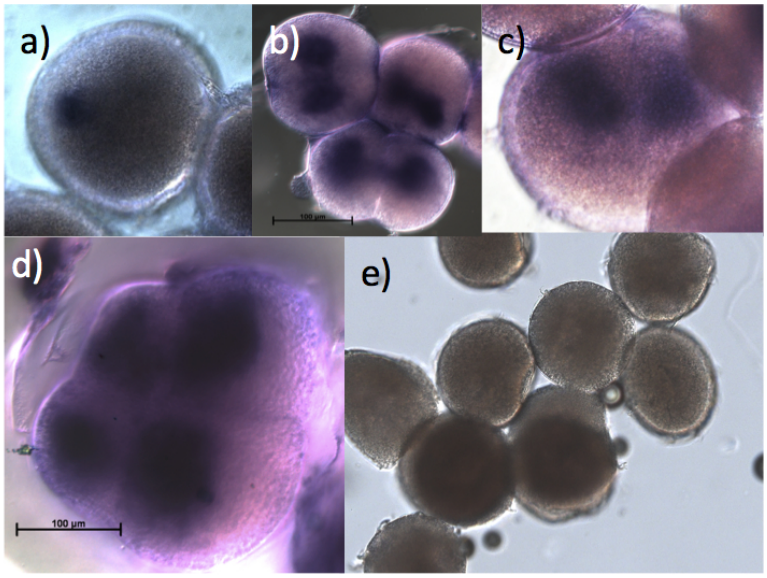


Figure 4.3: *Expression of $(H/Na)^+/K^+$ ATPase α subunit mRNA in* P. vulgata *embryos at a) 30 minutes, b) c) e) Two hours post fertilisation and d) Four hours post fertilisation.(hpf) a) shows a single cell, where mRNA appears to be localised to a point. b) shows a two cell stage embryo, where mRNA is found near the centre of the embryo. c) shows an apparent four cell stage embryo, where expression is found in some embryos only in two cells. d) shows an eight cell embryo, shown side on, with two micromeres visible at left and two macromeres visible at right. e) shows a sense mRNA control image, taken of multiple cells from the same parent stock as c).*

Fig. 4.3c, show expression in only two cells, but upon repetition of this experiment it seems that these are the exception rather than the rule. More generally, expression is found to be perinuclear in all four cells at this stage, and continues to be perinuclear through to the eight cell stage, and thereafter spreads throughout the developing embryo, with no observable localisation at the 16 and 32 cell stage.

At the 32 to 64 cell stage, however, expression begins to coalesce, and expression is eventually only found at one pole of the embryo at the 64 cell stage. This may represent the onset of embryonic expression with the loss of earlier maternal transcripts - the embryos are by this stage at least 10 hours old. For instance, the images shown in Fig. 4.5 show embryos of 13 and 24 hours post fertilisation. The images in Fig. 4.5a and b show late blastula stage embryos, with Fig. 4.5b in particular showing a blastula stage embryo with a macromere (indicated with red arrow) taking up a sizeable portion of the blastocoel, potentially the D lineage cell in the process of being induced. It should be noted that these embryos developed slowly, and the blastula would normally be seen earlier in development than 13 hours post fertil-

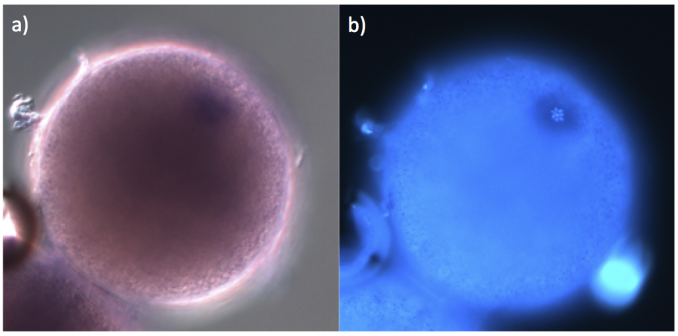


Figure 4.4: $(H/Na)^+/K^+$ ATPase α *mRNA expression (left) and Hoechst stain (right) images of a 30 minute post-fertilisation* P. vulgata *embryo, of the same stock shown in Fig. 4.3a. Note that expression of ATPase* α *mRNA is strongly perinuclear, mirrored by the location of the nucleus in the Hoechst stained image.*

isation (hpf) - this will be a consequence of the ambient temperature being lower than that seen in other samples. Expression of *(H/Na)$^+$/K$^+$ ATPase* α mRNA seems to be at the apical region opposite the ingressing macromeres. By 24 hpf, expression has spread more generally around the embryo, and is found in the shell field, at the apical tuft, and along the ventral and posterior extremities of the embryo. It is, however, absent from the central portion of the embryo as a whole.

(H/Na)$^+$/K$^+$ ATPase $\beta 1$ expression is minimal in embryos of all stages examined, with no staining seen in embryos up to the 8 cell stage (one/two cell stage shown in Fig. 4.6a) and only diffuse, nonspecific staining seen in embryos at the 13 hpf stage (Fig. 4.6b). No specific expression of this gene was seen in any embryo stages examined through to the 24 hpf stage, and I estimate that this subunit plays a role later in the life history of *P. vulgata*.

(H/Na)$^+$/K$^+$ ATPase $\beta 2$ expression, however, mirrors the expression of *(H/Na)$^+$/K$^+$ ATPase* α, especially at later stages. While only extremely diffuse perinuclear expression can be seen in 30 minutes pf and two hpf embryos (Fig. 4.6c, d) strong expression can be seen at one side of the developing blastula stage embryo at around 13 hpf (although, as noted earlier, these embryos developed slowly, and the blastula would normally be seen earlier in development). This is similar to the patterns seen for *(H/Na)$^+$/K$^+$ ATPase* α mRNA, and suggests a functional role for the assembled protein at one end of the developing embryo.

These expression domains do not provide support for the supposition of Levin et al. (2002) in regards to the asymmetric localisation of ATPase transcripts in early

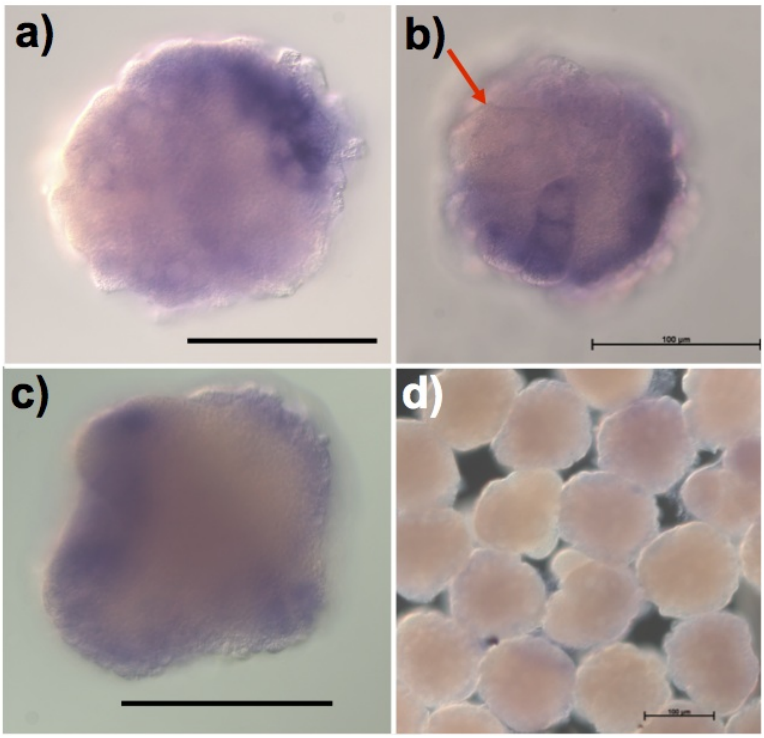


Figure 4.5: $(H/Na)^+/K^+$ ATPase α *mRNA expression in* P. vulgata *embryos at 13 (a, b, d) and 24 (c) hpf as revealed by DIG labelled RNA probes and AP antibody staining. a)* P. vulgata *blastula-stage embryo, with expression localised to one pole. b) Blastula stage embryo, with arrow indicating presumptive ingressing macromere. c) side view of 24 hpf* P. vulgata *embryo, oriented with apical tuft at top right and shell gland at left of image. d) Sense control, 13 hpf* P. vulgata *embryos. Black lines represent scale, 100 microns in length.*

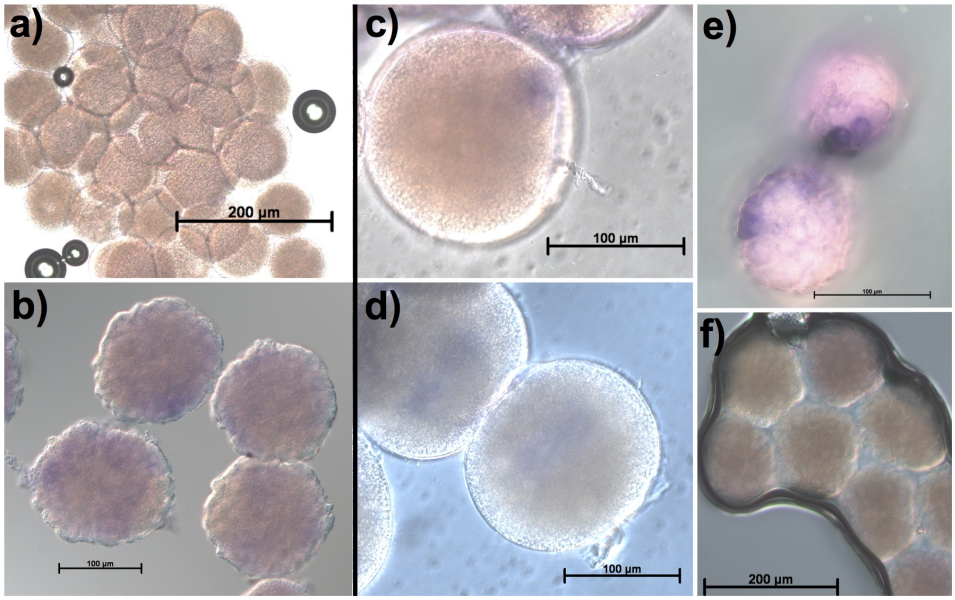


Figure 4.6: *H/Na)$^+$/K$^+$ ATPase β expression in* P. vulgata. *a)* (H/Na)$^+$/K$^+$ ATPase β1 *expression is not detected in early embryos - embryos at two hpf shown. b) Expression in 13 hpf embryos is similarly lacking, with only diffuse nonspecific staining visible. c) Weak perinuclear expression of* (H/Na)$^+$/K$^+$ ATPase β2 *mRNA is seen at 30 minutes pf and d) two hpf. e) at 13hpf, expression of* (H/Na)$^+$/K$^+$ ATPase β2 *is seen at one side of blastula-stage embryos (two of which are shown). f)* (H/Na)$^+$/K$^+$ ATPase β2 *sense control, 13 hpf. Black line represents scale as stated on images.*

embryos. This does not mean that they could not have an effect - for instance, mRNA localisation does not necessarily correlate with protein localisation. Unfortunately there are no antibodies to lophotrochozoan ATPases yet commercially available. If the results of functional testing in this clade, as described in section 4.4 of this report, are seen as sufficiently strong to warrant further investigation, the raising of such antibodies may be worth the cost involved in doing so.

Later expression (at around the blastula stage) suggests that a role may be played by (H/Na)$^+$/K$^+$ ATPases in establishing anterior identity. Both *(H/Na)$^+$/K$^+$ ATPase α* and β mRNA is found at this region, and the resulting protein product could be involved in cellular specification, generation of gradient within the blastocoel, or some aspect of specification of the D lineage macromere and its derivatives, as discussed later in this chapter. Functional work and protein localisation studies are needed to verify this, but it could be an interesting avenue for future research.

4.2.3 Serotonin and Possible Morphogens

A number of possible small molecule morphogens, potentially capable of being differentially distributed around an embryo under the influence of charge differences created by ATPase pump action have been putatively identified in vertebrates, including serotonin, inositol polyphosphates and Ca^{2+} ions, and retinoic acid, which may perform an upstream role. To test for a potential role for these in lophotrochozoan L/R asymmetry establishment, experiments were performed using immunohistochemistry.

Serotonin

Immunohistochemistry work against serotonin (using Abcam ab10385, with the protocol described in Chapter 2) has been performed in both *P. vulgata* and *P. lamarckii*, although results have been somewhat inconclusive. Figure 4.7 shows the results of experimentation on *P. vulgata* embryos through to the four cell stage of development.

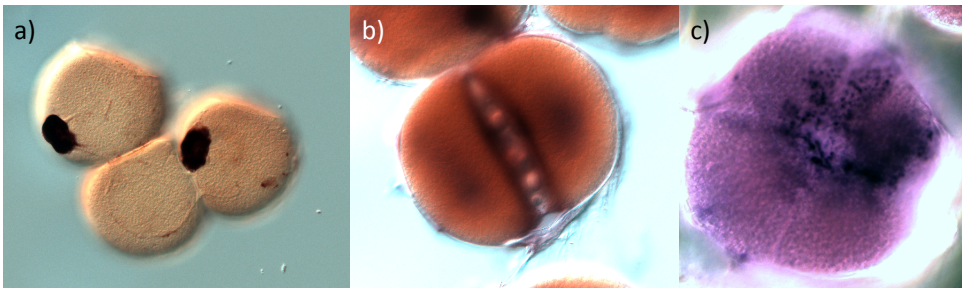


Figure 4.7: *Putative serotonin localisation in* P. vulgata *embryos. a) In single cell-stage embryos, serotonin is only found in polar bodies and extracellular tissue. b) By the two cell stage, serotonin appears to be concentrated around the point of cytokinesis. c) in a proportion of four cell stage embryos, serotonin appears to be concentrated asymmetrically, with bias towards a single cell. After this time point, expression of serotonin is found across the entirety of the developing embryo.*

Whether patterns are the result of transient changes in distribution (which could be causative) or are instead purely the result of artefacts of the staining process is difficult to discern. Only some embryos within any given sample show the patterns seen in Fig. 4.7c. Unless consistent results are gained it is unlikely that this work will reach a publishable standard. *P. lamarckii* work in this regard has also

now been performed by another student in the Shimeld laboratory, which provides a possible point of comparison, should consistent staining patterns be observed in that species.

4.2.4 ATPases and Asymmetry

The results of expression-based analysis give little support to the ion channel hypothesis as proposed in Levin et al. (2002). I cannot draw any conclusions as to a role for either ATPases or serotonin in the establishment of asymmetry based on this evidence alone. However, as noted earlier, the presence of pharmacological inhibitors of these molecules also allowed us to test these functionally *in vivo*. These results are described further in section 4.4, but in order to assess whether asymmetry had formed correctly, I required a positive read-out of this process. To provide this, as well as for a range of further insights, I studied the expression of two primary markers of symmetry breaking in other species- Nodal and Pitx expression.

4.3 The Initial Signs of Symmetry Breaking - *Nodal* and *Pitx* Expression

As well as performing a range of roles in patterning the developing embryo, *Nodal* and *Pitx* expression represents the first sign of molecular asymmetry described in mollusc embryos (Grande & Patel, 2009), which generally do not display morphological L/R asymmetry until the onset of torsion, several days after fertilisation. I set out to describe the normal expression of these genes, to ensure that the expression described previously by Grande & Patel (2009) is not restricted to the species examined in that work, to determine the normal mechanism of establishing asymmetry in my model species, and to provide a robust read-out of the correct establishment of asymmetry.

4.3.1 Cloning of *Nodal* and *Pitx*

The lophotrochozoan expression of the key signalling genes *Nodal* and *Pitx* are still only known to science from their original description in two species in Grande &

Patel (2009). It was therefore important to confirm that expression in my model organisms was not significantly different from that described in that paper. The first step in doing this was to clone the genes from my species, a process which was not trivial.

P. lamarckii

Initial work on cataloguing the lophotrochozoan Nodal signalling pathway for this thesis was based on the use of degenerate PCR to clone the *Nodal* and *Pitx* genes of *P. lamarckii*. Initial degenerate PCR oligo sequences were provided by Dr. Cristina Grande (Grande, C, personal communication), and additional oligos designed by alignment and comparison with known *Nodal* and *Pitx* gene sequence from other species. PCRs were performed as described in the Methods section of this thesis. Fragments of the coding sequence for *Pitx* were acquired by this method, but *Nodal* was not amplified and cloned, despite repeated attempts. RACE PCR was then performed to determine sufficient coding sequence for the construction of DIG labelled RNA probes for the *P. lamarckii Pitx* gene. This proved successful, and gene fragments of sufficient length were cloned and probes prepared for RNA in *in situ* hybridisation as described in Chapter 2.

The advent of sequence data as described in Chapter 3 simplified matters immensely. My assemblies of transcriptomic data did not provide a *P. lamarckii Nodal* homologue, but by tBlastn comparison of known *Nodal* sequence with that of the 51bp paired end read data I was able to identify a likely *P. lamarckii Nodal* homologue. This was cloned and the resultant 270 bp fragment identified by blastx as a clear *Nodal* homologue, spanning the characteristic 3′ region of the *Nodal* molecule. RACE PCR was performed to determine the full coding length of the mRNA for this gene, allowing the production of probes for RNA *in situ* hybridisation. 3′ RACE also revealed the presence of several potential alternative splice variants,. Another student in my laboratory has taken on responsibility for *P. lamarckii* RNA *in situ* hybridisation work, which is therefore not reported in the present thesis.

P. vulgata

The *Nodal* genes of *P. vulgata* share some similarity with its published *L. gigantea* homologue, and a portion of the sequence of one *Nodal* homologue had been cloned by degenerate PCR prior to my arrival in the laboratory. Before the advent of genomic data for this species, I undertook RACE PCR to determine more of the coding sequence of this gene. The *P. vulgata Pitx* homologue had similarly been identified prior to my arrival in the laboratory from a sample amplified usingRT PCR and degenerate primers, and from a RNA sequencing experiment eventually published as Werner et al. (2012), and was already cloned upon my arrival.

After completion of transcriptomic and genomic work, it became apparent that *Nodal* is present in two copies in *P. vulgata*. For phylogenomic confirmation of the identity of these Nodal homologues, please refer to Fig. 5.4. A portion of the sequence of this additional *Nodal* gene, henceforth referred to as '*Nodal 2*' (with the earlier-discovered gene dubbed *Nodal 1*) was also cloned for use in further experimentation as described in the Methods section of this thesis.

4.3.2 *Nodal* Expression

The *Nodal* duplication event is of great interest, as subfunctionalisation or neofunctionalisation can result after duplications, and any division of the role of *Nodal* in Molluscs could ease functional dissection of the roles of this gene in the Lophotrochozoa. Preliminary insight into the role of *Nodal* in *P. vulgata* was provided by RNA *in situ* analysis of mRNA expression, as shown in Figs. 4.8 and 4.9.

The first figure shown here, Fig. 4.8, shows that the *Nodal 1* gene may be responsible for the establishment of asymmetry in *P. vulgata* embryos. The patterns seen are similar, although not identical, to those noted in *L. gigantia* by Grande & Patel (2009). While not shown in my figure, experiments in my laboratory have shown expression begins in *P. vulgata* in a single cell at the 32 cell stage, before spreading to another adjacent cell (V. Fernandes, personal communication). This is identical to the onset of expression in *L. gigantia*, where expression of the single *Nodal* homologue is first found in the micromeres 1c and 2c. Later expression diverges from

this common beginning.

In *L. gigantia* at the trochophore stage, "there are two ectodermal domains of nodal expression on the right side of the embryo: one cephalic region plus a lateral domain near where shell formation initiates" (Grande & Patel, 2009, p. 1008). In *P. vulgata*, in contrast, expression is initially found more generally throughout the right-hand side of the embryo, before fading to more restricted domains. This can be seen most clearly in Fig. 4.8e, where expression is seen in large patches in three locations on the trochophore. By 24 hpf, in Fig 4.8f, expression has narrowed to a point at the presumptive cephalic region. It has, however, been reported by other members of my laboratory that expression can mirror that seen in *L. gigantia* at the 24 hpf stage, and the differences in this may reflect slight differences in developmental stage between the samples used for expression analysis.

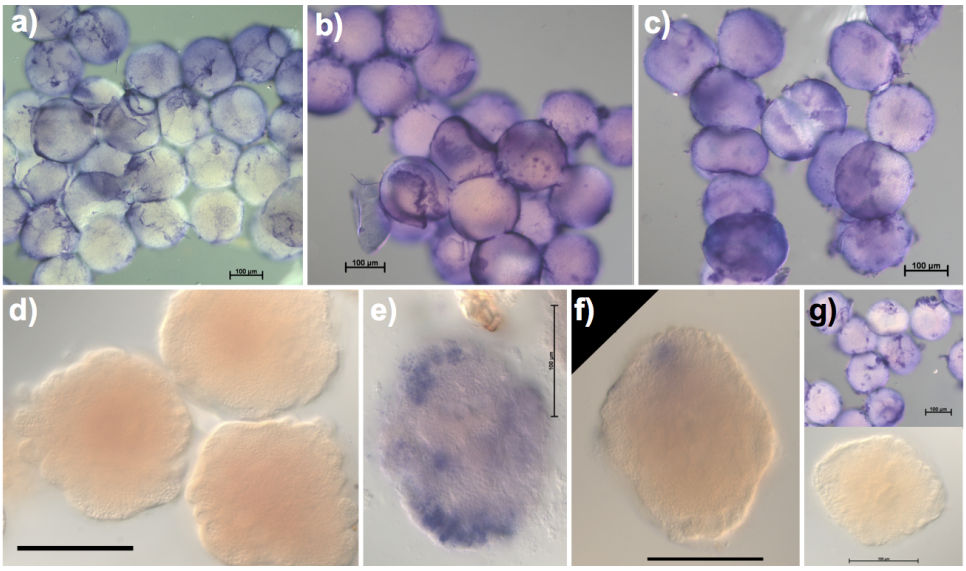


Figure 4.8: Nodal 1 *expression in* P. vulgata. *a) No specific expression is seen in unfertilised eggs, even when over-exposed. (Expression seen is on the outside of all embryos, and results from trapping of probe - this can also be seen in the sense control, g) b) Embryos 30 min pf and two hpf (c) also exhibit no consistent signs of expression, even when over-exposed. d) by the 32 cell stage, at approximately 10-13 hpf, no expression can be seen in these embryos. See text for discussion. e) at the 18 hpf stage, expression is found in the right hand side of the apical region, near right and centre of the 'foot' of the developing embryo, and in a right-hand sided, central location. f) At 24hpf expression has faded in many areas, but is still observed to the right of the apical region. g) sense controls at two hpf (top) and 24 hpf (bottom) show no specific staining.* P. vulgata *in e) and f) oriented with apical tuft uppermost and mouth oriented toward the reader. Scale bars represent 100 microns in length.*

The expression seen in the *Nodal 1* paralogue therefore seems to corroborate in general the findings seen in Grande & Patel (2009). A similar role for this gene in establishing L/R asymmetry seems likely, if yet to be confirmed by functional analysis.

In contrast with my findings for the *Nodal 1* paralogue, the *Nodal 2* gene of *P. vulgata* seems to show a markedly different expression pattern. This second *Nodal* paralogue shows expression from early embryonic stages (Fig. 4.9a), but this expression is perinuclear from the single cell stage through to the eight cell stage (Fig. 4.9b, c). By the 32 cell stage this expression (Fig. 4.9d, uppermost) has faded, which, together with the expression even in unfertilised eggs, suggests strongly that this early expression represents maternal deposition of mRNA transcripts. Embryonic expression then begins in earnest, with strong expression in a single cell at the 32 cell stage (Fig. 4.9d, lower image) and symmetrical expression in many regions of the body through the trochophore stage (Fig. 4.9e).

Both the maternal and symmetrical trochophore stage expression of this gene represents a novelty compared to the expression described in Grande & Patel (2009). Maternal deposition of *Nodal2* could suggest that in this species this gene plays a role in specifying a portion of the early processes of development in this species, but to confirm this protein localisation should be observed, as there are several notable examples of mRNA and protein localisation being drastically different in embryonic development (for example, *bicoid* in *D. melanogaster*).

These expression patterns suggest that the *Nodal 1* gene plays a role in the establishment of L/R asymmetry, but the role of *Nodal 2* is more difficult to define. The contrasting apparent binding sites seen in these loci (see Chapter 5) will be of interest for functional dissection of genomic regulation of these two genes. It should be noted that *L. gigantia Nodal* could have lost ancestral domains of expression itself, with *Nodal 2* expression mirroring that found ancestrally in other molluscs - only wider phylogenetic sampling in the Mollusca in particular and Spiralia in general will reveal if this is the case.

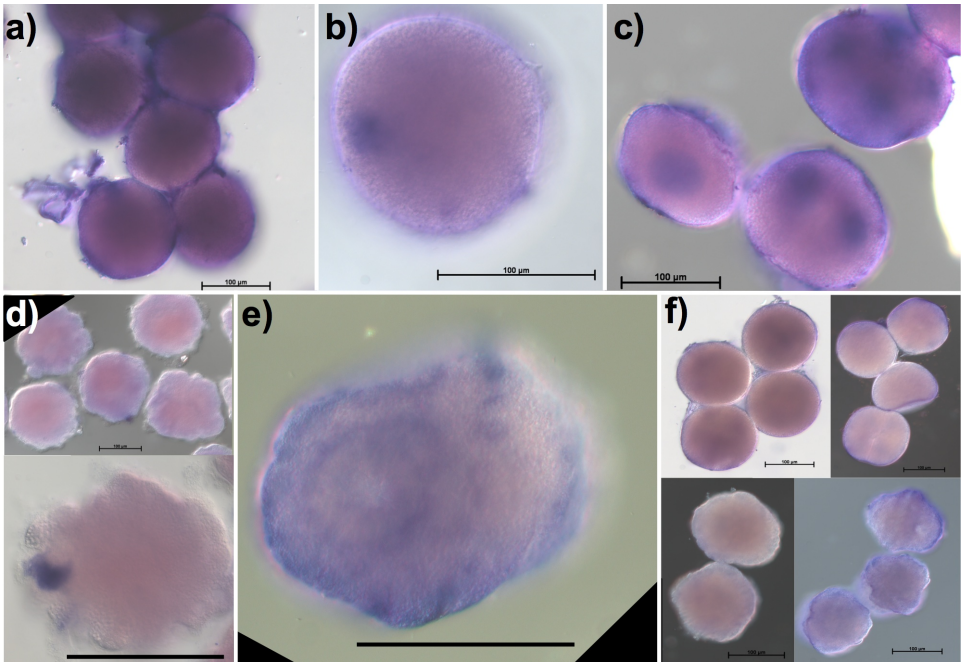


Figure 4.9: Nodal 2 *expression in* P. vulgata *embryos through to 24 hpf. a) expression is found perinuclearly in unfertilised eggs. Similar expression can be seen through the one (b, 30 minutes pf) two and four cell stage (c, two hpf) and begins to fade, and is eventually lost at around the 32 cell stage (d, uppermost, 10 hpf). However, expression is found in a single cell towards the end of the 32 cell stage (d, lower). By 18 hpf (not shown) expression is symmetrical, found around the developing shell gland in two concentric rings, in a line across the trochophore, and in apical and 'foot' regions of the embryo. e) shows this pattern in a 24hpf* P. vulgata*, oriented with apical tuft at far left and shell gland pointing toward the reader. f) shows sense controls, clockwise from top left unfertilised, two hpf, 24 hpf, 10 hpf. Scale bars represent 100 microns.*

4.3.3 *Pitx*

To confirm the identity of my cloned Pitx product, phylogenetic analysis was carried out, as can be seen in Fig. 4.10. This figure shows the presence of homologues of *Pitx* in a number of lophotrochozoan species, and this gene appears to be highly conserved, both in sequence identity (note short branch lengths) and in presence even in species which have undergone a high degree of gene loss (for instance, in *C. elegans*, whose *unc 30* gene is a clear *Pitx* homologue).

The conservation of *Pitx*, which performs a number of roles in neural development as well as in L/R asymmetry establishment, across the Bilateria is perhaps not surprising, and it will be interesting to see if its role in asymmetry is conserved more generally across the Lophotrochozoa. The *P. vulgata Pitx* homologue is found embedded within a clade containing other known *Pitx* sequences and with rela-

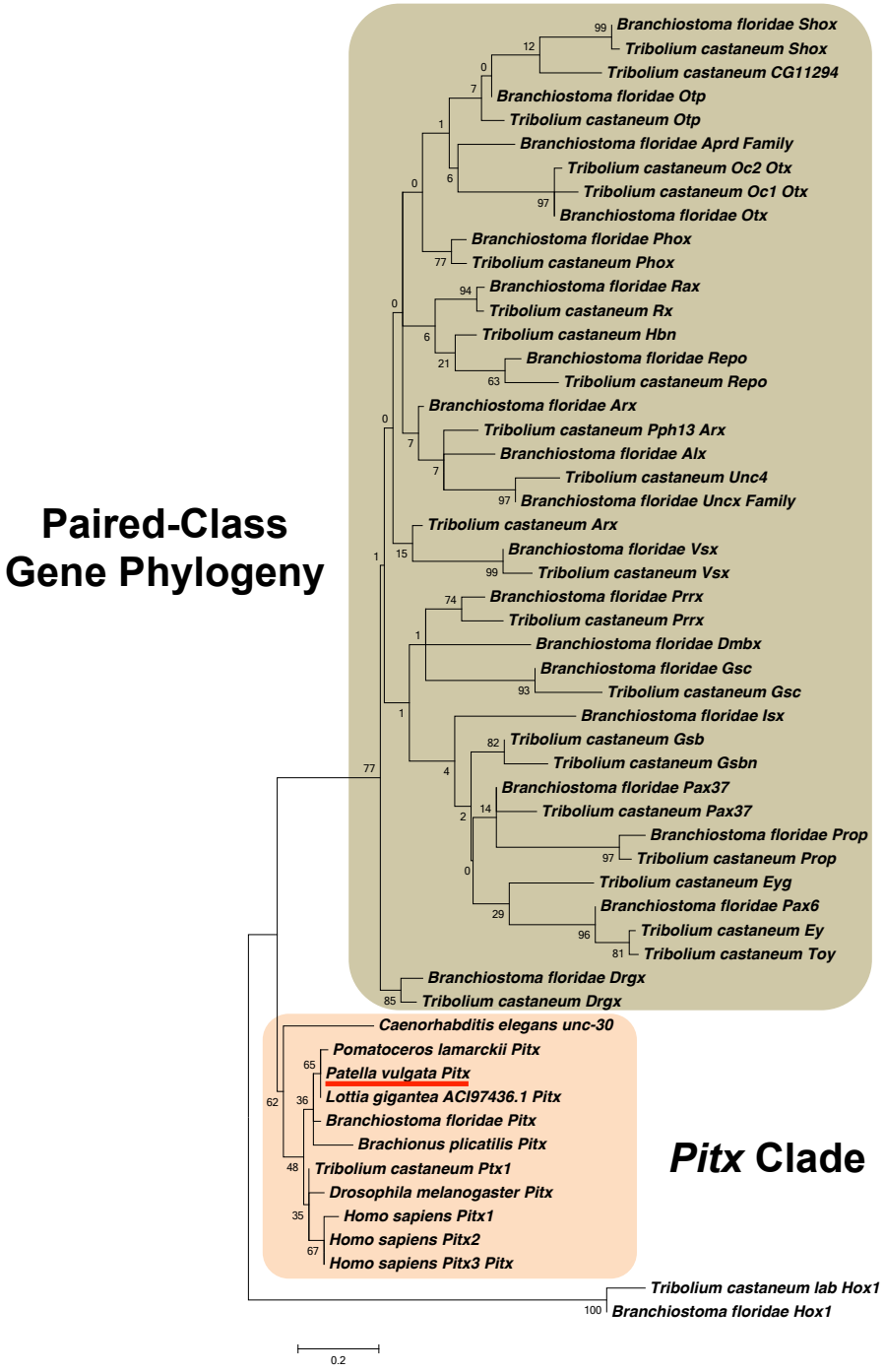


Figure 4.10: *Maximum likelihood* Pitx *and other Paired-class homeodomain containing gene phylogeny, reconstructed using MEGA5 with 1000 bootstrap replicates under the WAG+4G model, rooted with Hox 1 genes from* B. floridae *and* T. castaneum. *Homeodomain gene sequences from HomeoDB, with the exception of* L. gigantia *sequence, from NCBI, and* P. vulgata, P. lamarckii *and* B. plicatilis *sequences, taken from resources described in Chapter 3. Scale bar represents substitutions per site at given length of branch. Alignment can be seen in Fig. 4.11.*

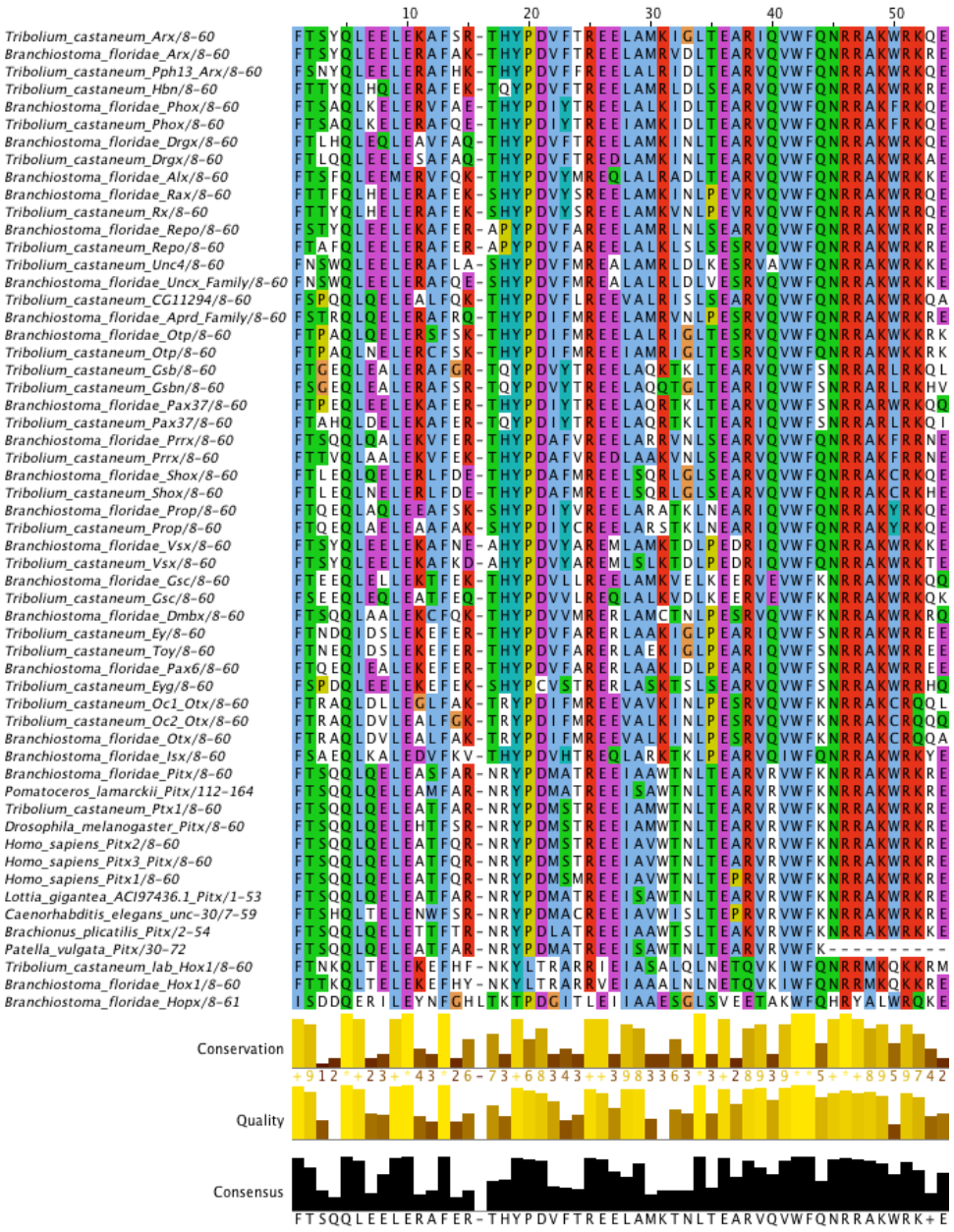


Figure 4.11: Pitx *Gene Alignment Used in Fig 4.10. Aligned in MAAFT under the L-INS-i model as described in Methods, and displayed using Jalview with ClustalX colouring.*

tively high (65) bootstrap support given the short alignment length. I can therefore be confident that this sequence represents the *P. vulgata Pitx* gene.

4.3.4 *Pitx* Expression

My probes for *P. vulgata Pitx* proved to be excellent, and the results of my investigations into *Pitx* expression at a range of developmental stages can be seen in Fig. 4.12. I focussed my investigations on the first 24 hours after fertilisation, as after this point L/R asymmetry has already been established in these embryos.

No early expression is observed, through until the 32 cell stage. Fig. 4.12b shows the first signs of *Pitx* expression at the 32 cell stage. It should be noted that the stock of embryos used for these *in situ* hybridisation experiments developed exceptionally fast, likely due to high temperatures in the laboratory after the cooler ceased to operate, and the 32 cell stage normally occurs several hours later in development. This expression of *Pitx* appears to be earlier than that detected in *L. gigantia* (Grande & Patel, 2009), where expression is first seen at the 64 cell stage. It may be that my probes, which performed robustly at many stages of development with little background, were able to detect weak expression earlier than those used in the previous study.

Strong expression in 32 and 64 cell stage embryos is followed by very strong expression in early trochophore stage embryos. This mirrors that seen in Grande & Patel (2009, p. 1008), where in *L. gigantia* "*Pitx* expression is seen in a group of ectodermal cells on the right side of the larvae, adjacent to those that express nodal, as well as in the developing gut". *Pitx* therefore seems to have a conserved role in patterning the gut and the ectoderm on the right-hand side of the embryo. Fig. 4.12e, where *P. vulgata* is shown in profile view, shows the expression of *Pitx* in the developing gut, and potentially, in some parts of the shell field.

Pitx does not seem to be at all maternally provisioned, and expression begins within the developing embryo at approximately eight hours post fertilisation, at the 32 cell stage. The expression at this stage seems to mirror that seen in Grande & Patel (2009). However, it must be noted that this is based on an assumption of cellular homology, as without prominent polar lobes I am unable to infer the

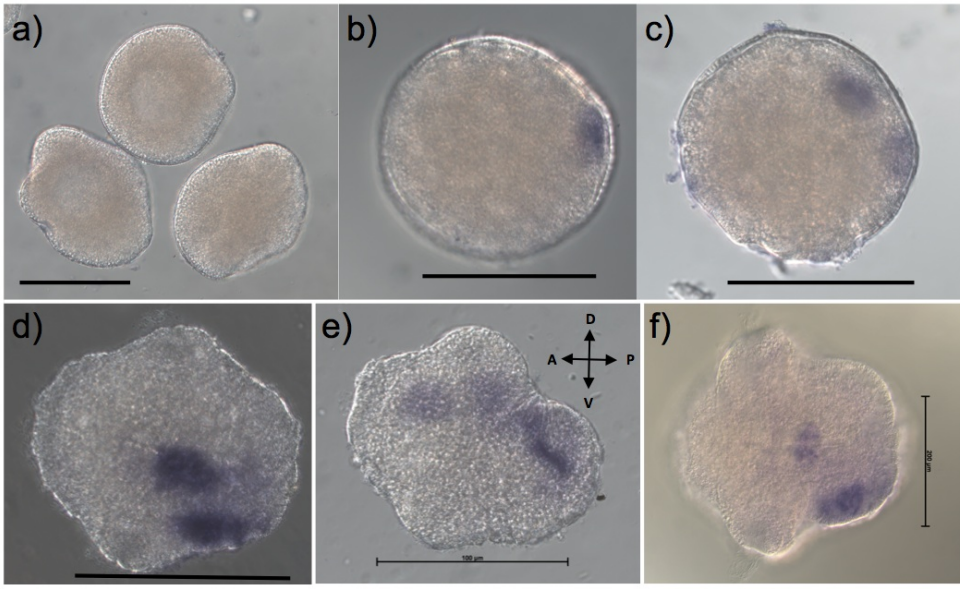


Figure 4.12: Pitx *mRNA expression in* P. vulgata *Embryos. Expression shown in embryos a) immediately after fertilisation (0 hpf), where no specific expression is observed, b) at a 32 cell stage embryo (8 hpf), where expression is seen progressively in one and c) then two cells (in a 64 cell stage embryo). d, e) at 16 hpf, where expression is seen in a central domain and at the right hand side of the developing emberyo, towards the shell field, and f), in 24 hpf embryos, where domains have narrowed compared to that seen at 16 hpf, but remain in approximately the same position on the embryo. No expression is observed in sense controls (data not shown here - see Fig. 4.15 for an example). All embryos shown with anterior to the left and mouth uppermost with the exception of the 16 hpf side view, with orientation as indicated on that image. Scale bars represent 100 microns of length.*

absolute orientation of the embryo before the formation of the mouth and (more easily) when the shell gland is established.

Functional proof of the role of *Pitx* in the Lophotrochozoa is still completely absent from the literature, and some form of knock-down of function would provide more evidence for the downstream role of this gene in this clade. Some form of knock-down, followed by analysis of asymmetrical growth similar to that found in Kurita & Wada (2011) would prove the role of *Pitx* and its connection to *Nodal*, and would be well worth doing in the future.

4.3.5 Summary, *Nodal* and *Pitx* Expression

In general, the expression of *Nodal 1* and *Pitx* seem to mirror that seen in *L. gigantia* and previously reported. Differences, when they exist, are slight and difficult to discern from natural variation in morphology. However, the presence and expression of the *Nodal 2* paralogue alters the landscape somewhat. Significant functional investigation is required to disentangle the relative roles of *Nodal 1* and *Nodal 2*, but the duplication could have significant consequences for the development of *P. vulgata*.

The lack of maternal provisioning of *Pitx* suggests that any role for *Nodal 2* early on in development must be independent of the *Pitx* transcription factor. The patterning of the mesoderm by Nodal has, in other species, been shown to be independent of Pitx, and it is possible that the *Nodal 2* gene begins to perform this role before *Nodal 1* is expressed. It is also possible this *Nodal 2* gene acts to impair the signalling of other ligands, by binding competitively to receptors without activating any other signalling pathway (as *Pitx* is not present to act as a downstream signal).

Maternally provisioned *Nodal 2* could also activate *Nodal 1*, as Nodal protein is self-activating via Smads. If any differential distribution of Nodal 2 protein was shown in early stage embryos, this could represent an earlier, symmetry breaking event in this species. However, it is unlikely such a role would be ancestrally shared within Bilateria, or even in the Mollusca - very few species have duplicated *Nodal* genes, and maternal deposition itself would be a novelty. This is, however, purely

supposition at this stage, and far more investigation is needed before firm conclusions can be drawn about the roles of these genes in *P. vulgata*, or the Lophotrochozoa more generally.

4.3.6 Early Markers of Asymmetry

As in *L. gigantia*, both *Nodal* homologues appear to be the earliest signs of asymmetrical expression of a transcript in developing embryos, with my data (*Nodal 2* and data in our lab (*Nodal 1*) showing the presence of asymmetrically distributed transcript as early as the blastula stage. In *P. vulgata*, unlike in *L. gigantia*, *Pitx* is also present in the 32 cell stage embryo, although double *in situs* or carefully timed RT PCR would be needed to determine whether the expression of either *Nodal* precedes that of *Pitx* in this cell. The lack of characteristic markers of a D lineage, such as polar lobes or a larger D cell, however, meant that interpreting the localisation of this staining relative to the future left and right hand side of the body would be difficult.

I wished to discern a clear and reproducible signal for the correct establishment of L/R asymmetry, for use in determining whether functional perturbation had resulted in asymmetry defects. The expression of both *Nodal* and *Pitx* homologues were considered as such markers, as there is no outward morphological sign of the correct establishment of asymmetry in early *P. vulgata* embryos, and the correct performance of torsion can be difficult to interpret, especially as faulty torsion is likely to be deleterious to continued growth.

I decided that the expression of *Pitx* at 24 hpf represented the best read-out of the normal establishment of asymmetry for my purposes, despite the lack of a causal link between *Nodal* and *Pitx* expression in my species. The strong, robust expression and lack of background seen in my *Pitx in situ* hybridization experiments gave us confidence in its ability to detect changes in expression relative to morphology and, even if *Pitx* is not immediately downstream of Nodal signalling in my species, it was a reliable read-out of established asymmetry in any case, being found reliably on the right in embryos examined at the 24 hpf stage.

4.4 Functional Testing of Symmetry Breaking Events

The discovery of a functional role for ATPases in the establishment of L/R asymmetry in a member of the Lophotrochozoa would represent the first evidence of a wider conservation of mechanisms for establishing asymmetry upstream of the Nodal cascade outside of the Deuterostomia. Both Na^+/K^+ and H^+/K^+ ATPases have been implicated in the control of L/R asymmetry establishment. As seen in section 4.2.1, lophotrochozoans seem to possess Na^+/K^+ ATPases, which may ancestrally play the role performed by H^+/K^+ ATPases in L/R asymmetry in some vertebrates.

4.4.1 $(H/Na)^+/K^+$ ATPase Transporter Inhibitors

Before using functional inhibitors of ATPase activity, it was vital to consider whether target sites for inhibitors was present. The inhibitors I aimed to use, lansoprazole, omeprazole and ouabain, target different sites within functioning ATPases. This diversity of mode of action is important, as corroborating results from each of these would reinforce my findings. Pharmacological inhibitors are noted for causing a range of off-target effects, which could obfuscate my analysis, but by using a range of inhibitors on independent samples, any concordant results would suggest strongly that a specific L/R asymmetry defect was being caused by these drugs.

Omeprazole binds covalently to the α subunit of H^+/K^+ ATPases (Mukherjee et al., 2001; Jiang et al., 2002) and halts their activity. The covalent binding site of omeprazole to ATPases has been narrowed down to a small portion of the α subunit, between the TM5 and TM6 transmembrane domains (Lambrecht et al., 1998; Munson et al., 2000). This site contains one or more cysteine residues, to which omeprazole can attach itself. The sequence of this area in $H/Na^+/K^+$ ATPases in humans, *Ciona intestinalis* and the two protostome species studied in this work are shown in Fig. 4.13. Not all ATPases possess the requisite cysteine residues in this area, but it is clear that the two protostome species possess the same cysteine residues available to bind omeprazole as *Ciona intestinalis*, in which this drug has already been shown to have an effect (Shimeld & Levin, 2006). Lansoprazole also

attaches to these residues, but has a slightly altered structure and changed chemical properties as a result.

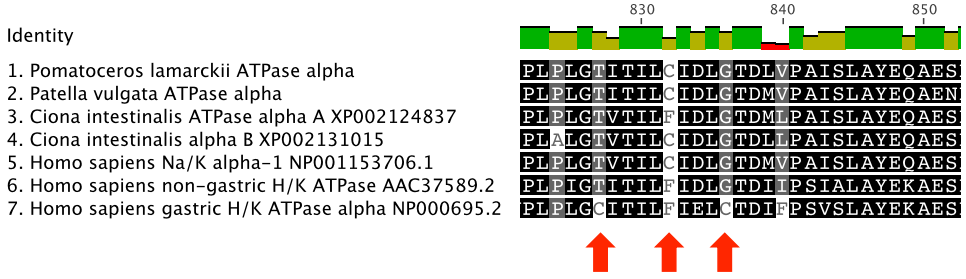


Figure 4.13: *MAFFT alignment of* P. lamarckii, P. vulgata, Ciona intestinalis *and* Homo sapiens *ATPase α subunits, across the active site and M6 domain. Red arrows indicate locations of Omeprazole target cysteines, where these exist.*

Ouabain targets Na^+/K^+ ATPases, and has previously been shown to bind to squid (*Loligo opalescens*) homologues of these proteins between the TM1, TM2, TM4, and TM6 transmembrane domains Sandtner et al. (2011). As stated in that paper, the structure of these domains is extremely well conserved in Na^+/K^+ ATPases across the Bilateria, and ouabain is likely to target these genes in the same place in all species that possess them.

By comparing the results of inhibition of these three drugs, it was hypothesised that I could gain clear evidence for or against a role for $(H/Na)^+/K^+$ ATPases in establishing asymmetry. First, however, I checked the effective exposure levels of these drugs, to avoid off-target and broadly teratogenic effects whenever possible.

4.4.2 Pharmacological Inhibition - Titration

To confirm any potential role for $(H/Na)^+/K^+$ ATPases in the establishment of asymmetry, developing *P. vulgata* were treated with ion channel inhibitors at a range of concentrations until 24 hpf stage, before being fixed, as described in Chapter 2. To ensure statistical power in assessing the results of this experiment, treatments were repeated five times, each using separate individual male and female *P. vulgata* as sources of gametes, to ensure that results are biologically replicable and consistent.

When using inhibitors such as omeprazole, teratogenic effects will occur above

a certain threshold. To ensure that the correct quantity of omeprazole was used as a treatment, a variety of concentrations were assayed, and their effects on development contrasted, with absolute numbers given in Table 4.1 while relative proportions can be seen in Fig. 4.14. The numbers of positively phototactic/negatively geotactic embryos represents a good shorthand for normal development, as these embryos will have acquired the capacity to detect and respond to environmental influences, while abnormally developing embryos cannot swim correctly in orientation to these cues.

	1			2			3			4			5		
Omeprazole	Photo/Geotactic	Defects	Undeveloped	Photo/Geotactic	Defects	Undeveloped	Photo/Geotactic	Defects	Undeveloped	Photo/Geotactic	Defects	Undeveloped	Photo/Geotactic	Defects	Undeveloped
0 µg/mL	25	30	14	25	17	12	27	29	4	18	31	16	22	28	3
20 µg/mL	15	22	8	25	29	22	8	33	21	15	17	5	15	32	15
40 µg/mL	54	38	22	20	46	16	22	42	10	26	48	18	17	23	9
80 µg/mL	10	31	19	29	103	35	20	62	21	14	52	23	8	38	31

Table 4.1: *Figures from the use of omeprazole on inhibiting the development of P. vulgata embryos. Three columns give number of embryos exhibiting normal (positively phototactic, negatively geotactic), abnormal development resulting in an inability to swim and respond to light/gravity, and no development at 24 hpf for each of five independent fertilizations.*

The difference between the numbers of positively phototactic/negatively geotactic embryos and abnormally developing embryos in my samples was compared statistically. Significantly poorer developmental outcomes (χ squared test of independence, probability of observation by chance alone, $2.98e^{-04}$, $2.45e^{-05}$ and $8.34e^{-11}$ respectively) were observed in embryos exposed to 20 μg/mL, 40 μg/mL and 80 μg/mL omeprazole compared to DMSO-only treated controls (omeprazole is stored in DMSO, which must therefore be provided in controls for proper comparability). This is similar to results seen in other species. No significant difference ($p = 0.0243$) is observed between 20 μg/mL and 40 μg/mL, but further significantly impaired outcomes exist between 20 and 80 and 40 and 80 μg/mL ($5.76e^{-08}$ and 5.07684^{-09} respectively).

This suggests that as little as 20 μg/mL omeprazole is necessary to interfere with normal embryonic development in *P. vulgata*. All treatments were, however, subjected to RNA *in situ* hybridisation of *Pitx* mRNA expression, in order to gain the maximum possible amount of data, for reasons that will soon become apparent.

I wished to repeat this analysis for the two further ion channel binding drugs mentioned earlier, lansoprazole and ouabain. Unfortunately for my ongoing exper-

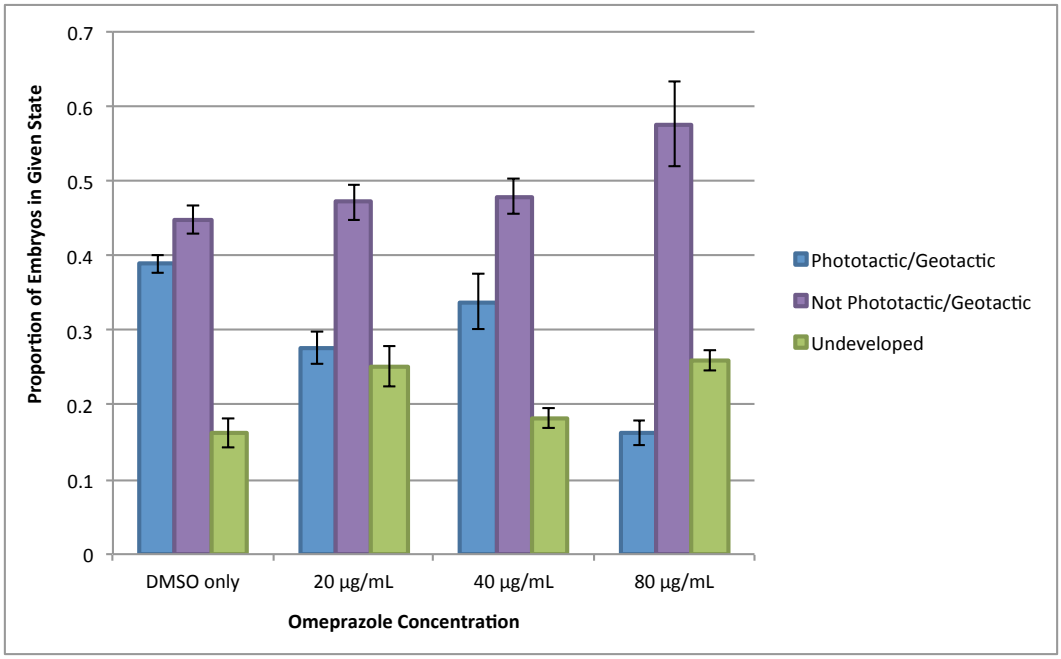


Figure 4.14: *Proportion of* P. vulgata *embryos exhibiting normal (positively phototactic, negatively geotactic), abnormal and no development at 24 hours post fertilisation in the combined values of five replicate trials at different omeprazole concentrations as listed. Error bars represent one standard deviation from the average proportion.*

imentation, the 2012 *P. vulgata* breeding season was particularly poor. The reasons for this are still unknown, but the breeding success of many wild-caught animals was compromised by extreme levels of rainfall in the UK, and particularly in the southwest - the second wettest year on record. This resulted in low salinity in the bays where *P. vulgata* were collected, and may have delayed the maturation of gametes in my model species. In any case, the fertilisations used for ouabain and lansoprazole functional tests proved to be inadequate.

	1		2		3		4		5	
Ouabain	Photo/Geotactic	Defects	Photo/Geotactic	Defects	Photo/Geotactic	Defects	Photo/Geotactic	Defects	Photo/Geotactic	Defects
0 µg/mL			13	35	1	14	8	30	4	17
6.25 µg/mL			6	46	0	22	3	26	9	24
12.5 µg/mL			0	39	2	13	5	15	1	12
31.25 µg/mL			0	13	0	13	2	17	0	14
62.5 µg/mL	EXTENSIVELY DEFORMED									
Lansoprazole	Photo/Geotactic	Defects	Photo/Geotactic	Defects	Photo/Geotactic	Defects	Photo/Geotactic	Defects	Photo/Geotactic	Defects
0 µg/mL					0	20	2	11	4	8
20 µg/mL					1	14	3	17	2	17
80 µg/mL					3	11	2	19	4	12
160 µg/mL					1	13	1	14	4	16

Table 4.2: *The results of experimentation using lansoprazole and ouabain to inhibit the development of P. vulgata embryos. Numbers of unfertilised or non-developing gametes not shown, as this was in triple figures for all samples.*

The results of fertilisations treated with lansoprazole and ouabain can be seen

in Table 4.2. Several fertilisations resulted in no viable embryos whatsoever at 24 hpf. Others resulted in only a few positively phototactic and negatively geotactic embryos being observed at this time point. While there is empirically a difference in the number of properly-developed embryos in the treated and untreated ouabain samples, numbers are not sufficient to make this statistically significant. In the lansoprazole dataset no difference can be detected, either empirically or statistically. This is disappointing and stymied further work. These experiments will be repeated in the Shimeld laboratory in the near future, with the aim of gaining sufficient positively phototactic and negatively geotactic embryos for *in situ* hybridisation analysis.

The expression of *Pitx* in the omeprazole treated samples was, however, compared and quantified. While, without the variety of inhibitors desired, off target effects are a possibility, if ATPases control the normal patterning of L/R asymmetry, I would expect to see significantly more ectopic or reversed expression of these genes in treated samples than untreated samples.

4.4.3 Pharmacological Inhibition - Expression Results

All positively phototactic and negatively geotactic embryos from the omeprazole treatments were subjected to *in situ* hybridisation, and the raw numbers of embryos recovered for each fertilisation and their condition can be seen in Table 4.3. While I expected to see ectopic expression of *Pitx* or *situs inversus*, in a proportion of embryos which increased as omeprazole concentration raised, a radialized phenotype was more prominently observed in embryos as omeprazole concentrations increased. These embryos completely lacked a mouth and shell field at 24hpf. Also noted were some embryos in which no expression was seen at all, and some where morphology was so disrupted by mechanical forces during the *in situ* hybridisation analysis that no conclusions could be drawn as to symmetry.

While I expected some attrition due to loss in the course of the *in situ* hybridisation protocol, and this is seen in most samples, a few samples, marked with a '$\wedge$' in Table 4.3, seem to have more embryos than Table 4.1 would suggest possible. This is likely the result of other, developmentally deficient, embryos being picked up

	1				2				3			
Omeprazole	Normal	Radialised	No Expression	Severe Damage	Normal	Radialised	No Expression	Severe Damage	Normal	Radialised	No Expression	Severe Damage
0 µg/mL	15	0	2	3	17	0	3	3	16	1	1	5
20 µg/mL	8	0	7	0	13	1	1	1	7	3	0	1
40 µg/mL	27	4*	4*	4	11	1	1	0	12	1	3	10
80 µg/mL	6	3	0	1	6	3	1	2	7	2	0	3
	4				**5**				* = one embryo has no expression and is radialised. This is counted once in each column noted.			
	Normal	Radialised	No Expression	Severe Damage	Normal	Radialised	No Expression	Severe Damage				
0 µg/mL	21	0	3	4	18	0	2	1				
20 µg/mL	9	0	0	1	12 ^	1	1	2	^ = some samples seemed to contain more embryos than they should. The possible reasons for this are discussed in the text.			
40 µg/mL	4	1	0	6	7 ^	2	0	11				
80 µg/mL	6 ^	6	0	3	2	2	0	1				

Table 4.3: *Morphology and* Pitx *expression in omeprazole treated embryos. Given in columns by fertilisation number are the number of embryos with apparently normal morphology and* Pitx *expression, the number exhibiting a radialized phenotype (for examples see Fig. 4.15), the number where no expression was observed whatsoever, and the number whose morphology was so disrupted by mechanical forces that no conclusions could be drawn. Notes at bottom right detail the meaning of ∧ and * characters.*

in error when I removed motile embryos from my samples. It is also possible that I mis-counted the number of positively phototactic and negatively geotactic embryos in these datasets. If the former is correct, these may be some of the embryos counted amongst the 'Severe Damage' data column.

No examples of classical *situs inversus* were seen in these samples. Anecdotally, more widespread ectopic expression of *Pitx* was observed as omeprazole concentration increased, but this was difficult to quantify, and so long as this expression was predominantly on the right and a mouth or shell gland was seen, these embryos were counted as 'normal'. It may be worthy of more specific study in the future, and higher levels of omeprazole might also be tried in order to generate a more classical *situs inversus* phenotype. It is also possible that severe defects inL/R patterning, rather than inverting the embryonic axis in my organism, result in completely deleterious effects early on in development. These would not be detected by my analysis, as only positively phototactic and negatively geotactic embryos were used in *in situ* analysis.

Examples of radialized embryos can be seen in Fig. 4.15. In some embryos, *Pitx* expression still appears to be asymmetric - some of these are shown in the top row of this figure. Other embryos appear to have lost the L/R asymmetric distribution of *Pitx* mRNA. The still-asymmetric expression of *Pitx* mRNA may mean that this portion of the L/R asymmetry cascade is still active, and is uncoupled from the radialised phenotype. The fact that *Pitx* expression is seen at the 32 cell stage, ahead

of the induction event, may hint that this is the case. Alternately it could mean that some signal is still getting through to the right-hand side of the embryo in these cases, and the radialisation defect is not as severe.

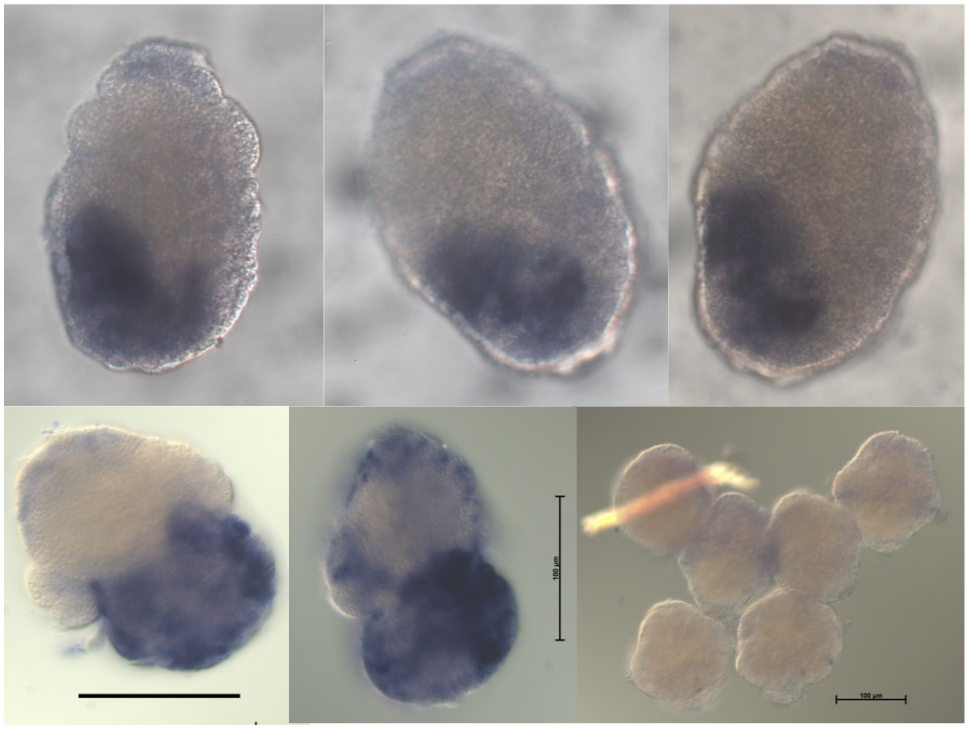


Figure 4.15: *Examples of radialised* P. vulgata *embryos after omeprazole treatment. Upper three images show radialised 24 hpf embryos with some asymmetric* Pitx *expression, but no sign of a mouth or shell field. Two images at lower left show embryos with ubiquitous* Pitx *expression in the lower portions of their bodies, and no sign of a mouth or shell field. At bottom right is shown an image of an* in situ *using a sense* Pitx *mRNA probe, in this case on embryos from fertilisation one of the 20 µg/mL omeprazole treatments, with no specific staining observed.Scale bars represent 100 microns when shown.*

The radialised phenotype, coupled with some embryos still exhibiting molecular L/R asymmetry despite it, seems to suggest that while ATPases are effecting the normal onset of L/R asymmetry, this is occurring through an indirect route - the incorrect establishment of the wider trochophore body plan. Uncoupling this from a purely L/R asymmetry effect will be difficult, but is possible. This can be accomplished by providing omeprazole only at specifically limited time points during development. In *P. vulgata*, for example, gastrulation occurs at around the 13 hour post fertilisation stage. By providing omeprazole at first only before or after this point, and then for even more restricted time points thereafter, the exact time point at which ATPases have their effect on L/R asymmetry, if any, can be pinpointed.

Before doing this, however, fertilizations should be performed in the coming breeding season to treat with ouabain and lansoprazole. These will confirm that any results are not simply the result of off-target effects from omeprazole, and provide a broader basis for this work to stand upon.

4.4.4 Possible Modes of Action

While expression studies seem to rule out a role for $(H/Na)^+/K^+$ ATPases in establishing asymmetry via the mechanism described in Levin et al. (2002) in early embryos, later expression patterns and the functional results described here seem to suggest that ATPases at least help in the normal establishment of the axes of spiralian embryos. The mechanism for this is as yet unknown, but as the first bilateral cleavage observed in *P. vulgata* is immediately after the induction event in the D lineage, if this is disrupted radialisation could well result.

Indeed, it has been noted in *P. vulgata* that when the D lineage is prevented from being established, radialised embryos are seen (Kuhtreiber et al., 1988; Damen & Dictus, 1996). This was done both with ablation, and with the use of the molecule monensin, which is known to interfere with both Na^+ and H^+ transport (Kuhtreiber et al., 1988). Monesin is said to disrupt processing of secretion by causing swelling of the Golgi system, interfering with intra-Golgi transport operations (Kuhtreiber et al., 1988). The authors of this work noted that the extracellular matrix might be disrupted by such an event, causing a loss of induction. When taken in combination with my results, however, this finding raises further questions. $(H/Na)^+/K^+$ ATPases are found on the outside of cells, rather than intracellularly, and it is possible that instead of interfering with intra-Golgi transport operations, monesin interfered with some form of current or charge establishment within the cells themselves.

MAP kinase (ERK) inhibition with U0126 also causes radialisation of embryos (Lartillot et al., 2002b). MAP kinase is activated in the D lineage, and MAP kinase activation is itself a key part of the induction event (Lambert & Nagy, 2001; Lartillot et al., 2002b). How these are phosphorylated is an interesting question, and could potentially be potentiated by charge, or at least by molecules responding to it. The

use of immunohistochemistry against activated MAPK in my radialised embryos could reveal if the D lineage has been specified in these samples.

Recent work from the Levin laboratory has been used to suggest an ancestral role for ion channels in establishing all three body axes (Aw et al., 2008) in vertebrates. Experiments could suggest that this role could be conserved more broadly in the Bilateria, whatever the connection with the establishment of L/R asymmetry. This could be by the establishment of a voltage gradient across a large portion of the whole embryo, or established within the blastocoel or in the D lineage upon contact with the roof of the blastocoel.

Another possible means for the communication of messages between ATPases and TGF ligands is more direct. Dahal et al. (2012) have shown a direct relationship between perturbation of K^+ levels and BMP functionality in *D. melanogaster*, perhaps by hindering Smad phosphorylation, or alternately by inhibiting Dpp (BMP2/4) production or receptor complex stabilization. The exact mechanism is yet to be determined, but it is easy to see how similar fluctuations in potassium level could inhibit TGF signalling in a similar way in the *P. vulgata* embryos seen in the present experiment.

Whatever their mode of action or role in the embryo, it is clear that $(H/Na)^+/K^+$ ATPases play a key role in establishing the broader body plan of *P. vulgata*. How this role relates to L/R asymmetry, or to what extent it reflects a shared ancestral role, remains to be deciphered, but such a key mechanism is nonetheless worthy of further consideration in the future.

4.4.5 Other ATPase Inhibitors

Alongside this work, a number of other ATPase inhibitors were assayed for a potential effect on the development of asymmetry in *P. vulgata*. The effect of these, and the numbers of embryos developing normally at given concentrations, is stated in Table 4.4. The concentrations used were based on the results seen in Shimeld & Levin (2006, Fig. 2A), and similar results were seen when concentration thresholds were exceeded.

These results suggest that ion channels play a variety of early roles in embryos,

and that this role may be conserved, at least between chordates and molluscs. What these roles are, however, is still somewhat of a mystery, and will require more focussed experimentation. The relatively low dosages required for broadly teratogenic effects to take hold, and the early onset of an effect, suggests that their role is not limited to the establishment of L/R asymmetry but is more fundamental to embryonic patterning and homeostasis.

		Replicate 1	Replicate 2	Replicate 3	Replicate 4	Replicate 5
Skelid (V-ATPase)	0	6	6	11	3	4
	25 µg/mL	FATAL ARREST, CIRCA 8 CELL STAGE				
	50 ≤ µg/mL	FATAL ARREST, CIRCA 8 CELL STAGE				
Concanamycin A (V-ATPase)	0	4		9	3	11
	0.5 ng/mL	6		5	0	2
	1 ≤ ng/mL	FATAL ARREST, CIRCA 8 CELL STAGE				
Barium Chloride (K Channel)	0	13	8	4	8	2
	50 µg/mL	FATAL ARREST, CIRCA 2 CELL STAGE				
	100 µg/mL	FATAL ARREST, CIRCA 2 CELL STAGE				
9-anthrocene Ca Acid (Cl Channel)	0	4	54		1	17
	15 µg/mL	1	2		4	5
	30 µg/mL	FATAL ARREST, CIRCA 2 CELL STAGE				

Table 4.4: *The results of pharmacological inhibition on the growth of* P. vulgata *embryos using inhibitors of a variety of other ATPase and ion channel proteins. Fertilisations of five independent* P. vulgata *pairs were used as the source of gametes. Numbers shown represent positively phototactic/negatively geotactic individual larvae at 24hpf in given treatment. Effect is stated in capitals when seen at concentration noted. In greyed-out boxes, no normal embryos observed at 24 hpf. Control rows contain DMSO if and as stated in Chapter 2.*

It would be possible to decrease dosage and attempt to find the minimum possible dosage of these drugs required to effect development without arresting cleavage. If this work was done, assays for *Pitx* expression could then be used to detect if these genes play any role in establishing L/R asymmetry in *P. vulgata*. Time constraints, and particularly the limited time window provided by the *P. vulgata* spawning season, prevented this work from occurring in time for inclusion in this thesis. This preliminary evidence will however be useful if work on the wider conserved role of ion channels in embryonic development is pursued in the future.

4.4.6 Calcium and Asymmetry

As noted in the introduction to this thesis, Ca^{2+} localisation has been noted as playing a number of roles in the establishment of asymmetry in vertebrates. Lanthanum chloride is a potent blocker of Ca^{2+} channel activity, and, in concert with experiments on ion channel activity detailed above, some embryos were also exposed to

this chemical during early development.

	1		2		3		4		5	
Lanthanum Chloride	Photo/Geotactic	Defects	Photo/Geotactic	Defects	Photo/Geotactic	Defects	Photo/Geotactic	Defects	Photo/Geotactic	Defects
0	4	9			1	14	4	26	8	4
25 µM	5	21			0	22	3	17	2	19
50 µM	2	16			2	13	4	8	4	16
100 µM	0	16			0	13	1	22	2	20

Table 4.5: *The results of pharmacological inhibition of Ca^{2+} channel activity on the growth of* P. vulgata *embryos. Fertilisations of five independent* P. vulgata *pairs were used as the source of gametes. Numbers shown represent positively phototactic/negatively geotactic individual larvae at 24hpf in given treatment on left, and number of abnormally developed embryos on right of each paired column.Numbers of undeveloped embryos not shown, in the hundreds for each treatment (including controls). In greyed-out boxes, no normal embryos observed at 24 hpf.*

The results of this experiment can be seen in Table 4.5. Empirically, a difference was observed in the viability of *P. vulgata* embryos between the controls and treatments seen in this table, but the sample sizes are sufficiently small that this difference is not statistically different under χ-squared analysis. I can still assay for a role in the establishment of L/R asymmetry by comparison of *Pitx* expression. Unfortunately, the low fertilisation efficiency of *P. vulgata* in the 2012 season meant that these results may be unusable for further analysis. Given the widespread role of calcium localisation in the establishment of asymmetry, I would recommend further investigations in this regard, perhaps in concert with live imaging of calcium ion flux, both with and without pharmacological inhibition.

4.4.7 Summary, Functional Testing

This work is not complete, and will be augmented by ongoing investigation in the Shimeld laboratory - both in an attempt to narrow the temporal bounds of ATPase action and to increase phylogenetic sampling by performing the same experiments in *P. lamarckii*. If a discrete time can be pinpointed for the action of ATPases in development, it could be well worth raising antibodies to investigate protein localisation, allowing firm conclusions to be drawn as to the role of these proteins in patterning Spiralian embryos early on in development.

The insights that have been drawn, however, are of interest when compared with results in the Deuterostoma and *D. japonica*. Both of these organisms use AT-

Pases for certain aspects of asymmetry establishment, and at least in the Deuterostomia ATPases are present early in development, even if their role is disputed. It therefore seems likely that a broad role in embryonic development for ATPases is shared ancestrally. Finding such evidence for a shared early role for ATPases in the establishment of bodily axes is in itself an interesting offshoot of my research. Uncoupling the role for these in establishing L/R asymmetry from these broader roles will be difficult, but is well worth doing, as their ubiquity at this key life stage makes them well and truly worthy of further investigation.

4.5 Conclusions, Breaking Lophotrochozoan Symmetry

This section of my research has been the most speculative, and has raised as many questions as it has answered. ATPases, while not displaying the marked asymmetry at early stages noted by Levin et al. (2002), do seem to play a role in establishing the axes of embryos, as when their action is perturbed a radial phenotype is observed. Far more functional work is required to confirm this, and to narrow down the point at which this occurs. If their effect is seen at the point of induction of the D lineage, as speculated, this would be interesting, and not just for L/R asymmetry establishment.

The expression of *Nodal* paralogues and *Pitx* as noted here also provides a building block for future investigations. These expression patterns can be used as readouts of the normal establishment of L/R asymmetry, as described above, as the basis for functional tests by gene knock-down, or in comparison with protein localisation work. The *in situ* images shown here suggest that some form of sub- or neo-functionalisation has occurred in the *Nodal 2* gene, at least compared to *L. gigantia* expression. While it should be noted that *L. gigantia Nodal* could have lost ancestral domains of expression itself, the duplication event in *P. vulgata* will have opened the possibility for subdivision of roles between prologues, which could be very useful for the further functional dissection of the role of *Nodal* in the Lophotrochozoa as a whole.

Chapter 5

Lophotrochozoan TGF β Pathways and Regulation

5.1 Patterning and Maintaining Asymmetry - the Nodal Pathway

As noted in the general introduction to this thesis, the *Nodal* pathway is of central importance to the proper establishment of L/R asymmetry once initial symmetry is broken. This has been long established in vertebrates, but the characterisation of *Nodal* expression in molluscs (Grande & Patel, 2009) has only begun to address the conservation of the entire *Nodal*-related signalling cascade in the Lophotrochozoa. As part of my investigations, therefore, I have investigated the conservation of the TGF β pathway in this superphylum, providing a broad underpinning to our knowledge of this vital cascade upon which future work can be based. I have also briefly investigated the regulation of *Nodal* signalling. This work suggests that the broadest aspects of regulation of the Nodal pathway may have been established in the Urbilaterian stem lineage, and will provide fertile ground for a variety of investigations into this field. In short, the work described in this chapter:

- Provides the first insight into transcriptional regulation of Nodal genes in the Lophotrochozoa
- Constitutes a systematic overview of the TGF β signalling cascade across a previously neglected Superphyla
- Allows the identification of fruitful avenues for future research into the reg-

ulation of Nodal and other TGF β ligands, which is vital for further understanding of the conservation of the process of L/R asymmetry establishment in the Lophotrochozoa

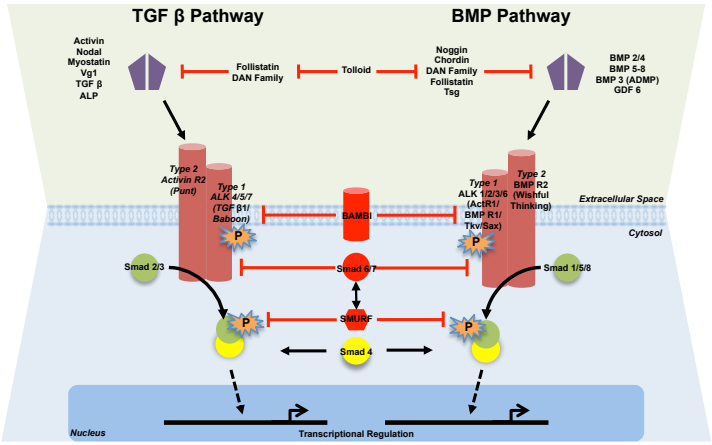


Figure 5.1: *Reproduced from Fig. 1.5 for convenience. Representation of the canonical signalling pathways for TGF β-like and BMP-like cascades with inhibitors of signalling shown in red and operational signalling shown in black. Refer to Fig. 1.5 for full figure legend.*

Nodal transcriptional regulation has been studied in depth in a number of vertebrate species, as noted in the introduction to this work. As a number of canonical signalling pathways and their known binding domains (including, notably, the Notch pathway) are known to be involved in genomic-level regulation of *Nodal* transcription these represented excellent targets for investigation using my novel genomic data. We also wished to discern how Nodal ligands could be regulated by the wider TGF β pathway, as to date very little is known about this important cascade within the Lophotrochozoa as a whole.

The TGF β signalling pathway (Fig. 5.1) has been well studied in a wide variety of traditional model systems, and is regarded as a module (Wagner, 1996), capable of regulating homeostasis, growth, and differentiation in a range of contexts (Derynck & Miyazono, 2008; Moustakas & Heldin, 2009). While the TGF β pathway most likely arose in something resembling its canonical form in the metazoan common ancestor (Pang et al., 2011), its divergence across the Metazoa, and the ancestral roles played by its components, are still in many ways unknown. With

the resources presented in Chapter 3 of this thesis, along with recently published sequence resources, I was able to catalogue the entirety of the Nodal signalling pathway and wider TGF β cassette across the Lophotrochozoa, allowing for the first time a Metazoa-wide understanding of the evolution and divergence of this crucial signalling pathway.

Attempts have been made to categorise the ancestral metazoan TGF β cassette (Herpin et al., 2004; Adamska et al., 2007; Huminiecki et al., 2009). According to such analysis, it is known that the original Metazoan repertoire consisted of at least four TGF β receptors and four Smads (Suga et al., 1999; Huminiecki et al., 2009), although the extent of the original ligand cassette and regulatory repertoire remains uncatalogued. While fully-fledged TGF β signalling components have not yet been found outside the Metazoa, the choanoflagellate *Monosiga brevicollis* has been shown to possess a gene with an MH2 domain similar to that of a Smad family protein (Srivastava et al., 2010; Pang et al., 2011). Both *Trichoplax adhaerens* and *Mnemiopsis leidyi* have a fairly complete central signalling pathway complement (Huminiecki et al., 2009; Pang et al., 2011) and sponges possess at least the basic receptor and Smad complements that can be found generally in the Bilateria (Suga et al., 1999).

The evolution of the regulatory elements of the cassette are less well understood. In order to understand what could be receiving and modulating Nodal signals in the Lophotrochozoa, a rigorous examination of the entire cassette was needed to discern the true conservation of this pathway in this superphylum, as there was no *a posteriori* guarantee that the findings of these papers would be mirrored in this clade. The TGF β signalling pathway has been well studied in traditional ecdysozoan (Patterson & Padgett, 2000) and deuterostome (Massagué et al., 2000) model systems such as *D. melanogaster* and *M. musculus*. The few attempts that have been made at categorizing elements of the lophotrochozoan TGF β cassette have typically been limited to individual elements and/or single species (for example, Herpin et al. (2005); Freitas et al. (2007); Fleury et al. (2008)).

In essence, while the TGF β pathway regulates a number of highly complex processes in animal tissues, its mode of action is simple, and can be seen schematically

in Fig.1.5 . For more information about the canonical TGF β signalling pathway and its operation, please refer to section 1.4.2 of the introduction to this thesis. For all its importance and ubiquity, the TGF β pathway within the Lophotrochozoa has yet to be satisfactorily documented. Some studies have found evidence of diversification at the ligand level within the leech *Helobdella robusta* (Kuo & Weisblat, 2011) and Platyhelminthes (Gavino & Reddien, 2011; Freitas et al., 2007), while other aspects of the signalling pathway may be conserved (Molina et al., 2011) although this is unclear (Kuo & Weisblat, 2011). The identification of the full signalling complements of a number of lophotrochozoan species should provide a springboard for the disentanglement of this network.

The advent of transcriptomic and genomic data in my model species as described in Chapter 3, coupled with the publication of key datasets, allowed a more comprehensive approach to be taken to understanding Nodal and TGF β signalling across the Lophotrochozoa. To further investigate and determine the conservation of this key signalling pathway across the Lophotrochozoa, the genomes of *L. gigantea* and *C. teleta* (Simakov et al., 2013) and the planarian *Schistosoma mansoni* (Berriman et al., 2009) were searched as described in section 2.2.8 of this thesis, along with novel genomic resources for the rotifer *B. plicatilis*, limpet *P. vulgata* and *P. lamarckii* . Here I demonstrate the existence of the majority of the core TGF β cassette in the Lophotrochozoa, and the maintenance of a substantial portion of the *Nodal* regulatory cassette, confirming the conservation of this module across evolutionary time, albeit with extensive diversification of the ligand complement in this lineage and loss of genes encoding extracellular inhibitors in some species. This may have implications for Nodal signalling, as described further below.

5.2 *Nodal* Duplication and Consequences: Genomic Regulatory Elements

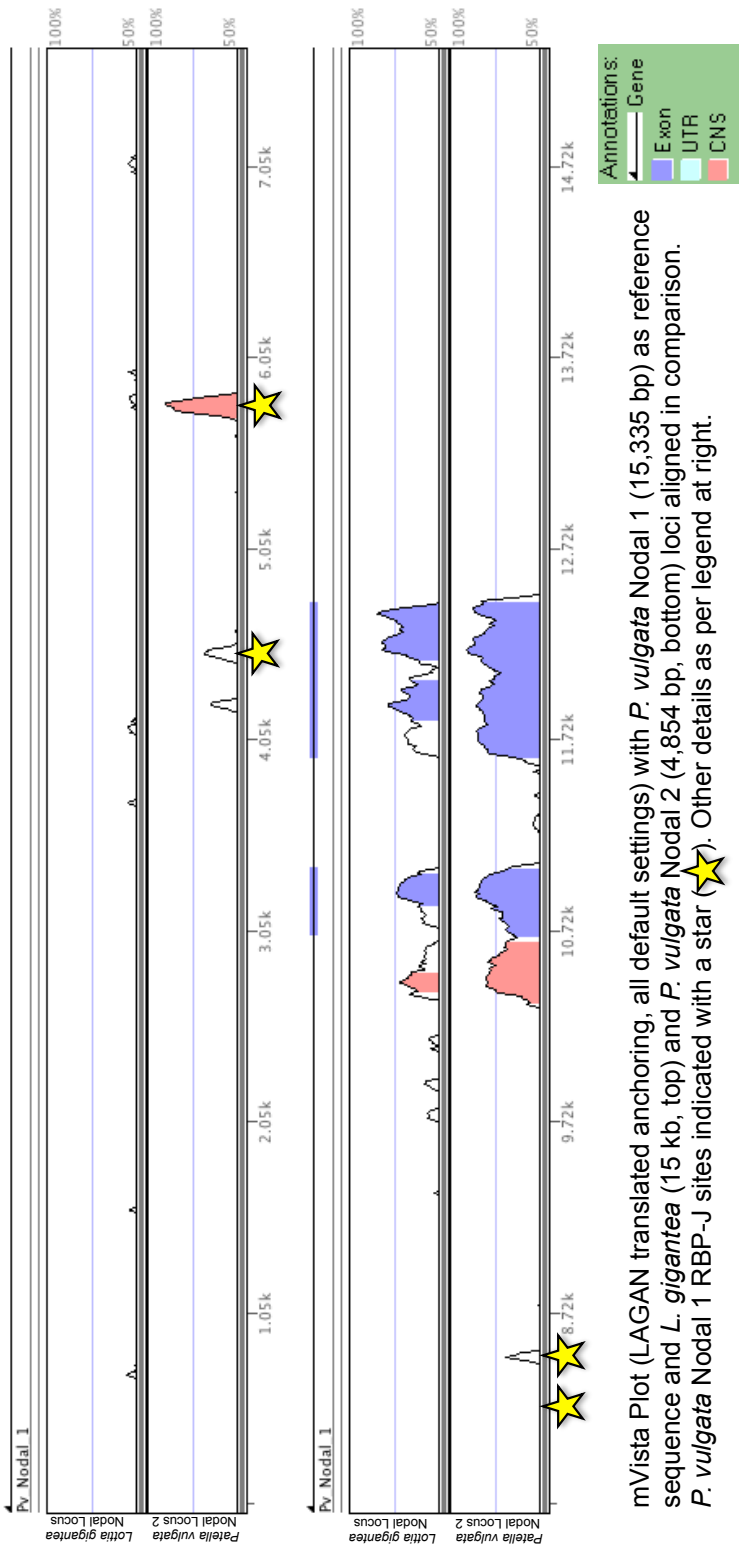
Nodal is not only regulated post-translationally. A key element in examining the conservation of the mechanisms of establishing L/R asymmetry between deuterostomes and lophotrochozoans is considering the regulation of *Nodal* at the genomic

level. If many of the elements that act to pattern the expression of this gene in chordates are also found in *P. vulgata* I can be very confident of a deeply conserved L/R asymmetry patterning cascade in the Metazoa.

As noted in the introduction to this report, there are four regions of genomic regulation noted around the *Nodal* gene in chordates. I aimed to test the presence and function of these elements, to determine if the regulation of *Nodal* has been conserved since the deuterostome/protostome lineage diverged. The duplication of the *P. vulgata Nodal* locus is particularly useful for this, as it provides a relatively closely related point of comparison, with the *Nodal* locus of *Lottia gigantia* as more distantly related species to test conservation of non-coding sequence. These species diverged approximately 180 million years ago, in the early Jurassic period (Nakano & Ozawa, 2007, Fig. 5), and as such they are already relatively derived from one another.

The Vista alignment of the two loci can be seen in Fig. 5.2, with *L. gigantia* Nodal (top) and *P. vulgata Nodal 2* (bottom) aligned to *P. vulgata Nodal 1* . It can clearly be seen that the *P. vulgata Nodal 2* is almost as different to *P. vulgata Nodal 1* at the coding nucleotide level (shaded blue) as *L. gigantia* Nodal. This extreme differentiation underlines further that these are not allelic variants, despite the lack of a fully contigged genome, and suggest that the duplication that gave rise to them is extremely ancient - it could even have occurred before the last common ancestor of *P. vulgata* and *L. gigantia*. Outside the coding region, *P. vulgata Nodal 2* has clearly more numerous and better-conserved noncoding elements in common with *P. vulgata Nodal 1* than *L. gigantia* Nodal.

Most interestingly for my purposes, these conserved noncoding elements coincide with the location of Notch response elements. Notch signalling is perhaps the best categorised cell-cell communication process, having been categorised as early as 1917 Morgan (1917). The Notch pathway as been shown to be involved in binary cell fate choice, cell lineage specification, boundary formation and stem cell maintenance. As noted earlier, recently it has been posited as having a role in the specification of left/right asymmetry in vertebrates, perhaps modulated by asymmetric extracellular calcium ion concentration, to produce expression of the



mVista Plot (LAGAN translated anchoring, all default settings) with *P. vulgata* Nodal 1 (15,335 bp) as reference sequence and *L. gigantea* (15 kb, top) and *P. vulgata* Nodal 2 (4,854 bp, bottom) loci aligned in comparison. *P. vulgata* Nodal 1 RBP-J sites indicated with a star (★). Other details as per legend at right.

Figure 5.2: *Comparison, using Vista (translated anchoring in LAGAN, and all default settings) of P. vulgata and L. gigantia Nodal loci.*

signalling molecule *Nodal* on the left side of the node in developing embryos. In mice, it does this via several upstream enhancers, located approximately 10 kb upstream of the *Nodal* locus (Krebs et al., 2003).

As noted in Fig. 5.2 canonical RBP-J sites ((C/T)GTGGGAA), the binding location for the downstream elements of Notch signalling, are found four times immediately upstream of the *P. vulgata Nodal 1* locus. RBP-J binding sites have been shown to be conserved in sequence and function from humans to *D. melanogaster* (Honjo, 1996; Taniguchi et al., 1998), and as such are likely to play a conserved role in my species.

Only three RBP-J sites are found upstream of the *P. vulgata Nodal 2*, which could be a consequence of the reduced contig length of this sequence. However, the first site, when aligned to *P. vulgata Nodal 1* corresponds with the region of exceptionally high conservation found at 5.8 kb into the alignment in Fig. 5.2. The latter two align to the to closely spaced RBP-J elements falling in the less well conserved area at 8.5kb on this graph. These represent strong circumstantial evidence for a role for Notch in regulation of L/R asymmetry, and as such are good candidates for cloning and testing with a reporter element. In addition, this role can be further tested with the use of DAPT, an inhibitor of Notch signalling. If expression of the GFP reporter element can be driven by the RBP-J containing region upstream of *Nodal* in the same area that Nodal is itself expressed, and then knocked down by DAPT treatment, firm inferences can be made about Notch's role in driving expression of Nodal in the Lophotrochozoa.

A microinjection protocol has already been established for this organism and proven successful (Hui, J. personal communication). Suitable expression vectors, using a GFP reporter construct driven by the regions 5' of the *Nodal* start site up to and including the start codon have been constructed. The practicality of utilising expression vectors in *P. vulgata* embryos has been shown previously (Damen et al., 1994), albeit with a LacZ fusion reporter. These have been constructed in two lengths for each *P. vulgata Nodal* gene, one incorporating the RBP-J sites, and one without. Unfortunately a disastrous breeding season for *P. vulgata* in the 2012/13 season precluded microinjection of this construct into viable embryos. Electropora-

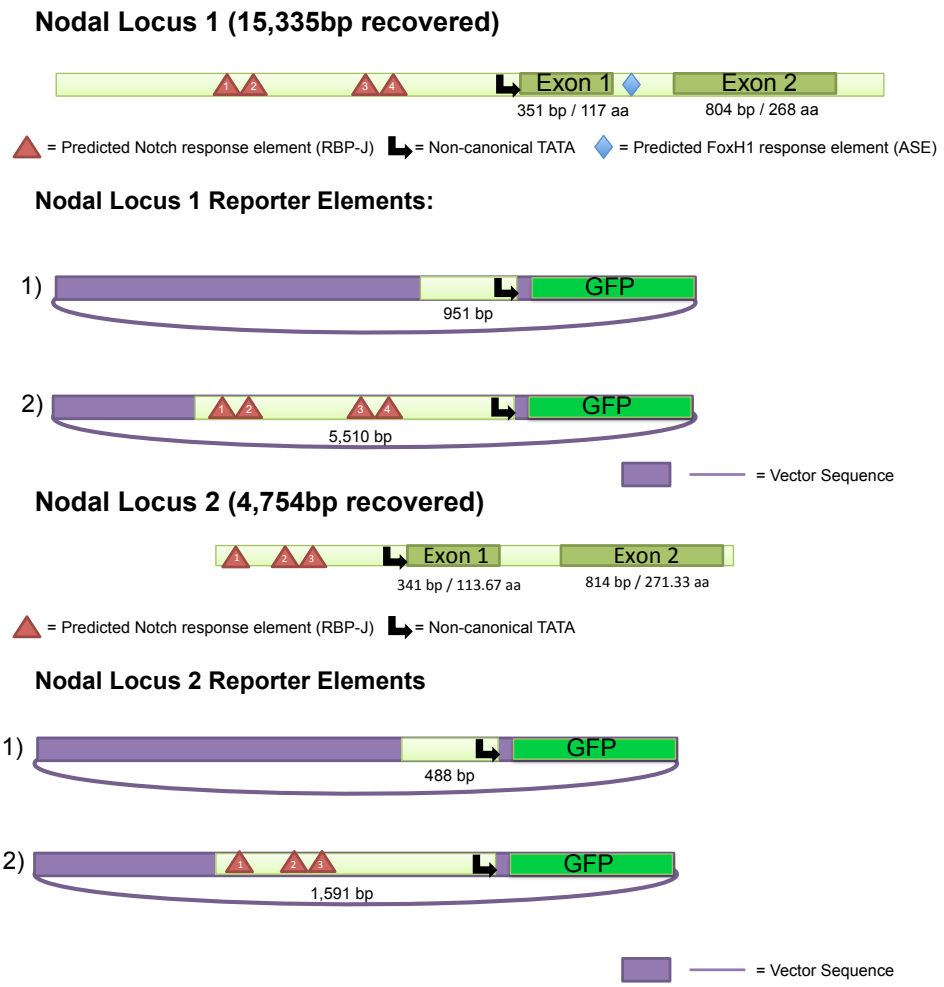


Figure 5.3: *Reporter elements constructed to test putative* P. vulgata Nodal *loci enhancers. For each of the two* P. vulgata Nodal *loci, two reporter constructs were made. All reporter elements are based on a CiCrys-EGFP construct with* Ciona intestinalis *regulatory sequence excised using Sal1 and BamH1. CiCrys-EGFP vector is that used in Shimeld et al. (2005)*

tion of these contracts into *C. intestinalis* has been attempted multiple times without success. it is hoped that the 2013/14 breeding season will be more consistent, allowing another attempt at microinjections to take place.

5.2.1 FoxH1

It has been shown in mice that *Nodal* is regulated by a enhancer within its large conserved first intron, to which the transcription factor *FoxH1* binds (Saijoh et al., 2000; Norris et al., 2002). This conserved enhancer sequence, a left side-specific enhancer dubbed the ASE, modulates patterns of *Nodal* expression throughout early

development. When the ASE is specifically removed normal L/R asymmetry is disrupted (Norris et al., 2002). While this has thus far only been demonstrated in mice, it is of interest to see whether it is conserved through to the Lophotrochozoa, particularly due to the presence of a putative *FoxH1* homologue within my *P. vulgata* genome data and that of *L. gigantia*, suggesting a conserved presence of this gene as far back as the Urbilateria (data not shown).

Examination of the *P. vulgata* intronic region for the Nodal 1 gene reveals a single site with a canonical ASE element sequence (AATCAACAT), antisense to the direction of transcription. This site is not seen in the intron of *P. vulgata Nodal 2*, although two slightly derived forms are seen (AATCAACTC and CATCAACGT). Given their short length, which could indicate chance similarity, little can be inferred as to a possible function at this time. The intron of *L. gigantia*, which is very truncated at 198 bp, much shorter than the 0.5 - 1 kbp intron normally observed at the *Nodal* locus, also possesses no apparent ASE sequence. While presence of an ASE element is not confirmation of a role in regulating L/R asymmetry, it represents an intriguing potential regulative element.

Further functional testing of *FoxH1*'s binding to this enhancer is also possible. If *FoxH1* encoding dsRNA was injected alongside the reporter element, canonical RNAi could be used to knock down the expression of endogenous *FoxH1* - canonical RNAi already having been attempted with positive results in *P. vulgata* (Hui, J., personal communication). If FoxH1 is driving expression, as would be likely (Norris et al., 2002), I would then expect expression of Nodal, assayed via RNA *in situ* hybridisation, to be knocked down accordingly. This would require careful consideration of proper controls, but could act as confirmation of a conserved role for FoxH1 in the establishment of L/R asymmetry, especially if confirmed by RNA *in situ* hybridization, on downstream target genes such as *Pitx*, which would be expected to show reduced expression as a result. This experiment has been attempted by another member of the Shimeld laboratory on the basis of my results, but the 2012/13 breeding season, which was disrupted due to low salinity in the cachements where *P. vulgata* is collected, resulted in exceptionally poor (< 0.5 %) fertilisation efficiencies that rendered microinjection of dsRNA impracticable. This

experiment will again be re-attempted in the coming months.

5.3 TGF β Signalling Cascades and Regulation

5.3.1 TGF β Ligands

Nodal is by no means the only TGF β ligand found in my model species, and as science is presently unaware of the diversity in this regard, I set out to catalogue the Lophotrochozoan cassette and confirm the true orthology of *Nodal* in the Lophotrochozoa. TGF β ligands participate in a diverse range of mechanisms in cellular specification and functionality. They are synthesized as relatively long precursor proteins, but an N-terminal propeptide is cleaved during processing, leaving a short, 110-140 amino acid mature ligand. This mature ligand can be recognized by the characteristic conservation of at least six cysteine residues that when folded form a structure known as a cysteine knot, stabilized by three disulfide bonds (Sun & Davies, 1995; ten Dijke & Arthur, 2007).

There are 33 distinct genes encoding TGF β ligands in humans, seven in *D. melanogaster* and five in *C. elegans* (Huminiecki et al., 2009). These ligands can be split into two broad families, the TGF β/Activin subfamily and the BMP (bone morphogenetic protein) families (Yamamoto & Oelgeschlager, 2004). Generally these subfamilies are mirrored by the signalling pathway through which they operate (see Fig. 1.5 for details) but this is not always the case - Nodal, for instance, is generally said to belong to the BMP subfamily, but Nodal signals are transduced via the TGF pathway.

In many ways the complements of TGF β ligands found in lophotrochozoans are similar to those found in more classical model organisms. This similarity breaks down, however, in the Activin/TGF β subfamily, which appears to have undergone extensive divergence in this superphylum. Initial attempts at making phylogenetic trees for TGF β ligands were confounded by the diversity of ligand sequence in this subfamily, which resulted in poor alignments and multiple gaps, and, in turn, in poorly-supported trees. To better ascertain phylogenetic relationships within and between TGF β ligands, I have analysed the interrelationships of the TGF β family

in two steps - firstly, by assigning, on the basis of the preliminary tree and by Blast identity, ligands to either the BMP subfamily or the Activin/TGF β subfamily, and secondly by analysing these two subfamilies separately. Maverick genes, despite falling into the Activin/TGF β subfamily in some phylogenetic reconstructions, were analysed with the BMP clade, due to historical evidence placing them here with more robust support (Van der Zee et al 2008 Fig 2, Pang et al 2011 Fig 2).

The results of maximum likelihood and Bayesian inference of their phylogenetic relationships for the BMP subfamily can be seen in Fig 5.4. Both means of phylogenetic inference recover clear familial relationships between lophotrochozoan sequences and orthologues of many well described genes. Bootstrap values are not always high, most likely due to the relatively short dataset (88 amino acid alignment) from which these samples were drawn. Posterior probabilities, however, clearly support many nodes on the Bayesian tree, for example, the ADMP clade has a posterior probability of 1 under Bayesian analysis, but bootstrap support of 58 under ML analysis. Nonetheless, I am confident that this clade has been recovered correctly, as bootstrapping is known to suffer when aligned regions are short due to the limited data available for re-sampling.

The existence and expression of several members of the BMP subfamily of the TGF β ligand family in the Lophotrochozoa has already been established by prior studies, although the phylogenetic distribution of these studies has been scattered and the full Lophotrochozoan complement was unclear (Nederbragt et al., 2002; Render, 1997; Freitas et al., 2007; Grande & Patel, 2009; Kin et al., 2009; Wang et al., 2010a). It seems that annelids and molluscs have retained the majority of the diversity of the BMP subfamily family found in the Ecdysozoa and Deuterostomia, and in many cases better conservation is found than in ecdysozoan models. In contrast *B. plicatilis* and, in particular, *S. mansoni*, have apparently lost many of the more BMP-like members of the of the TGF β ligand cassette.

The apparent Ecdysozoan-specific loss of *Nodal* and *BMP3/GDF10* is confirmed by my analysis, with well-supported nodes containing lophotrochozoan orthologues from a number of species seen for these gene families. Particularly intriguing for the present study was the apparently lineage specific duplication of *Nodal*

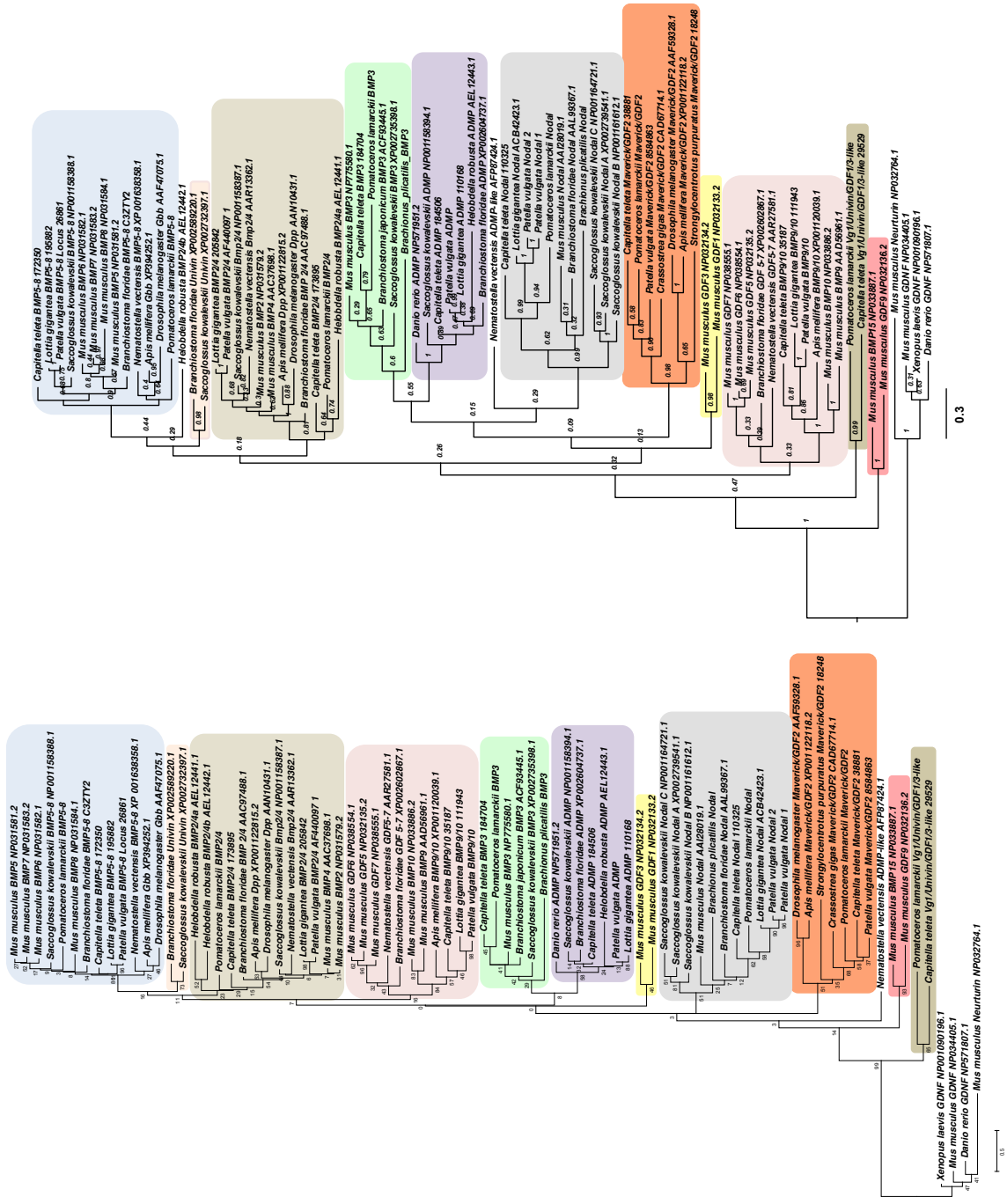


Figure 5.4: *Phylogeny of BMP-like ligand subfamily interrelationships across the Metazoa, as determined by maximum likelihood (Tamura et al 2011, left) and Bayesian (Huelsenbeck and Ronquist 2001, right) methods. Alignment generated by MAFFT (Katoh & Standley, 2013) using the L-INS-i strategy resulting in an 88 amino acid informative alignment of the mature ligand domains after the exclusion of gaps, as seen in Fig. 5.5. Both phylogenies determined using the WAG mode (Whelan & Goldman, 2001). Bootstrap percentage (of 1000 replicates) and posterior probabilities (after 5,000,000 generations) can be seen at the nodes of ML and Bayesian trees respectively. Scale bars represent substitutions per site at given distances.*

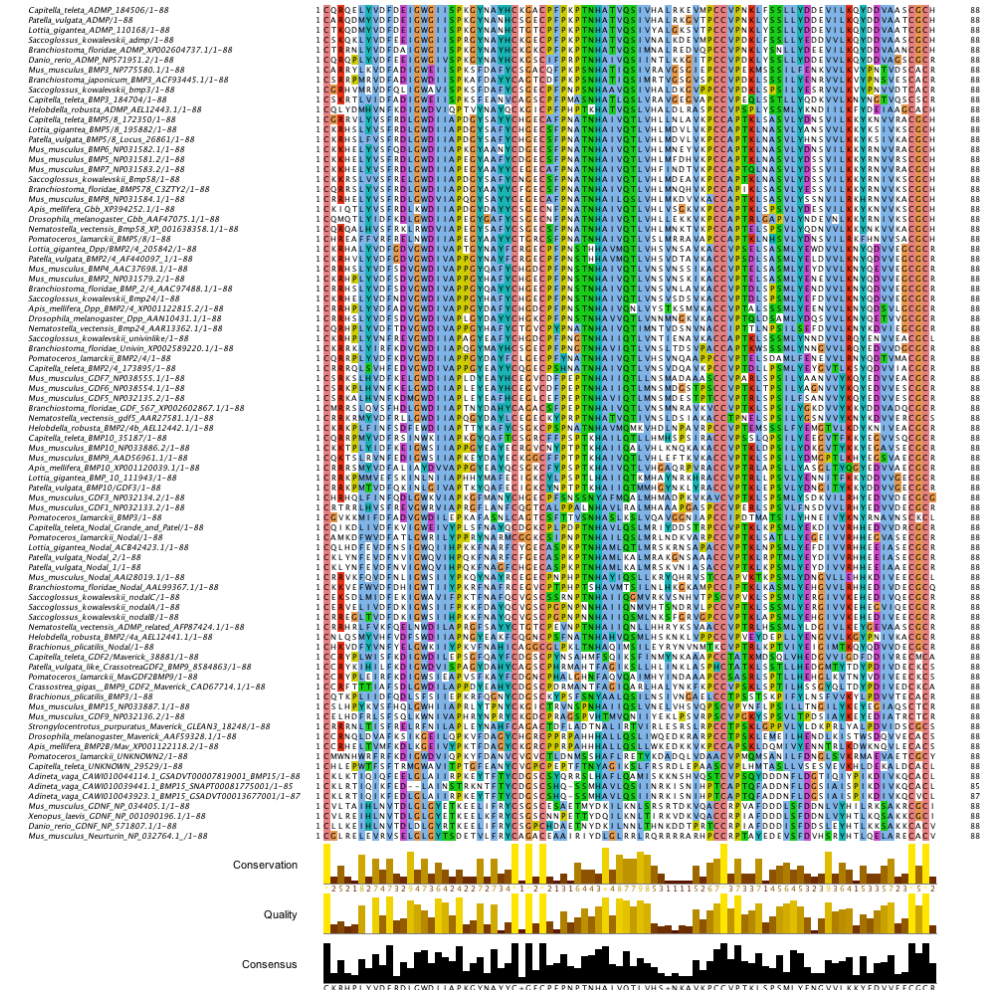


Figure 5.5: *BMP Family Gene Alignment Used in Fig 5.4. Alignment generated by MAFFT under the L-INS-i model as described in Methods, and displayed using Jalview with ClustalX colouring.*

which has occurred in *P. vulgata*. These two copies are reasonably dissimilar at the amino acid level (14 differences out of 119 amino acids from the proteolytic cleavage site onwards, as can be seen in 5.6), and even more dissimilar at the nucleotide level, suggesting that this duplication is reasonably ancient, and does not represent allelic variation. Both of these genes were cloned from cDNA derived from mixed embryonic samples, indicating that they are both transcribed into RNA, and their expression can be seen in Chapter 4. This duplication has provided a range of data at the locus level, which is discussed later in this thesis.

Figure 5.6: *Jalview alignment of Patella vulgata Nodal protein sequence, coloured with Blosum 62 score. Note 14 differences between two paralogues*

Nodal genes have been lost in the schistosome species of Platyhelminth examined, but their presence in rotifer indicates that they were ancestrally shared until the divergence of Trochozoa and 'Platyzoa' - although the true phylogenetic nature of the clade known as the Platyzoa (also known flippantly as the Longbranchozoa) is uncertain, the retention of *Nodal* suggests that it may retain its importance outside the Trochozoa as well. The *Nodal* and *Pitx* genes of *B. plicatilis* were cloned by the author during a visit to the Dearden laboratory in January 2013, and expression data will be reported as soon as these experiments are performed by my collaborators in that lab.

More complex is the case of the Univin/VG-1/GDF1/3 like family. While these are not recovered in my dataset as a monophyletic grouping in Fig. 5.4, they have appeared in previous studies as a well-supported clade of deuterostome-specific genes, for example in Lapraz et al. (2006). The *Helobdella robusta* (Leech) *BMP 2/4b* sequence (AEL12442.1) has also been noted as showing some resemblance to the Univin/VG-1/GDF1/3 like family. Some of my analyses tentatively suggest that annelids may in fact contain an orthologue of this family, which I have named Vg1/Univin/GDF1/3-like on my trees in Fig. 5.4 Better sampling is needed across the Lophotrochozoa in order to confirm their true relationships, and estab-

lish whether these are truly orthologous to the Vg1/Univin/GDF1/3 genes in the Deuterostomia.

Canonical homologues of *BMP2/4* (*Dpp*), *BMP 5-8* (*Gbb/Scw*), *BMP9/10* (*GDF5-7*), *ADMP* and *Maverick* are also found in the Lophotrochozoa as reported by my trees (Fig. 5.4). These appear to be far better conserved in the Mollusca and Annelida than in other lophotrochozoan lineages examined, and are almost ubiquitous in my sampled mollusc and annelid models. No homologues for *GDF9/BMP15* can be found in my lophotrochozoan datasets, suggesting that this is a deuterostome innovation.

Figure 5.7 shows the inferred identity of the genes of the activin/myostatin/inhibin-like clade as determined by phylogenetic analysis. Clear and reproducible signal was found for canonical Myostatin and Activin/Inhibin clades, especially in the case of Bayesian analysis, where Myostatins cluster together with a posterior probability of 1. Clades for TGF β and Lefty homologues in deuterostomes are consistently recovered with good support. On the basis of the tree presented here, Lefty and TGF β ligands sensu stricto seem to be deuterostome innovations, with the previous report of a tentative TGF β homologue in *M. leidyi* (Pang et al., 2011) standing in evidence against this. We have included this sequence in the tree shown in Fig. 5.7, and, as stated in Pang et al. (2011), it is only weakly supported as a homologue of the TGF β clade sensu stricto. Wider taxon sampling at the base of the Metazoa, and particularly in the Ctenophora, would allow us to test whether TGF β ligands (in the strictest sense) are indeed a deuterostome novelty.

It should also be noted that two sequences are not shown in my analysis in Fig. 5.7. The first of these sequences is a *P. vulgata* Myostatin-like homologue, whose sequence was partially recovered from my genomic and transcriptomic data, but which covers only a portion of the mature ligand sequence and was therefore excluded from analysis. The second is a *C. teleta* gene, protein ID 198732, which appears to be a markedly divergent TGF-like ligand. Its annotation in the *C. teleta* genome suggests that it has been found to be expressed, but it appears to have lost a significant portion of the mature ligand region. This may be the result of

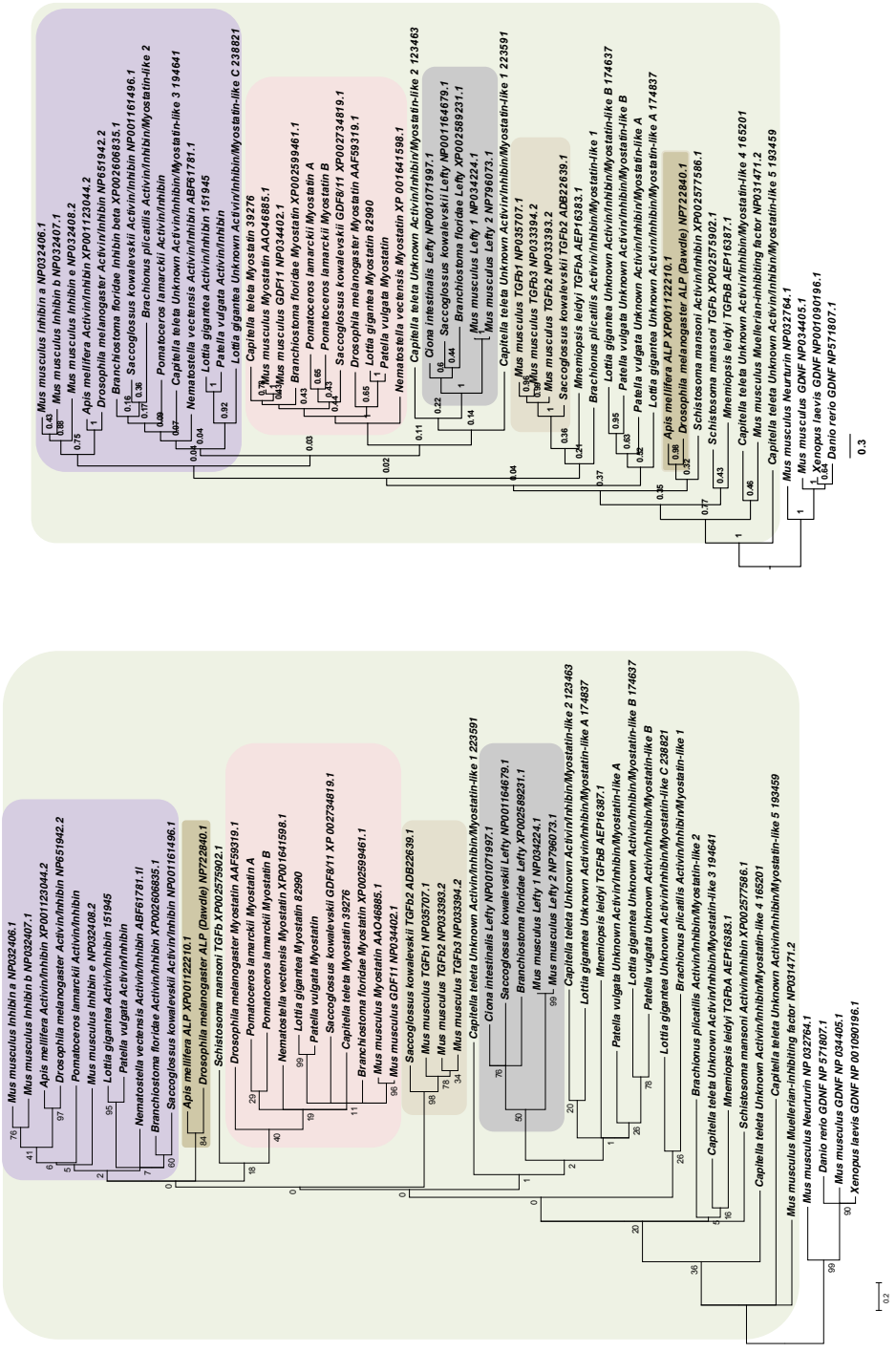


Figure 5.7: *TGF β-like subfamily phylogenetic interrelationships across the Metazoa, as determined by maximum likelihood (Tamura et al 2011, left) and Bayesian (Huelsenbeck and Ronquist 2001, right) methods. Alignment generated by MAFFT (Katoh & Standley, 2013) using the G-iNS-i strategy, resulting in a final 71 amino acid informative alignment as seen in Fig. 5.8. Both phylogenies determined using the WAG mode (Whelan & Goldman, 2001). Bootstrap percentage (of 1000 replicates) and posterior probabilities (after 10,000,000 generations before convergence) can be seen at the nodes of ML and Bayesian trees respectively. Scale bars represent substitutions per site at given distances.*

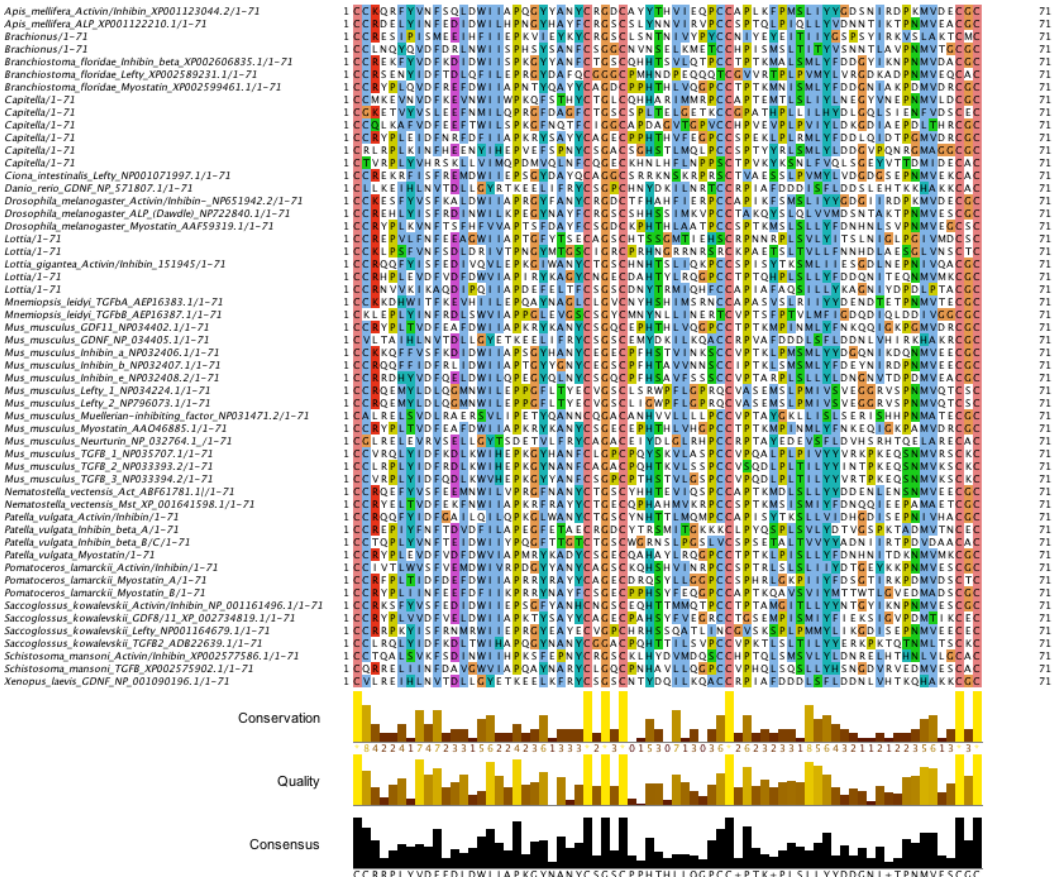


Figure 5.8: *TGF Family Gene Alignment Used in Fig 5.7. Alignment generated by MAFFT under the L-INS-i model as described in Methods, and displayed using Jalview with ClustalX colouring.*

pseudogenisation in progress, as without the portions of the ligand domains, it is unlikely the protein it encodes is fully functional in the same manner as canonical TGF-like ligands. The oyster *C. gigas* also shows evidence of diversification in TGF-like ligands (Fleury et al., 2008), although when these sequences are added to my analysis no further structure was added to my tree - indeed, these again seem to be fast-evolving, highly derived sequences, with uncertain homology to the other lophotrochozoan data presented in this paper.

ALP seems to be an ecdysozoan innovation, forming a clade with clear support and no orthologue seen in any lophotrochozoan species. *Muellerian inhibiting factor* has been suggested to be a deuterostome or even vertebrate-specific ligand, but weak support groups the *C. teleta* Unknown Activin/Inhibin/Myostatin-like 4 with this sequence in the Bayesian analysis. Whether this is a true relationship, or instead is a result of long branch attraction is uncertain, but could be tested with increased genomic sampling across the Lophotrochozoa.

Some support is found for a mollusc-specific clade of ligands, encompassing *L. gigantea* Unknown Activin/Inhibin/Myostatin-like A and B and similarly named sequences in *P. vulgata*. The roles for these genes in vivo is as yet uncatalogued, and no homologue is found in the genome of *C. gigas* by Blastp search. These sequences therefore could represent a gastropod or patellogastropod novelty.

Outside of these clades, however, little signal can be recovered to support relationships between a diverse range of other ligands in the Lophotrochozoa. On the basis of the lack of clades forming, even between such relatively closely related species as *P. vulgata* and *L. gigantea*, it seems that these ligand sequences (if orthologous) are either very unconstrained by evolution, and are thus changing in sequence rapidly, or alternately they are the product of multiple lineage-specific duplications, and therefore are independently changing under their own evolutionary pressures. These sequences are named without inference to their evolutionary relationships, and the terms A, B, C etc in multiple species are used to allow within-species numeration rather than any inference of orthology.

Some additional insight could potentially be drawn from the number of cysteine residues found in these sequences, as these have a characteristic distribution

in model organisms. However even this is little help. In some vertebrates, TGF β ligands sensu stricto and inhibin β are said to have nine cysteines, while inhibins (along with BMPs and GDFs in the BMP-like clade) have seven (Chen & Shen, 2004; Derynck & Miyazono, 2008; Moustakas & Heldin, 2009). These pair to form four and three disulfide bonds respectively, with the remaining cysteine residue forming such a bond when the ligands dimerise to signal. Lefty proteins, as well as GDFs 3, 9 and BMP 15 are said to have only six cysteines - they do not bond covalently to form dimers (Chen & Shen, 2004; Derynck & Miyazono, 2008; Moustakas & Heldin, 2009). However, in the *M. musculus* and deuterostome sequences used in my analyses, Lefty and Muellerian-inhibiting factor proteins possess seven cysteines (lacking the fifth and second respectively when counting from the N terminus of the mature ligand), while all other deuterostome ligands in my dataset possess at least eight cysteine residues. This may represent the ancestral condition, with the more derived forms studied in detail by those papers referenced above. These cysteines can be seen in positions 1, 2, 28. 32, 57, 58, 85 and 87 when present in my BMP alignment, or in positions 1, 2, 27, 31, 42, 43, 69 and 71 of my TGF - like alignment (data not shown).

The majority of my Lophotrochozoan ligand sequences possess eight cysteine residues, Of the uncategorised sequences seen in Fig. 5.4, the four *L. gigantea* and *P. vulgata* sequences mentioned as forming a weakly supported clade earlier have 7 cysteines (lacking the fifth as counted from the N terminus), as do *C. teleta* Unknown Activin/Inhibin/Myostatin-like 4 and *C. teleta* Unknown Activin/Inhibin/Myostatin-like 1, which lack the second from the N terminus. One sequence, *C. teleta* Unknown Activin/Inhibin/Myostatin-like 5, has only six cysteines, lacking both the second and the fifth. When cysteines are lost from ligands in vertebrates, they too are lost from the second and/or fifth position, which further reinforces that some cysteines are vital for maintaining the cysteine knot form of the active ligand, while others can be lost.

Unfortunately, given the uncertainty regarding the number of cysteines found in canonical groups, this character cannot be used to further classify my ligands. A better understanding of the interrelationships between these ligands is proba-

bly only to be drawn from denser taxon sampling across the Lophotrochozoa. At present, the short length of the mature ligand sequence and lack of conservation of sequence in the longer propeptide means that phylogenetic inference, by whatever means, is limited by a lack of information.

The lack of constraint on the non-cysteine portions of the sequences of these ligands is also interesting from a structural perspective. The sequence of BMP-like ligands seems highly constrained, probably because of their vital interactions with receptors and regulators of their activity. That TGF-like ligands in the Lophotrochozoan clade can diversify so much, even between closely related species, when the remainder of the signalling cascade remains relatively stable raises questions about how these ligands can successfully maintain their secondary and tertiary structures in the face of relatively large sequence changes, when so many other TGF ligands have, presumably under purifying selection, maintained a relatively stable sequence over evolutionary time across the Metazoa.

As with the BMP-like ligands (shown in Fig. 5.4) TGF β-like ligands seem to have been lost from both *B. plicatilis* and *S. mansoni*, with those that are found not falling into identifiable clades. This echoes the findings of previous studies in schistosomes (Freitas et al 2007). It has been suggested that *S. mansoni* could use host ligands as part of its signalling cassette (Osman et al., 2006), which would explain low diversity of these ligands in this species. The lack of *B. plicatilis* TGF β ligands is more unusual, and future sequencing efforts in the Rotifera will reveal whether this loss is real, or an artefact of insufficient sequencing depth. It should be noted, however, that the *B. plicatilis* receptor complement is also slightly modified, and could reflect changes to ligand sequence and structure in concert with downstream aspects of this cascade.

The short alignments used to build the phylogenies shown here lead to the risk that genes are positioned as a result of stochastic changes in sequence, rather than as a reflection of true phylogenetic position. It is likely that this is seen in Fig. 5.7, where relaxation of selective pressure, and concomitant rapid loss of signal after duplication has lead to the lack of clades seen in the TGF beta tree, while other ligands, playing conserved roles, are kept in clades by conservation of sequence

important for ancestrally shared roles.

Short alignments are, however, unavoidable given the short length of ligands themselves. TGF ligands generally exhibit reasonable sequence conservation as a result of their well-conserved roles and pathways, and while the risk remains that results are misleading, the positioning of ligands within well-supported clades will generally reflect good evidence for conserved identity. For genes outside these clades, however, identity should be inferred only with caution.

In terms of L/R asymmetry establishment, while canonical *Nodal* genes have been found in the Lophotrochozoa, it is well worth exploring the novelty represented by the wide variety of *Activin/Inhibin/Myostatin-like* genes for a role in establishing or modulating this process. As these genes would be hypothesised to signal through the same receptor as Nodal, they could easily play a role in establishing or modulating the establishment of the same body axis as Nodal, similarly to how Lefty operates in the Deuterostomia. Unconfirmed reports suggest that, at least in *C. teleta*, one of these genes was expressed in a group of cells on the right side of the stomodeum (C. Grande, oral communication). Much further work is required to disentangle the functional roles of these novel ligands, both in the establishment of L/R asymmetry and more widely in these organisms, but with clear delineation of the diversity of these genes in this clade this can now begin in earnest.

5.3.2 Activin-receptor Like Kinase (ALK) Receptors

Each TGF β ligand pair binds to type 1 and 2 serine/threonine kinase family member receptors, which are often known as Activin-receptor like kinases (ALKs). When TGF β ligands form a complex with representatives of both types of receptor, the intracellular kinase domains of the receptors are brought together and the type I receptor is phosphorylated and activated (Massagué, 1998). In humans, a total of seven type I and five type II receptors have been described, while *Tribolium castaneum, Apis mellifera* and *D. melanogaster* possess a total of five, three type I, and 2 type II (Van der Zee et al., 2008).

The ALK receptor complements of the Lophotrochozoa have previously been the subject of investigation, with those of *C. gigas* already described (Herpin et al.,

2005). Coupled with my extensive knowledge of the diversity of these molecules in the freshwater sponge *Ephydatia fluviatilis* (Suga et al., 1999) and other non-bilaterian metazoans (Huminiecki et al 2009, Pang et al 2011) these are perhaps the best-catalogued components of the TGF β cascade. With some exceptions, the ALK receptor complement varies little across the Metazoa. Generally two Type 2 and three Type 1 receptors are found in any species, with TGF β-like signalling occurring through a set pair of dimerised Type 1 and Type 2 receptors (TGF β R1 and Act R2, also known by a diverse range of other names), while two Type 1 receptors (BMP R1 and the misleadingly named Act R1) can each be found in complex with BMP R2.

In vertebrates considerably more diversity of receptor number exists, most likely due to the 2R whole genome duplications. Larger numbers of receptor also exist in invertebrate deuterostomes and the cnidarian *N. vectensis*, most likely due to independent duplications in these lineages, some of which have been traced to specific nodes on the deuterostome tree of life (Huminiecki et al., 2009).

While the canonical five ALK receptors are found in all annelids and molluscs species examined (Fig. 5.9), these have diversified greatly in the rotifer *B. plicatilis*. There are only three apparent subunits present in this species, one a clear Activin R2 homologue, while the other two are more difficult to place, being clear type 1 receptors lying close together in my trees, possibly by long branch attraction. On the basis of blastp similarity, one sequence appears to be a divergent BMP R1, the other remains unclassifiable, being similar to both TGF β R1 and Act R1 classes. While apparent loss could be explained by insufficient recovery of gene sequences within the genome and transcriptome sequences examined, the divergent nature of the TGF β ligands found in this species may suggest that these sequences are instead evolving to signal in a derived fashion.

Schistosome receptor complements have been studied elsewhere in depth (Davies et al., 1998; Forrester et al., 2004), and in many ways their complements represent a surprising finding, given the presence in *S. mansoni* of only two TGF β-like ligands. These complements do not map exactly onto the canonical cassette, but, as hypothesised in Osman et al. (2006) and earlier in this thesis, their quantity

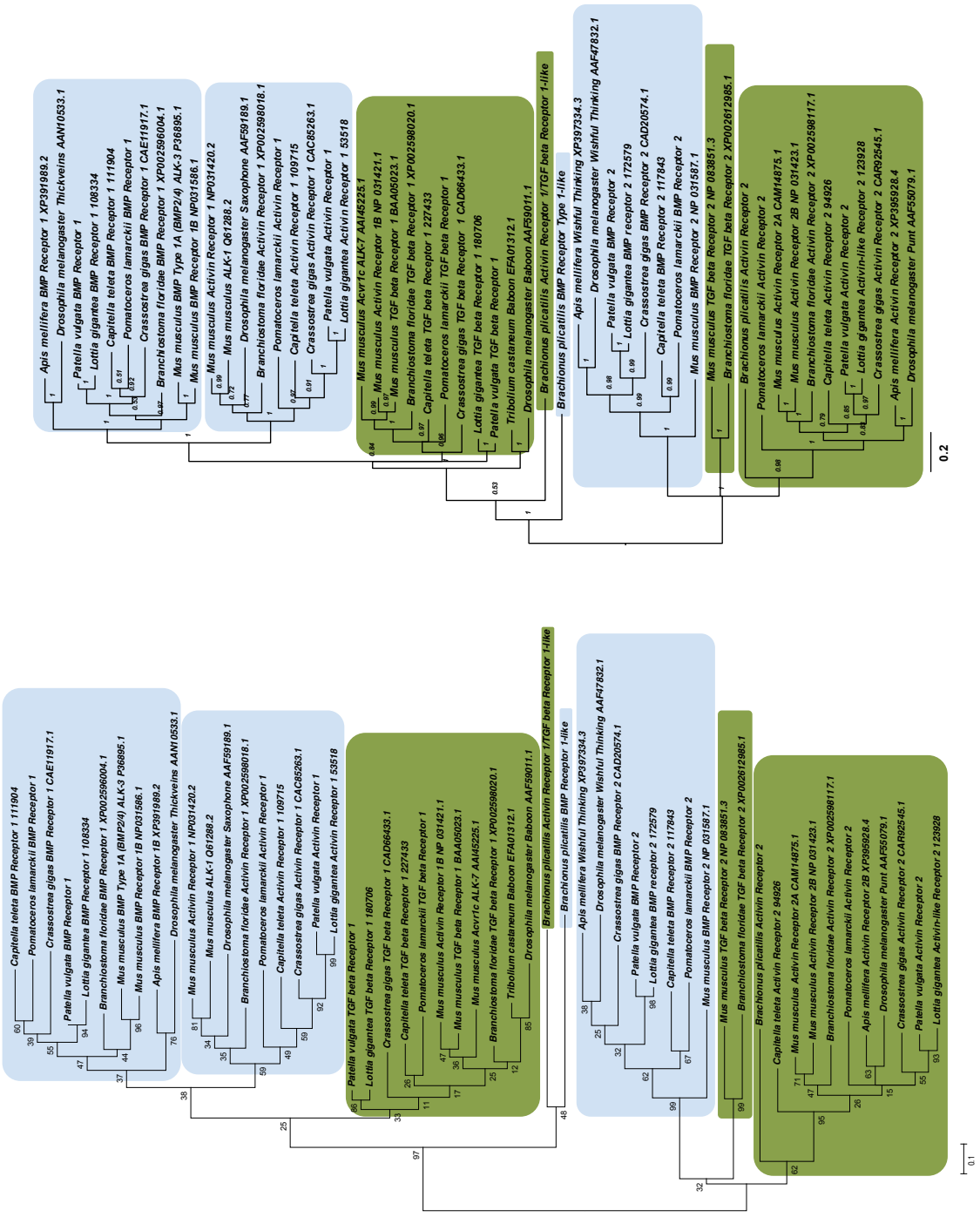


Figure 5.9: *TGF and BMP receptor molecule interrelationships across the Metazoa, as determined by maximum likelihood (Tamura et al 2011, left) and Bayesian (Huelsenbeck and Ronquist 2001, right) methods. Alignment generated by MAFFT (Katoh & Standley, 2013) using the G-iNS-i strategy resulting in a 137 informative amino acid alignment spanning the protein kinase domain (PFAM PF00069), which can be seen in Appendix C, Fig. 8. Both phylogenies determined using the WAG mode (Whelan & Goldman, 2001). Bootstrap percentage (of 1000 replicates) and posterior probabilities can be seen at the nodes of ML and Bayesian trees respectively. Scale bars represent substitutions per site at given distances.*

when compared to the few ligands encoded in its genome may suggest that these molecules respond to host, rather than endogenous, signalling cues.

The relative conservation of ALK receptor sequences within annelids and molluscs confirms the suggestion of Herpin et al. (2005) with regard to the broad conservation of ALK receptor diversity across metazoan evolution. It is particularly interesting that Schistosomes seem to have the apparatus for both BMP and TGF-like signalling, given that they only possess TGF-like ligands. As noted above, it has been suggested that they can utilise host ligands (Osman et al., 2006), a conclusion that is supported by my findings, given the extensive loss in schistosomes of other components of this pathway and in their genomes generally, it would be unusual for them to retain these receptors without using them.

Nodal itself signals via Activin Receptor 2 (Type 2) and ALK 4/5/7 /TGF β R1 (Type 1) receptors. Canonical forms of these are found in all annelid and mollusc species examined, with a canonical Activin Receptor 2 and divergent TGF β R1 found in *B. plicatilis*. This suggests, especially for the Trochozoans studied, that Nodal signalling operates through canonical ALK receptor operation, although this has yet to be functionally tested.

5.3.3 Smad Proteins

Smad proteins play a key role in transducing extracellular signals into an intracellular response. Of all the parts of the TGF β signalling cascade investigated in the present study, it is these molecules that show the least loss and disparity in number across the Metazoa, presumably due to the pleiotropic effects that would happen if loss was to occur. Smad molecules respond to a diverse range of incoming signals, as can be seen in Fig. 1.5, and even if loss of one ligand occurs in a lineage, a Smad may still be responsible for passing on the message brought by another ligand. Such pleiotropy means that loss may be less tolerated by natural selection at this step of the signalling cascade than others.

All lophotrochozoan species studied in the present investigation were found to contain at least one of each of the four major families of Smad molecule (5.10. The short branch lengths generally found outside the inhibitory Smad (Smad6/7 or

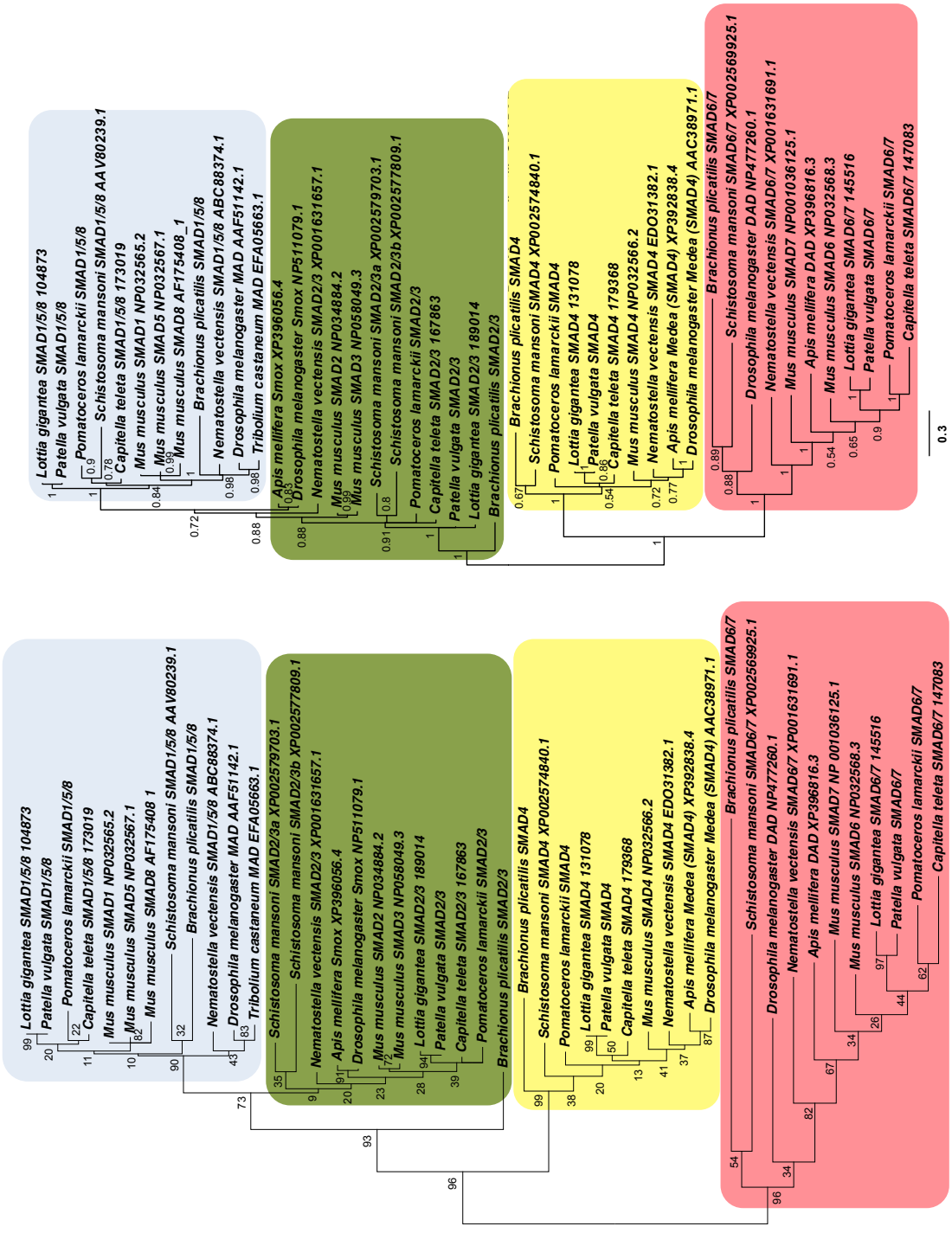


Figure 5.10: *Smad and Dad interrelationships across the Metazoa, as determined by maximum likelihood (Tamura et al 2011, left) and Bayesian (Huelsenbeck and Ronquist 2001, right) methods. Alignment generated by MAFFT (Katoh & Standley, 2013) using the G-iNS-i strategy, with the section used for analysis a 139 informative amino acid region spanning the MH2 domain (Pfam PF03166) which can be seen in Appendix C, Fig 9. Both phylogenies determined using the WAG mode (Whelan & Goldman, 2001). Bootstrap percentage (of 1000 replicates) and posterior probabilities can be seen at the nodes of ML and Bayesian trees respectively. Scale bars represent substitutions per site at given distances.*

Dad) clade also point to generally constraining selection on these molecules. Both *B. plicatilis* and *S. mansoni* show longer branch lengths for these sequences, however, which may be a result of co-evolution to interact with the divergent receptor cassettes also seen in these phyla. We note that I do not recover a monophyletic clade of Smad 2/3 sequences in my Bayesian analysis, which is largely due to the small differences in sequence between the R-Smad clades. I can confidently hypothesise, however, due to the high degree of conservation in the Smad family, that Smad 2/3 proteins play a conserved role in transmitting Nodal signals through to the nucleus of lophotrochozoan cells.

5.3.4 Cripto/Cryptic

The Cripto/Cryptic/TDGF1/EGF-CFC family of membrane anchored proteins are vital co-receptors for Nodal activity. They have traditionally been seen as a deuterostome novelty, although it is clear from examination of my datasets that they are present in the Lophotrochozoa and probably reflect an Urbilaterian stem lineage novelty that has been lost in the Ecdysozoa along with the wider Nodal cascade. Figure 5.11 shows the phylogeny of metazoan Cripto family members, as recovered by my analysis.

Cripto functionality in the Lophotrochozoa could be particularly interesting for the study of the establishment of L/R asymmetry due to the restricted receptor complement found in this clade as compared to the Deuterostomia. While activins and other TGF β like ligands signal through exactly the same receptors as Nodal in the Lophotrochozoa, Cripto plays an important role in down-regulating Smad activation when Nodal is not the ligand bound to activin receptors (Gray & Vale, 2012). This means that activin-like signals can result in low levels of Smad activation, which activates specific genes more responsive to low levels of Smad2/3 signalling. With the aid of Cripto, Nodal signals result in high, sustained signalling, activating a different suite of genes. This allows two distinct signalling pathways to propagate from a single pair of receptors.

The presence of *Cripto* within the Lophotrochozoa is therefore far more significant than just the additional retention of a single ancestral gene. It implies that the

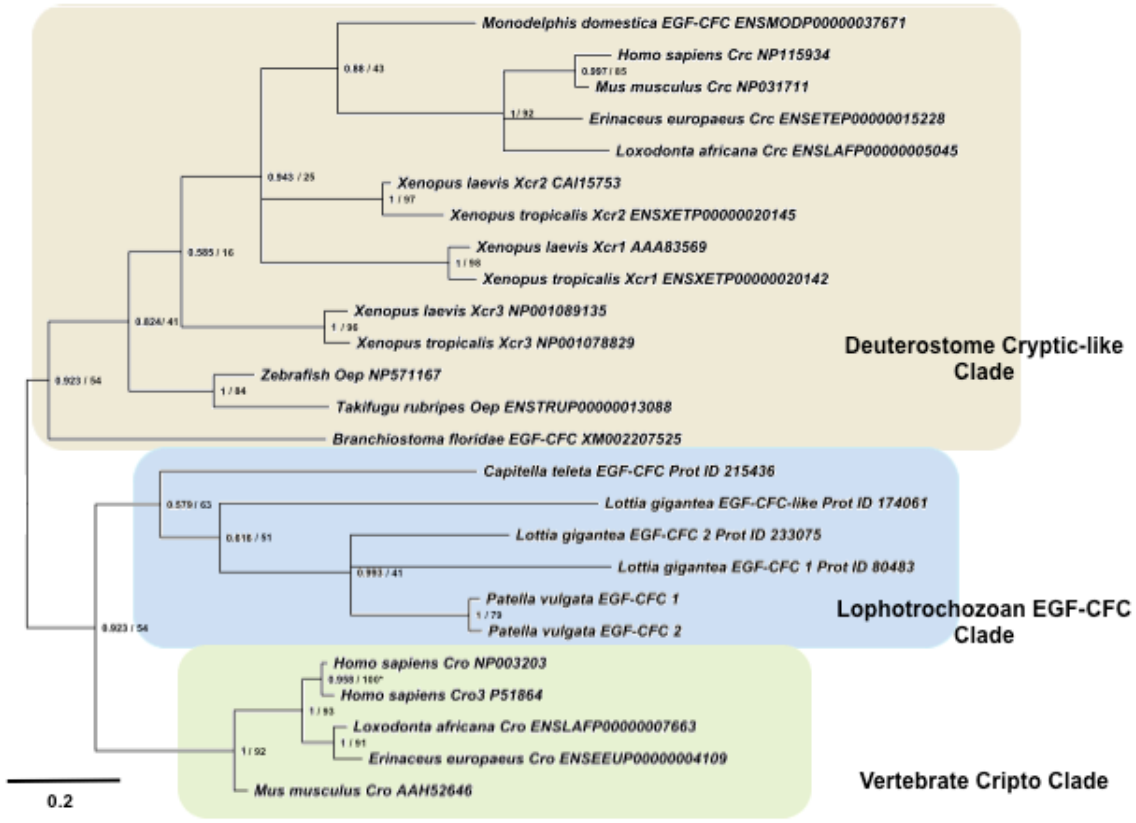


Figure 5.11: *Phylogeny of metazoan Cripto/Cryptic/TDGF1/EGF-CFC genes, based on a 64 informative amino acid alignment of an area spanning the CFC domain (Pfam ID 09443), which can be seen in Appendix C, Fig. 10. Phylogeny shown is the result of ML analysis. Both phylogenies determined using the WAG mode (Whelan & Goldman, 2001). Posterior probabilities/bootstrap percentage (of 1000 replicates) and can be seen at the base of nodes. Scale bars represent substitutions per site at given distances.*

nuances of Nodal signalling in establishing L/R asymmetry were already in place in the last common ancestor of Deuterostomia and Lophotrochozoa. It also acts as further confirmation, if any were needed, that *Nodal* and *Pitx* were not independently recruited in these superphyla for a role in establishing asymmetry.

The extensive duplication of mollusc *Cripto* genes is of some interest, particularly in light of the apparent duplication of Nodal ligand in *P. vulgata*. Whether this duplication is related to a role in modulating Nodal signalling will require confirmation of expression and functional testing, but represents a fascinating possibility. It should be noted also that mollusc Criptos also exhibit an apparent Pacifastin 1 (Pfam ID PF05375) domain upstream of the canonical EGF site. These domains are serine protease inhibitors (Simonet et al., 2002), and Criptos in the Mollusca could therefore enhance Nodal signalling by protecting ligands from inhibitors, although I note Follistatins have similar domains but act to inhibit signalling.

Cripto has also been noted as having a role independent of the cell surface when the GPI anchor, which normally holds it to the membrane, is cleaved enzymatically. This results in a soluble protein that can have an action at a distance from the cell that transcribed and translated it. The production of a reliable antibody to Cripto would allow the testing of this in the Lophotrochozoa, but at present I am unable to discern whether such an affect is found in this clade.

5.3.5 Dan/Cerl/Coco/Prdc/Cerberus/Gremlin

Members of the wider DAN-like family sequester ligands, preventing them from binding to receptors and activating signalling cascades. This family has diverged into a large and confusingly named clade of genes, especially in vertebrates, where the 2R whole genome duplication event likely allowed sub- and neo-functionalisation to occur. The results of phylogenetic analyses of members of this gene family from species across the Metazoa (Fig. 5.12) reveal how this often difficult-to-catalogue family has evolved.

It appears that the cassette of DAN-family members found in the common ancestor of deuterostomes and protostomes may have resembled that of *N. vectensis*, with two homologues giving rise to the diversity seen today. It is possible that the

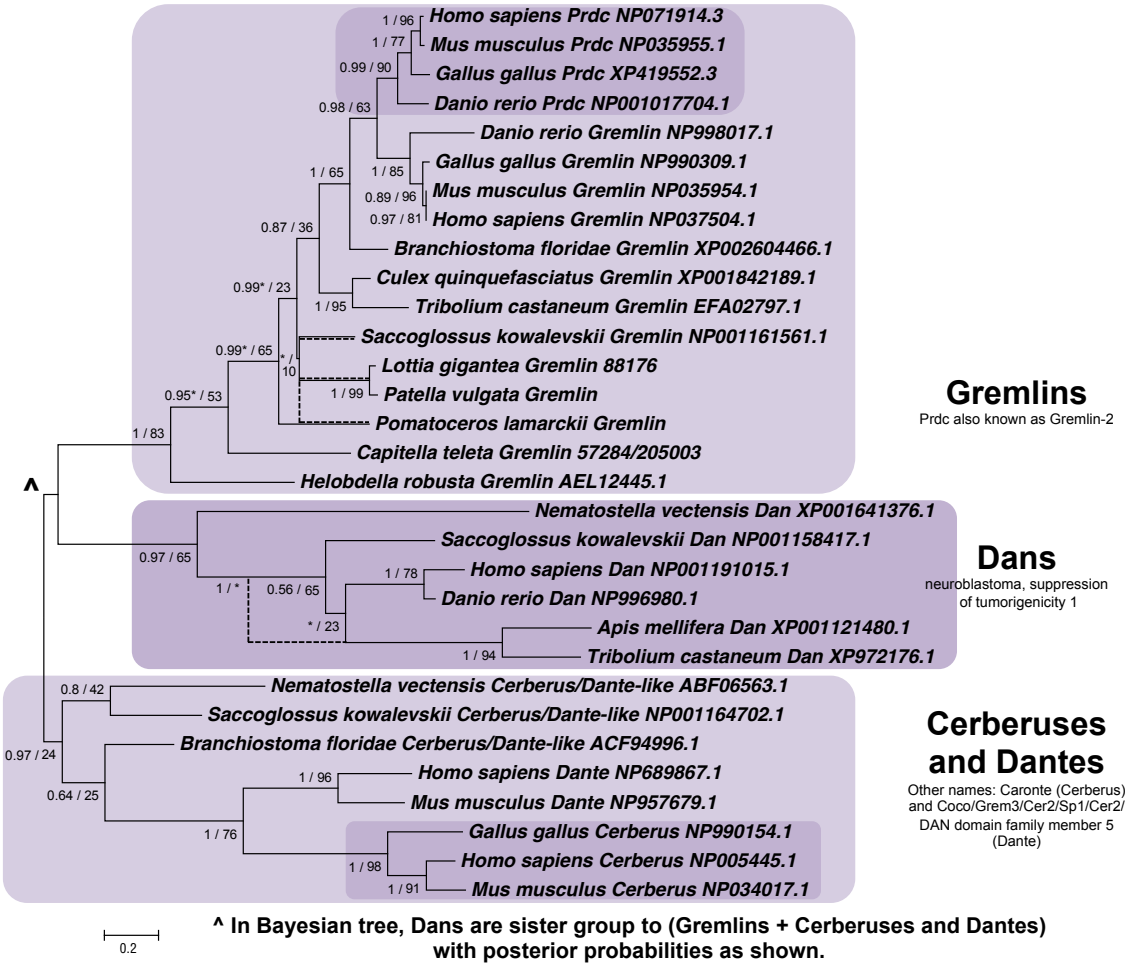


Figure 5.12: *DAN family interrelationships across the Metazoa, as determined by maximum likelihood (Tamura et al., 2011) and Bayesian (Huelsenbeck & Ronquist, 2001) methods. Alignment generated by MAFFT (Katoh & Standley, 2013) using the G-iNS-i strategy. Phylogenies calculated on the basis of an 87 informative amino acid alignment spanning the DAN domain (Pfam ID PF03045) which can be seen in Appendix C, Fig. 11. Phylogeny shown is the result of ML analysis, with differences in topology using Bayesian methods indicated with a dotted line. Both phylogenies determined using the WAG mode (Whelan & Goldman, 2001). Posterior probabilities/bootstrap percentage (of 1000 replicates) and can be seen at the base of nodes. Scale bars represent substitutions per site at given distances.*

Cerberus/Dante family represents a deuterostome innovation - despite the placement of *N. vectensis* Cerberus/Dante-like (ABF06563.1) at the base of this clade, given the widespread presence of Gremlins across the Metazoa, and previous study in this group in Cnidarians (Rentzsch et al., 2006), it perhaps would be more parsimonious to infer that it is in fact a Gremlin, rather than inferring loss of a cnidarian Gremlin and protostome Cerberus/Dante-like factors. We cannot distinguish between these alternatives definitively with the data available, but this hypothesis will be easily testable with the advent of broader sequence availability.

No wider DAN family genes can be found in schistosomes or in the rotifer *B. plicatilis*, and loss of portions of this family are prevalent in other species - *D. melanogaster*, for example, has no DAN family members in its genome (Van der Zee et al., 2008), and DANs sensu stricto have been lost across the Lophotrochozoa as can be seen in Table 5.1. DANs sequester a variety of ligands, with specificity varying depending on the DAN protein examined. Some are specific to the BMP-like signalling pathway, while others (such as Cerberus) can inhibit Nodal in the TGF-like pathway. Gremlin is known to be active in *H. robusta*, targeting BMP 2/4 preferentially (Kuo & Weisblat, 2011). Nodal in particular is known to be regulated by Cerl2, a DAN family member (Inacio et al., 2013).

Much investigation is required, however, before inference can be made as to the wider roles of the *Gremlin* genes that remain in the Lophotrochozoa, given the wide range of actions that these genes are known to have in more established model organisms. As DANs can regulate both BMP- and TGF-like ligands, it is imminently possible that the Gremlin homologues could be involved in Nodal regulation.

5.3.6 Chordin

Chordin is a BMP regulatory molecule, which sequesters ligands by binding to them. It is best known for its role antagonising BMP 2/4 (also known as Dpp) in dorsoventral patterning. A similar molecule, Chordin-like, has also been identified, initially as a BMP 4 antagonist in the chick (Sakuta et al., 2001).

Searches through the lophotrochozoan genomes examined in the present project suggest that both *Chordin* and *Chordin-like* genes are present within the Lophotro-

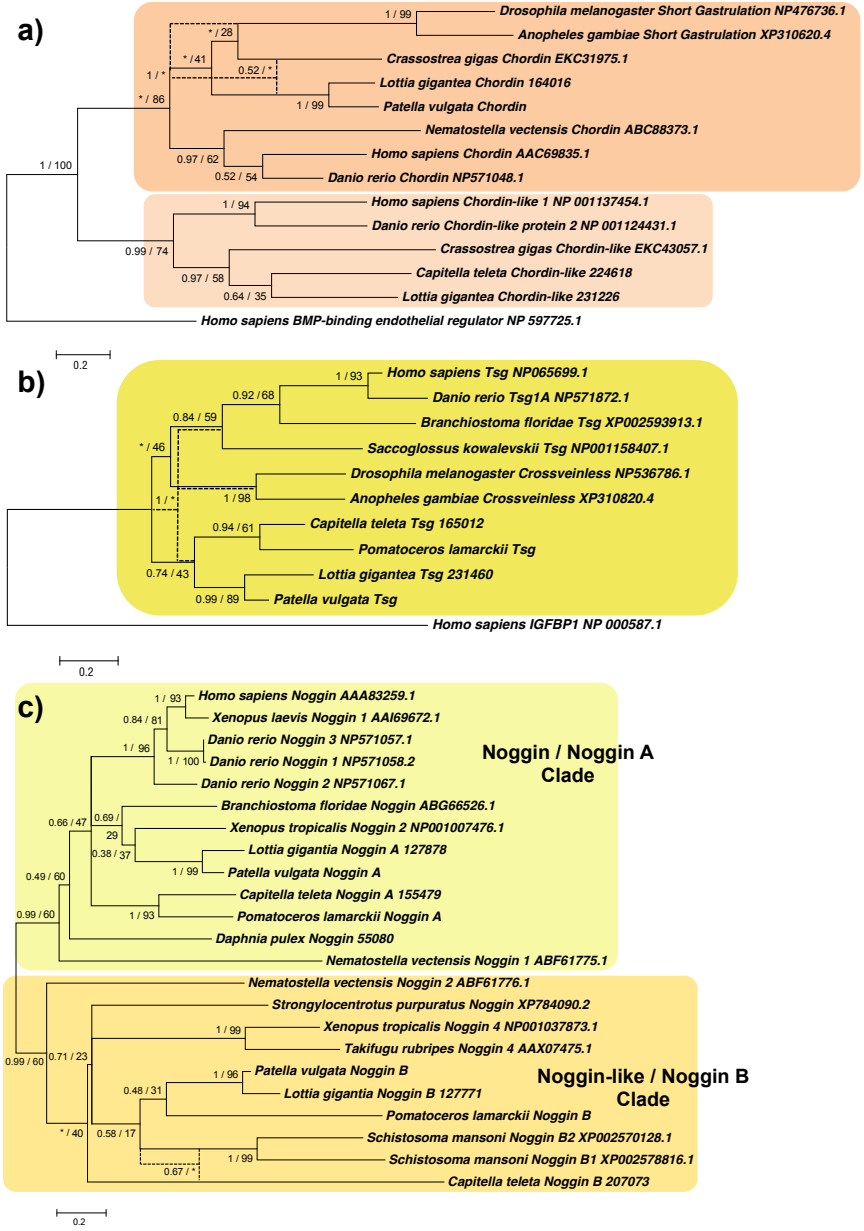


Figure 5.13: *Chordin (a), Twisted Gastrulation (b) and Noggin (c) interrelationships across the Metazoa, as determined by maximum likelihood (Tamura et al., 2011) and Bayesian (Huelsenbeck & Ronquist, 2001) methods. Phylogeny shown is the result of ML analysis, with differences in topology using Bayesian methods indicated with a dotted line. Chordin phylogeny based on a 120 informative amino acid alignment of the von Willebrand factor type C/D domains and C8 domains generated by MAFFT (Katoh & Standley, 2013) using the G-iNS-i strategy, analysed under the JTT model (Jones et al. (1992), ML) and Dayhoff model (Dayhoff et al 1978, Bayesian), which can be seen in Appendix C, Fig. 12. Tsg phylogeny based on a 146 informative amino acid global alignment generated by MAFFT using the G-iNS-i strategy which can be seen in Appendix C, Fig. 13, rooted with IGFBP after Vilmos et al 2001, with both phylogenies determined using the WAG model (Whelan & Goldman, 2001). Noggin phylogeny based on a 103 informative amino acid global alignment generated by MAFFT using the G-iNS-i strategy which can be seen in Appendix C, Fig. 14, followed by analysis using the WAG model in both cases. Posterior probabilities/bootstrap percentage (of 1000 replicates) and can be seen at the base of nodes. Scale bars represent substitutions per site at given distances.*

chozoa, as can be seen in Fig 5.13a. No Chordin or Chordin-like sequences were found in either the rotifer *B. plicatilis* or in the Platyhelminthes, although other more derived chordin-like families (eg CRIM) were not examined in the above analysis. The presence of clear chordin-like genes in the Lophotrochozoa confirms the hypothesis that chordin-like genes were present in the Bilaterian common ancestor, rather than being a deuterostome novelty, and corroborates the putative assignment of *T. adhaerens* Chordin-like homologue within this family rather than as a canonical Chordin. It is more likely that these *Chordin* genes are involved in direct regulation of BMP-like ligands rather than ligands that signal via the TGF β-like route, but this could feed back into L/R asymmetry if it is found that BMP2/4 antagonises Nodal signalling as is found, for instance, in sea urchins.

5.3.7 Twisted Gastrulation (Tsg)

Tsg is a modulator of BMP signalling, although its mode of action is yet to be fully understood. As well as binding BMPs, it has been suggested that Tsg might also act as a promoter of BMP signalling, by freeing ligands from Chordin after the Chordin-ligand complex has been cleaved by Tolloid (Oelgeschlager et al., 2000; Scott et al., 2001).

The phylogenetic relationships of a number of metazoan Tsg sequences can be seen in Fig 5.13b. Mollusc and annelid species all possess a single Tsg homologue, while all schistosomes examined and the rotifer *B. plicatilis* seem to have lost theirs, as none can be found in their genomes or transcriptomes. This could correlated to the markedly reduced BMP complements of these species - without the ligand diversity found in other species, it is perhaps unsurprising that regulatory mechanisms have also been lost. As with Chordins above, a direct role for Tsg in regulating Nodal is perhaps unlikely, but would be worth investigating in the context of a role for BMP antagonism of L/R asymmetry if this was found to occur.

5.3.8 Noggin

Noggins are proteins which act to interfere with canonical TGF signalling by sequestering ligands before they can bind to their receptors, as with Chordin. Noggins primarily have been shown to interact with BMP-like ligands. My data (Fig. 5.13c) confirms the study of Molina et al. (2011), who posited the existence of two major clades of Noggin across the Metazoa, a canonical Noggin clade, and a less well categorised Noggin-like clade. We find both kinds in all Lophotrochozoans examined with the exception of the rotifer *B. plicatilis*, which lacks Noggins entirely. It is well documented that canonical Noggins and Noggin-like proteins are absent from the genomes of some ecdysozoan model organisms (Van der Zee et al., 2008) although these are found in some other arthropods (Duncan et al., 2013): the presence of a Noggin in the crustacean *Daphnia pulex* and Noggin and Noggin-like in the hemipterans *Acyrthosiphon pisum* and *Rhodnius prolixus* suggests that this loss happened independently in some insects and nematodes.

The presence of *Noggin-like* genes in schistosomes again raises questions as to their function they could interact with the TGF-like ligands found in these species, quite unlike their role in BMP regulation in other superphyla, or instead they may be involved in regulating exogenous signals or other processes entirely. We do not detect the diversity of Noggin sequences in other lophotrochozoans that are found in *S. mediterreanea*, where eight *Noggin* and *Noggin-like* genes are found (Molina et al., 2011). We therefore posit that this expansion is lineage specific, and may be related to species-specific roles for ligands, perhaps in regeneration. As Nodal is missing from the platyhelminth species yet sequenced, I do not posit a role in the establishment of L/R asymmetry for these in this clade, but investigation in the Annelida or Mollusca is warranted given the differential loss which has occurred in this family of genes across the Bilateria, as it is possible that the restricted complements of Noggin genes in more well-studied phyla have lost a portion of an ancestral role in regulating TGF-like signalling.

5.3.9 Follistatin

Follistatin binds to and inhibits TGF β ligands, particularly Activins, although it can bind to other ligands, even those in the BMP-like family. Canonical Follistatin sequences are found in the Lophotrochozoa, as can be seen in the phylogenetic tree presented in Fig. 5.14a. We note that Platyhelminthes possess Follistatin, despite not having a canonical Activin ligand for it to bind to. As noted earlier, it is suspected that parasitic Platyhelminths can utilise host ligands, so it is possible that these Follistatins inhibit the action of these, but it is perhaps more likely that Follistatins bind the more derived Activin-like ligands found in these species. No Follistatin sequence could be identified within my *B. plicatilis* dataset, although why it is lost in the rotifer but retained in schistosomes is unknown, despite the similar loss of canonical activins.

Only Follistatin homologues with the classical Follistatin/Osteonectin-like EGF domain followed by three Kazal-type serine protease inhibitor domains were examined by the present study, with the exception of the *N. vectensis* homologue, which lacks a clearly identifiable EGF domain. Extensive diversification of this gene has taken place within the deuterostome lineage as early as the ambulacraria/chordate split, with loss, rearrangement and gain of various domains resulting in a total of up to five *Follistatin-like* genes, descended from an ancestral *Follistatin* sequence. Given the retention of *Follistatin* in schistosomes and the known affinity of Follistatin to both activin and BMP-like ligands, it is quite possible that Nodal signalling could be regulated by this gene in the Lophotrochozoa.

5.3.10 Tolloids

Tolloids are found extracellularly, and cleave regulators of TGF β signalling when they have formed complexes with free ligands, releasing the ligand they have bound and allowing signalling to occur. Fig. 5.14b shows the result of phylogenetic analysis of Tolloid sequences from a range of metazoan species, rooted with the homologue found in *N. vectensis*.

A single Tolloid homologue was found in both mollusc species examined, in

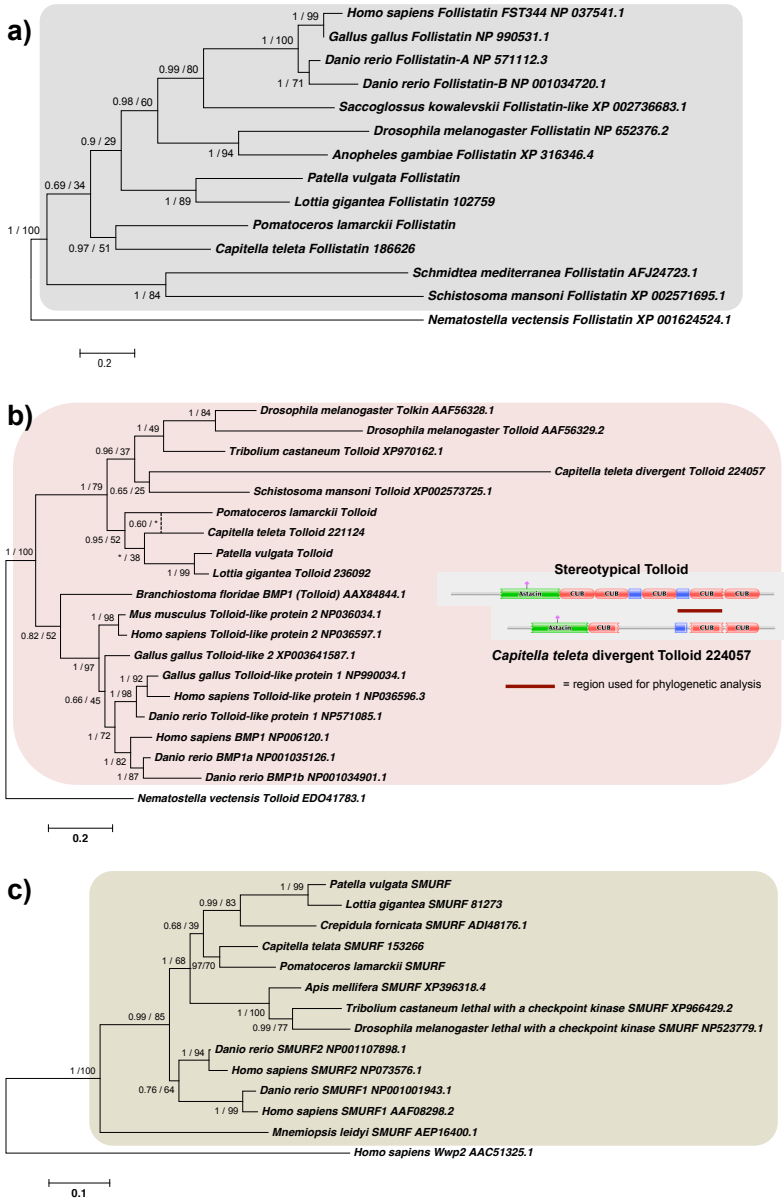


Figure 5.14: *Follistatin (a), Tolloid (b) and Smurf (c) interrelationships across the Metazoa, as determined by maximum likelihood (Tamura et al., 2011) and Bayesian (Huelsenbeck & Ronquist, 2001) methods. Phylogeny shown is the result of ML analysis, with differences in topology using Bayesian methods indicated with a dotted line. Follistatin phylogenies inferred on the basis of a MAFFT alignment (Katoh and Standley 2013, E-INS-i strategy) with 227 informative sites, which can be seen in Appendix C, Fig. 15, analysed under the WAG model (Whelan & Goldman, 2001). Tolloid phylogeny shown generated according to the JTT model (Jones et al. (1992), ML) / Blosum model (Henikoff & Henikoff (1992), Bayesian) from a 150 informative amino acid alignment spanning the calcium-binding EGF domain and immediately proceeding the Cub domain as shown on inset, which can be seen in Appendix C, Fig. 16, generated using MAFFT under the G-INS-i strategy. Inset shows the domain structure of canonical Tolloid proteins, along with that of the C. teleta divergent paralogue, presented from the Pfam database (Punta et al., 2012). The Smurf phylogeny was calculated under the JTT model, based on a 233 informative amino acid alignment generated using MAFFT using the G-INS-i strategy, which can be seen in Appendix C, Fig. 17. Posterior probabilities/bootstrap percentage (of 1000 replicates) and can be seen at the base of nodes. Scale bars represent substitutions per site at given distances.*

P. lamarckii, and in all schistosome species (*S. mansoni* shown on tree), although it appears to have been lost from the rotifer genome. *C. teleta* possesses two homologues, one with the canonical domain structure (inset, Fig. 5.14b) and one with a highly divergent structure, although this does not appear to be the result of misannotation of the *C. teleta* genome, and no equivalent is found in *H. robusta* or *P. lamarckii*. Two Tolloids (Tolloid and Tolkin) are found in *D. melanogaster*, and my analysis corroborates the suggestion of Van der Zee et al. (2008) that this Tolkin homologue is in fact a lineage specific paralogue, specific to the Diptera rather than found protostome-wide. It would be expected that this gene retains its ancestral functionality in the Lophotrochozoa, although the role of the divergent *C. teleta* gene will require more rigorous examination to discern.

5.3.11 SMURFs

SMURF (SMAD specific E3 ubiquitin protein ligase) proteins regulate TGF β signalling by a number of mechanisms, generally involving the targeting of R-Smads for degradation, but other roles have also been posited for these proteins. Two SMURF genes are characterised in vertebrates, but to date only single orthologues have been found in sequenced invertebrate species. My phylogenetic analysis (Fig. 5.14c) of these genes from species across the Metazoa, suggests a paralogous relationship between the two homologues found in vertebrate model species.

SMURF proteins appear to have originated within the early metazoan lineage, and clear homologues can be identified for these in *M. leidyi* and *T. adhaerens*, although not to date in *N. vectensis*. While single canonical SMURF orthologues are readily identifiable in a variety of protostome species, no SMURF proteins can be found in the genomes of *C. elegans*, *B. plicatilis* or the schistosome species. These species do, however, possess other E3 ubiquitin-protein ligases which readily cluster with nedd-4-like E3 ubiquitin-protein ligases by blast identity and in preliminary trees, and are probably homologues of this class of protein. We note that two *C. fornicata* SMURF homologues are present in the NCBI nr dataset, although one (ADI48175.1) is only a fragmentary sequence with 100% similarity to its homologue at the amino acid level, and as such was not used in my phylogenetic analysis. It is

very likely that the ancestral role of SMURF protein in regulating Nodal signalling at the Smad level is retained in the Lophotrochozoa.

5.3.12 BMP and Activin Membrane-Bound Inhibitor (BAMBI)

BMP and activin membrane-bound inhibitor is a pseudoreceptor that competes with true type 2 receptors for ligand binding (Onichtchouk et al., 1999). The presence of this gene has been noted in protostomes previously, particularly by Van der Zee et al. (2008) and Huminiecki et al. (2009), although the origin of these is less clear. Searches through the genomes of a variety of non-bilaterian metazoans revealed no trace of a BAMBI homologue, and it therefore seems likely that BAMBI emerged on the lineage leading to the last common ancestor of protostomes and deuterostomes.

Figure 5.15a shows the results of phylogenetic inference into the interelationships of BAMBI homologues from a variety of species in the Protostomia and Deuterostomia. No BAMBI orthologue could be identified in *C. elegans*, or the Schistosome species or Rotifera sampled.

We note that a putative BAMBI sequence has been described in a previous publication in *S. mediterranea* (Gavino & Reddien, 2011), but in my investigations this sequence (ADX42731.1) appears to more readily resemble canonical ALK receptors, clustering with these sequences in the course of phylogenetic analysis, although the partial sequence provided does not include the intracellular serine/threonine kinase domain required for signalling. We are unable to test whether this domain absence is the result of fragmentary sequence or is truly absent in the complete protein, but if it is the latter this may represent the re-evolution of a trait (a *BAMBI*-like gene) present more generally across the Metazoa. It is likely that the *BAMBI* genes found in the Lophotrochozoa perform a canonical role in regulating BMP signalling, but it is unlikely there is a direct link to regulation of Nodal signalling for this gene.

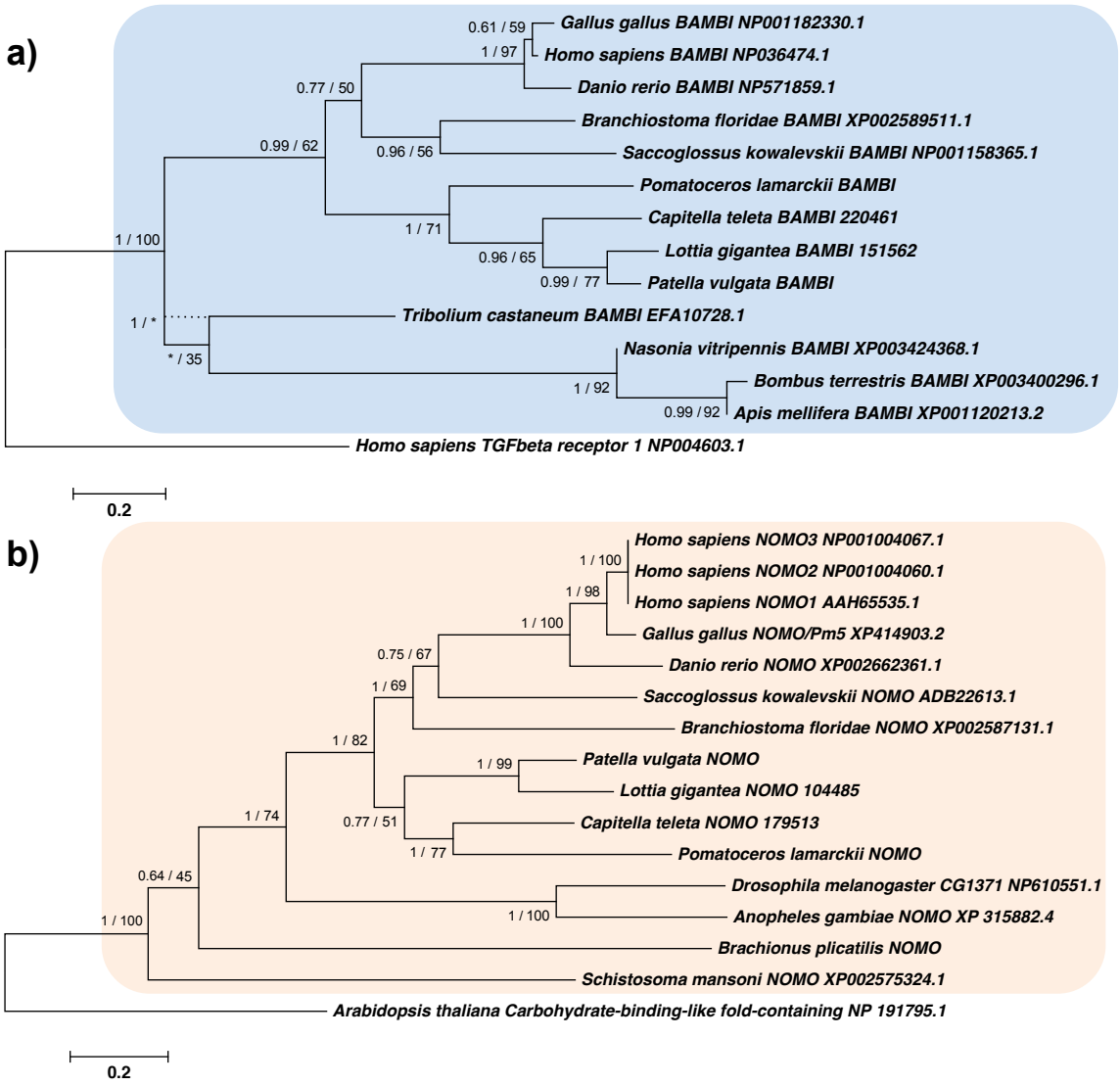


Figure 5.15: *Bambi (a) and NOMO (b) interrelationships across the Metazoa, as determined by maximum likelihood (Tamura et al., 2011) and Bayesian (Huelsenbeck & Ronquist, 2001) methods. Phylogeny shown is the result of ML analysis, with differences in topology using Bayesian methods indicated with dotted lines. Bambi phylogenies determined using the JTT model (Jones et al., 1992), NOMO phylogenies under the WAG model (Whelan & Goldman, 2001). Bambi phylogeny based on 73 informative amino acid alignment created using MAFFT (Katoh & Standley, 2013) under the L-INS-i model, spanning the transmembrane domain, incorporating some conserved regions both extra- and intracellularly, and can be seen in Appendix C, Fig. 18. NOMO phylogeny based on G-INS-i alignment by MAFFT, with 439 informative positions, rooted with an apparent NOMO sequence found in Arabidopsis thaliana (NP_191795.1), and can be seen in Appendix C, Fig. 19. Numbers at node reflect Bayesian posterior probabilities/bootstrap support (1000 replicates, JTT model/WAG model, all default priors) respectively. Posterior probabilities/bootstrap percentage (of 1000 replicates) and can be seen at the base of nodes. Scale bars represent substitutions per site at given distances.*

5.3.13 NOMO

NOMO (Nodal Modulator, previously known as pM5), is known for its role in directly antagonising Activin/Nodal signalling, in concert with Nicalin, with which it forms a transmembrane complex at the endoplasmic reticulum (Haffner et al., 2004). These complexes are similar to complexes that regulate γ-secretase activity, but do not perform the same roles, as shown in Zebrafish rescue assays. Instead they have been shown to regulate the formation of mesendoderm by attenuating Nodal signalling. NOMO is, however, still under-researched, with little known of its molecular mode of action (Dettmer et al., 2010).

Phylogenetic analysis of NOMO sequence from a range of metazoans can be seen in Fig. 5.15b, rooted with an apparent NOMO sequence found in Arabidopsis thaliana. NOMO is found in every genome examined in this work, suggesting a crucial role in cellular signalling, even in species where Nodal is not found (for example, in the Ecdysozoa).

NOMO-like sequences are well described outside the Metazoa and are suggested to be found throughout the Eukaryota (HomoloGene:13810). It seems that NOMO is particularly poorly named given its conservation outside of clades where Nodal exists, and in many cases where TGF β signalling is entirely absent. Some recent analyses have suggested that NOMO is involved in the regulation of nicotinic acetylcholine receptor functionality (Almedom et al 2009, for example), and this, along with the widespread conservation of NOMO throughout the Eukaryota, suggests that this complex has been recruited by metazoans for further regulation of TGF β signalling - although much mechanistic work is required to untangle how this regulation is performed.

5.3.14 General Lophotrochozoan TGF β Componentry and the Evolution of TGF β signalling

The results of my analysis of lophotrochozoan TGF β cassettes can be seen in Table 5.1 and for TGF β ligands in particular in Table 5.2. In general, the complements of the annelid and mollusc species considered in my analysis resemble that of in-

vertebrate deuterostomes more than they resemble those of ecdysozoan models. Ligand diversity seems more pronounced than that seen in the Ecdysozoa, and many regulatory components, such as the Noggin family, seem to be the result of Ecdysozoa-specific, rather than protostome-wide, loss.

The most striking differences between the cassettes of other metazoans and my sampled datasets are in the case of the rotifer *B. plicatilis* and the planarian *S. mansoni*. Extensive loss seems to have occurred at the ligand and regulatory levels in these phyla, with Smad and receptor diversity relatively unchanged, particularly in *S. mansoni*.

These apparent lost genes could be missing from my datasets due to inadequate assembly of these genomes. This is more likely for *B. plicatilis* than for *S. mansoni*, as this genome results from a much deeper sequencing effort. Other planarian genomes also exist, allowing assessment by comparison to outgroups in cases of loss. Analyses of the *B. plicatilis* dataset, however, suggests that the substantial majority of genes present in that species are recovered by the genome dataset presented in this work, and while some missing genes are perhaps to be expected from my analysis, I would not expect these to be numerous.

It should be noted that the TGF β componentry of *S. mansoni* is very similar to that of *S. japonicum* and in most cases to that of *S. mediterranea*, and it has been chosen as a representative and very well described member of its phylum for the purposes of comparison. While some differences exist - *S. mediterranea*, for example, has 8 Noggin sequences listed on Genbank - these appear to be the result of lineage specific loss or gain, and the *S. mediterranea* genome is still generally cited as a draft resource. It is possible that more basal planarian species will exhibit a more complete TGF β complement than the species currently sampled.

While the internal phylogeny of the lophotrochozoan clade is yet to be fully resolved, most studies place molluscs and annelids as sister groups, with planarian and rotifer data suggesting that these phyla are only distantly related to the Trochozoa sensu stricto. While my sampling may not allow us to trace the evolution of the lophotrochozoan TGF β complement across the entirety of its constituent phyla, I can be confident that the last common ancestor of the Trochozoa, Rotifer

		Ligands	Smads	Receptors	Chordin	Chordin-like	Noggin	Noggin-like	Follistatin	Gremlins	Dans	Dantes	Tsgs	Tolloids	BAMBI	NOMO	SMURFs	References:
Deuterostomes		33	8	13	1	2	1	0	1	2	1	2	1	3	1	3	2	*Huminiecki et al 2009*
	Ciona intestinalis	10	5	7	1	1	1	0	0	1	0	0	1	1	0	1	1	*Hino et al 2003/Huminiecki et al 2009*
Ecdysozoans	*Strongylocentrotus purpuratus*	14	4	6	1	1	0	1	1	1	1	0	1	3	0	2	1	*Lapraz et al 2006*
	Drosophila melanogaster	7	4	5	1	0	0	0	1	0	0	0	3	2	0	1	1	*van der Zee et al 2008*
	Apis mellifera	7	4	5	1	0	0	0	0	0	1	0	1	1	1	1	1	*van der Zee et al 2008*
	Tribolium castaneum	8	4	5	1	0	0	0	1	1	1	0	1	1	1	1	1	*van der Zee et al 2008*
	Caenorhabditis elegans	5	7	3	0	0	0	0	0	1	0	0	0	1	0	1	0	*Savage-Dunn 2005, Huminiecki et al 2009*
Lophotrochozoans	*Capitella teleta*	16	4	5	0	1	1	1	1	1	0	0	1	2	1	1	1	
	Pomatoceros lamarckii	9	4	5	0	0	1	1	1	1	0	0	1	1	1	1	1	
	Lottia gigantea	10	4	5	1	1	1	1	1	1	0	0	1	1	1	1	1	
	Patella vulgata	12	4	5	1	0	1	1	1	1	0	0	1	1	1	1	1	
	Brachionus plicatilis	4	4	3	0	0	0	0	0	0	0	0	0	0	0	1	0	
	Schistosoma mansoni	2	5	5	0	0	1	2	1	0	0	0	0	1	0	1	0	
Diploblasts	*Nematostella vectensis*	6	4	6	1	0	1	1	1	1	1	0	0	1	0	1	0	*Huminiecki et al 2009, Saina and Technau 2009*
	Mnemiopsis leidyi	9	5	4	0	0	0	0	0	0	0	0	0	1	0	1	1	*Pang et al 2011*
	Trichoplax adherens	5	4	4	0	1	1	0	1	0	1	0	0	0	0	1	1	*Huminiecki et al 2009*

Table 5.1: *TGF β complements of a range of metazoan species, as determined in the present manuscript or from previously published work as cited at right, where well annotated examples of these complements have been determined from other studies. NB while no whole-genome analysis of this pathway has yet been performed in the Porifera, I refer the interested reader to Suga et al. (1999), Adamska et al. (2007), and Fig 11 of Pang et al. (2011), where a number of the key families listed above are described in this phylum.*

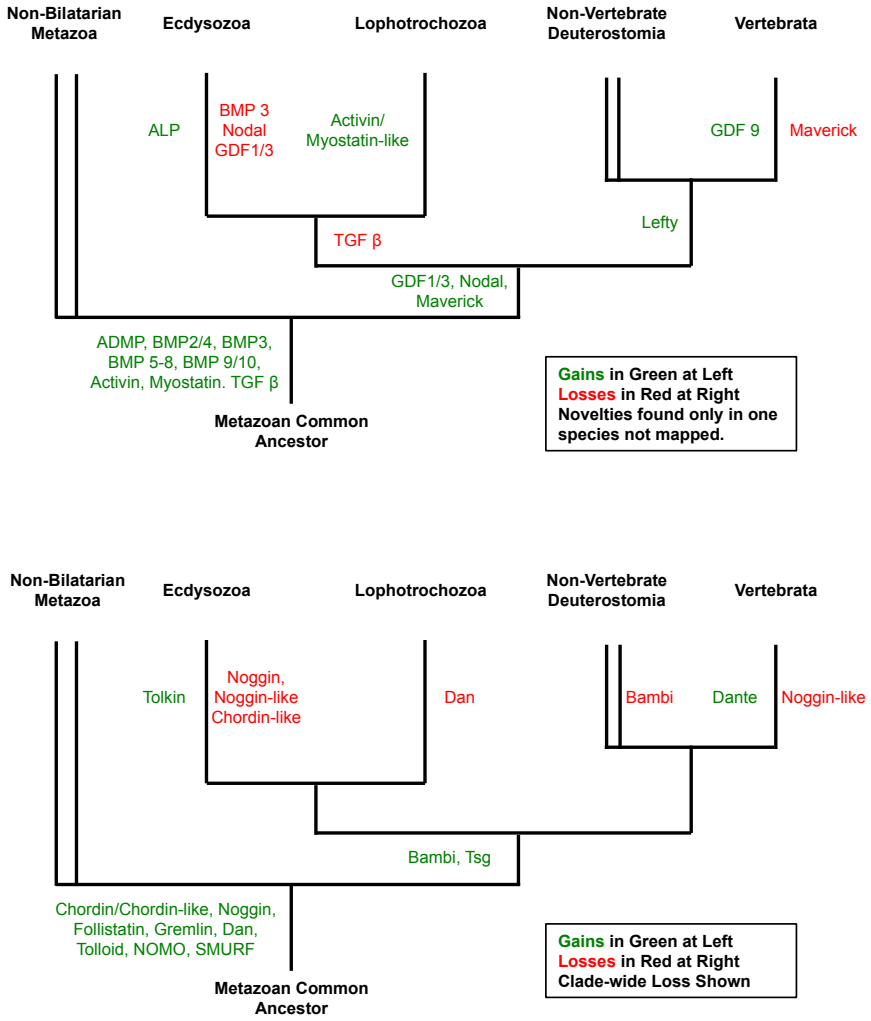


Figure 5.16: *Gain and loss of TGF β ligands (top) and modulators of TGF β signalling (bottom) across the Metazoa, mapped onto a schematic cladogram of the interrelationships of these superphyla. Position of TGF β itself is contingent on the true assignation of TGF β status to M. leidyi protein sequence by Pang et al. (2011), as this hypothesis is only weakly supported by my analysis.*

and Platyhelminthes will be relatively closely related in molecular complement to the Urlophotrochozoan, and my sampling thus allows us to draw inference as to the cassette of that hypothetical organism, by mapping gain and loss onto a schematic phylogeny (Fig. 5.16).

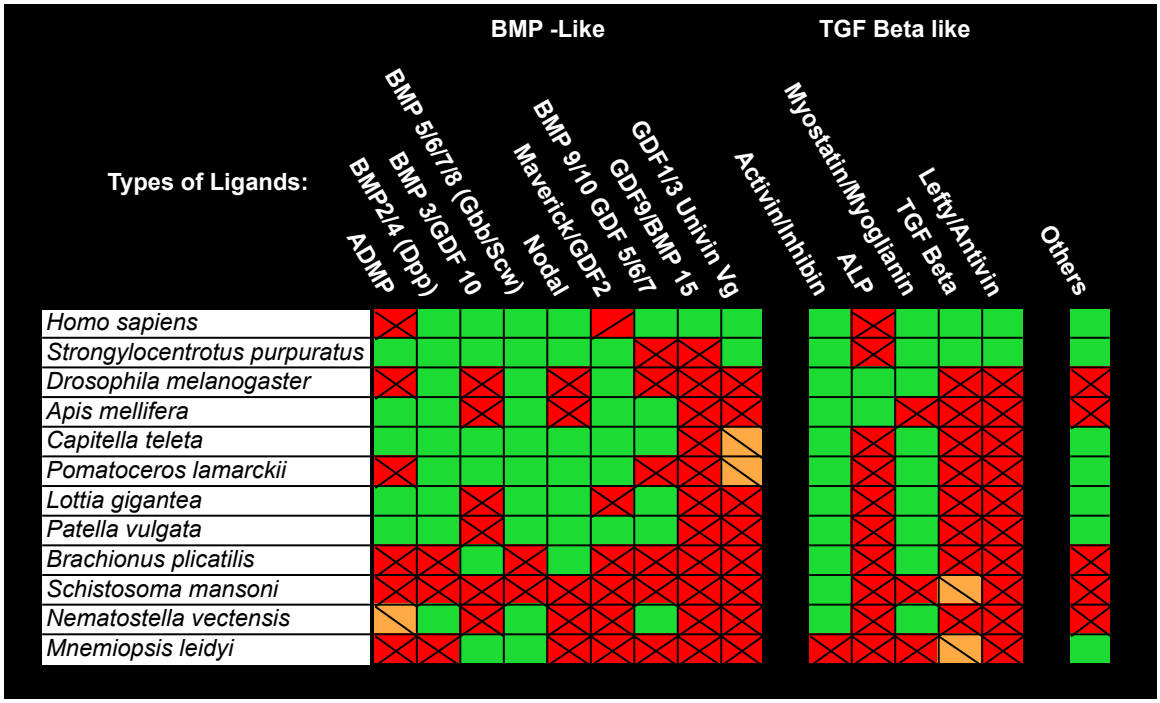


Table 5.2: *TGF β ligand subfamily presence and absence in a range of metazoan species, as determined in the present manuscript or inferred from previously published work. Only species with clearly informative TGF β ligand identity known are listed above. Confirmed identity is shown in green (no lines), tentative identity in orange (single diagonal line) and absence in red (two diagonal lines, forming a cross).*

My work therefore suggests that I might expect the TGF β signalling complement of the Urlophotrochozoan to be very complete, with only Dans *sensu stricto* completely missing from the ancestrally shared regulatory cassette across all lophotrochozoan phyla, and no lophotrochozoan-wide loss seen in the ligand complement. Loss is, however, prevalent in gene families in the planarians and in the rotifer species sampled, suggesting that a diverse TGF β complement may not be necessary for these clades, although the reasons why these genes have been lost while being so well conserved in other lineages remain unknown.

In general, however, the TGF β ligand signalling cassette is well conserved in the Lophotrochozoa, which will allow us to compare functional roles and expres-

sion of these elements between this superphyla and the Deuterostomia and the Ecdysozoa - a practice that has not always been possible when only these latter superphyla were described, as often one or other lacked a gene completely. This has and will continue to allow us to infer ancestral roles for a variety of genes, a vital step in understanding how animal life in general, and the TGF β signalling cascade in particular, has evolved.

This is particularly pertinent for my work in the Nodal cascade, which had previously been regarded as a Deuterostome innovation due to the absence of evidence from the Ecdysozoan superphylum. Now, however, I can reliably infer the presence of almost the entire core Nodal pathway as far back as the Urbilaterian (Fig 5.17). While Lefty appears to be a true Deuterostome novelty, the only other component of this pathway with question marks over its presence at this point is the gene *FoxH*, which acts as the mechanism for passing on Nodal signalling to a variety of downstream genes in the Deuterostomia. Some evidence (data not shown) suggests that this gene is present in the Mollusca, and work is ongoing in the Shimeld laboratory to test this hypothesis. This detailed description of the Nodal Pathway, and the wider the TGF β signalling cascade will allow a range of more hypothesis-driven work to begin in earnest, both in my lab and elsewhere, and some of this is already underway.

5.4 Summary

This section presents a wide variety of work that will form the basis for much systematic investigation into the conservation of the Nodal pathway in the Lophotrochozoa for many years to come. In this chapter, the first systematic treatment of the TGF β signalling cascade in the Lophotrochozoa, an investigation of the *Nodal* locus itself and its putative regulatory mechanisms in this clade, and more speculative, but perhaps high-impact, data from the *P. lamarckii Nodal* locus has been shown. While this work has been quite descriptive in nature, this was necessary for a firm underpinning of investigations of this nature going forward, and has revealed a wide range of tantalising threads for future research. This work will

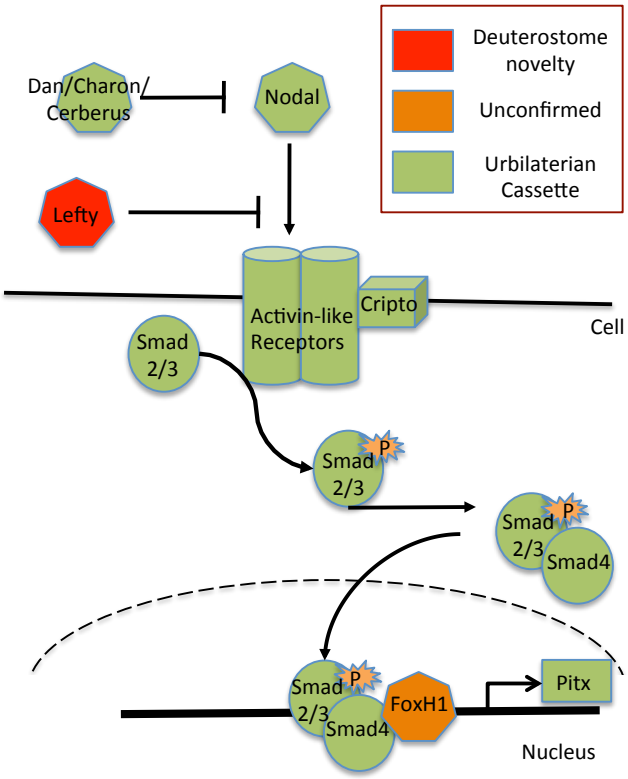


Figure 5.17: *Origin of core Nodal pathway genes across the Bilateria. The vast majority of the core cassette was present as early as the common ancestor of the Deuterostomia and Lophotrochozoa. Question marks remain over the presence of FoxH in the Lophotrochozoa.*

provide the building blocks to allow us to understand a variety of developmental and homeostasis-related signalling cascades in these species, and also provide a valuable resource for tracing the ancestral functionality of the TGF β pathway. It is particularly useful for hypothesis building in my efforts to understand the Nodal pathway within the Lophotrochozoa, as for the first time I have provided a holistic overview of the full complement found in representatives of this clade. This will allow for a range of functional and expression-related experiments to be performed, fully fleshing out how this vital pathway is involved in body plan establishment in organisms in this Superphylum.

Resulting Publications

Some of this work has already been published in peer reviewed journals, while other portions are in various stages of readiness for publication. In the interests of clarity, this is summarised below.

6.0.1 Published Work

Three publications directly relating to my doctoral studies have already been published and are publicly available. These are found in Appendix 7.0.4 of this thesis, and are published as:

- Kenny N, Shimeld S: Additive multiple *k*-mer transcriptome of the keelworm *Pomatoceros lamarckii*; (Annelida; Serpulidae) reveals annelid trochophore transcription factor cassette. Development Genes and Evolution 2012, 222(6):325-339.

Which comprises portions of chapter 3 of this thesis, and

- Kenny N, Quah S, Holland PWH, Tobe SS, Hui JHL: How are comparative genomics and the study of microRNAs changing our views on arthropod endocrinology and adaptations to the environment? General and Comparative Endocrinology 2013, 188:16-22.

Some details of which are found in the introduction and Chapter 3 of this thesis. Finally, the paper

- Namigai E, Kenny N, Shimeld S: Right across the tree of life: the evolution of left right asymmetry in the Bilateria: Genesis (Special Issue), 2014, Manuscript Accepted, In Press.

is available in pre-print online.

Two publications not relevant to my thesis has also came out in the last year:

- Kenny N, Dearden P: NMDA receptor expression and C terminus structure in the rotifer *Brachionus plicatilis* and long-term potentiation across the Metazoa. Invertebrate Neuroscience 2013:13(2) pp 125-134. doi: 10.1007/s10158-013-0154-0

- Wilson M J, Kenny N J and P K Dearden : Components of the dorsal-ventral pathway also contribute to anterior-posterior patterning in honeybee embryos (*Apis mellifera*), *EvoDevo*, 2014, Manuscript Accepted, In Press.

6.0.2 Planned Publications

Several publications are planned or are in the last stages of revision for publication with input from the present work. These are:

- Kenny N, Namigai E, Dearden P, Hui J, Grande C and Shimeld S: The Lophotrochozoan TGF β Signalling Cassette: Diversification and Conservation in a Key Signalling Pathway. The International Journal of Developmental Biology (Special Issue), Manuscript Submitted.

- Kenny N, Duncan E J, Hyink O, Benton M, Morgan S, Leask M, McCartney R, O'Neill M, Blommaert J, Wilson M J and Dearden P K: Insights into Gnathiferan and Lophotrochozoan Biology from Genomic and Transcriptomic Analysis of the Rotifer *Brachionus plicatilis*. Manuscript in prep.

- Grande C, Martin-Duran J M, Kenny N J, Truchado-Garcia M and Hejnol A Evolution, divergence and loss of the Nodal signalling pathway: new data and a synthesis across the Bilateria, *The International Journal of Developmental Biology*, Manuscript Submitted.

- Truchado M, Kenny N and Grande C: Deep Transcriptome of the Schistosomiasis Vector *Biomphalaria glabrata* (Gastropoda: Planorbidae) Illuminates Molluscan Disease-Related Pathways. BMC Genomics, Manuscript in prep.

- Kenny N, Namigai E, Shimeld S and Hui J: Comparison of microRNAs in the Lophotrochozoa facilitated by next-generation sequencing of novel polychaete and limpet genomes. Work in progress.

Summary and Future Directions

7.0.3 Synopsis of Research

This thesis represents a major substantive contribution towards our knowledge of lophotrochozoan biology in several regards. The Lophotrochozoa are still genomically under-sampled, and even resources with poor contiguity such as the ones represented here are invaluable for the mapping the presence of traits and genes across the tree of life. The utility of these resources for our own investigations has been covered in depth in this thesis, and will not be repeated here, but the power of these next-generation sequencing techniques has been an education in itself, and similar approaches are likely to be of repeated importance in my ongoing scientific career. This work is largely complete and is either published or being readied for publication, and will allow a range of investigations by biologists in many fields.

The work presented in Chapter 4 of this thesis represents the most speculative and high-impact part of this thesis, and unfortunately the most interesting results were also the most difficult to interpret. I was however able to identify and describe the expression of a number of key genes in this process, work necessary if *P. vulgata* is to be used for future investigations in this field. This work will form the basis for continuing work in the Shimeld laboratory, and with the addition of corroborating evidence from other species and more focussed investigation, it is hoped that we can gain real insight into the fundamental processes underlying the initial break of symmetry - whether through ion channels, or through any other mechanism.

The TGF β pathway description provided in Chapter 5 of this work is as fundamental to understanding many aspects of the establishment of L/R asymmetry as the sequences presented in Chapter 3. Without a broad overview of the diversity of

influences on TGF β signalling in the Lophotrochozoa it is impossible to interpret the results of functional work in the Nodal pathway, or in any other TGF β mediated cascade. Our description and cataloguing should provide a firm basis for this work to proceed on in the future.

7.0.4 Promising Avenues for Further Investigation

The establishment of L/R asymmetry is one of the most fascinating topics in modern developmental biology, and there are many avenues in which research could proceed in this field. This thesis has shone some light on aspects of this process, but, in many cases, more questions have been raised by the work presented here than have been answered.

The genomic and transcriptomic data presented here no doubt contain information useful in a range of biological investigations, but in particular provide a resource that may allow investigation of regulation of *Nodal* in the Lophotrochozoa. Our analysis has already pinpointed areas of interest in this regard, and with the increasing availability of molluscan genomic data the identification of GREs in this clade will increase in efficiency. It is hoped that mechanisms for reliably inducing temporary and permanent transgenesis are also soon established in this clade, as the robust tests of hypotheses regarding the regulation of Nodal (and other genes) will then become a reality.

The ubiquity of the TGF β pathway in metazoan growth and development means that our description of this cascade was as much a hypothesis-building exercise as a concrete scientific output in its own regard. The many specific questions posed by this part of this thesis have been noted in Chapter 5, but in my opinion, perhaps the most pressing question raised by this analysis regards the role of *Cripto* in the regulation of Nodal signalling in the Lophotrochozoa. The diversity of the ligand cassette of annelids in particular and lophotrochozoans in general also is of wider interest, but will take more work to discern, as will the interplay between ligands and regulators in the Spiralia. This work will shed much light on the ancestral roles of this pathway in the Bilateria, and how it has altered over evolutionary time, but is not as germane to my work as the specific role of *Cripto*.

Most tantalising of all, but unfinished in this work, are the results shown in the final portion of Chapter 4 of this thesis. The role of ion channels in establishing asymmetry is presently contentious, although a role for communicating information about asymmetry, if not in breaking symmetry itself, seems ancestrally shared in the Bilateria. Evidence from the Lophotrochozoa would provide new insight into the mechanism behind this role, and deserves more attention than provided here. While weather conditions stymied much of the work in this regard in past years, it is hoped that ongoing efforts in the Shimeld lab and elsewhere will illuminate this field further. To cut a long thesis short, I am confident that this work will not only stand on its own scientific merits, but will also form a solid building block for ongoing investigations by myself, my colleagues and collaborators, and by scientists of a range of backgrounds worldwide.

Acknowledgements

I would first like to thank my supervisor, Dr Sebastian Shimeld . Your sage advice, guidance and encouragement when I have needed it throughout the years have been much appreciated, and I apologise for all the times I have probably bulldozed my way through without realising my mistakes.

Secondly, I would like to thank all the members of the wider Evolution and Development Research Group. Your friendship, support and straight answers to stupid questions were instrumental in crafting this work into its present form, such as it is, and I hope that we remain both scientific colleagues and friends for the remainders of our lives to come. Everyone (especially Sarah Morgan and Mary Colasanto) who either volunteered, was bribed or coerced into proofing my drafts I sincerely thank you. And apologise. Sorry.

Financial support for this work has been provided by a number of sponsors over the years. The major support for my studies was provided by a Clarendon Scholarship, without which I would not have been able to be here at all. The Elizabeth Hannah Jenkinson Fund provided the funding for sequencing which formed the basis of my investigations. Supplementary support has at various times been provided by the BSDB, the Genetics Society, the Santander Academic Travel Award fund and the Marine Biological Association of the UK, and to them go all my thanks.

Thanks must go to my family - Mum and Dad, Lauren, and the wider whanau - your support and love has been a constant presence throughout in my life, and I couldn't have done half of this without your encouragement along the way.

Finally, Laurie, this work is dedicated to you - I couldn't have done it at all without you.

Bibliography

Adachi, H., Saijoh, Y., Mochida, K., Ohishi, S., Hashiguchi, H., Hirao, A., & Hamada, H. (1999). Determination of left/right asymmetric expression of Nodal by a left side-specific enhancer with sequence similarity to a Lefty-2 enhancer. *Genes and Development*, *13*(12), 1589–1600.

Adam, G., Perrimon, N., & Noselli, S. (2003). The retinoic-like juvenile hormone controls the looping of left-right asymmetric organs in *Drosophila*. *Development*, *130*(11), 2397–2406.

Adams, D., Robinson, K., Fukumoto, T., Yuan, S., Albertson, R., Yelick, P., Kuo, L., McSweeney, M., & Levin, M. (2006). Early, H^+-V-ATPase-dependent proton flux is necessary for consistent left-right patterning of non-mammalian vertebrates. *Development*, *133*, 1657 – 71.

Adamska, M., Degnan, S. M., Green, K. M., Adamski, M., Craigie, A., Larroux, C., & Degnan, B. M. (2007). Wnt and TGF β expression in the sponge *Amphimedon queenslandica* and the origin of metazoan embryonic patterning. *PLoS ONE*, *2*(10), e1031.

Adema, C., Luo, M., Hanelt, B., Hertel, L., Marshall, J., Zhang, S., DeJong, R., Kim, H., Kudrna, D., & Wing, R. (2006). A bacterial artificial chromosome library for *Biomphalaria glabrata*, intermediate snail host of *Schistosoma mansoni*. *Memorias do Instituto Oswaldo Cruz*, *101*, 167 – 177.

Afzelius, B. (1976). A human syndrome caused by immotile cilia. *Science*, *193*(4250), 317–319.

Aguinaldo, A. M., Turbeville, J. M., Linford, L. S., Rivera, M. C., Garey, J. R., Raff, R. A., & Lake, J. A. (1997). Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature*, *387*, 489–493.

Ahringer, J. (2003). Control of cell polarity and mitotic spindle positioning in animal cells. *Current Opinion in Cell Biology*, *15*(1), 73 – 81.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, *25*, 3389 – 3402.

Anderson, D. T. (1973). *Embryology and Phylogeny in Annelids and Arthropods*. Oxford, UK: Pergamon Press.

Andrews, S. (2011). FastQC. Available from `http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/`.

Ansorge, W. (2009). Next-generation DNA sequencing techniques. *New Biotechnology*, *25*, 195 – 203.

Antic, D., Stubbs, J. L., Suyama, K., Kintner, C., Scott, M. P., & Axelrod, J. D. (2010). Planar cell polarity enables posterior localization of Nodal cilia and left-right axis determination during mouse and *Xenopus* embryogenesis. *PLoS ONE*, *5*(2), e8999.

Arce, L., Yokoyama, N. N., & Waterman, M. L. (2006). Diversity of lef/tcf action in development and disease. *Oncogene*, *25*(57), 7492–7504.

Arendt, D., Denes, A. S., Jekely, G., & Tessmar-Raible, K. (2008). The evolution of nervous system centralization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1496), 1523–1528.

Armakolas, A., & Klar, A. J. (2007). Left-right dynein motor implicated in selective chromatid segregation in mouse cells. *Science*, *315*(5808), 100–1.

Armakolas, A., Koutsilieris, M., & Klar, A. J. (2010). Discovery of the mitotic selective chromatid segregation phenomenon and its implications for vertebrate development. *Current Opinions in Cell Biology*, *22*(1), 81–7.

Arnolds, W., van den Biggelaar, J., & Verdonk, N. (1983). Spatial aspects of cell interactions involved in the determination of dorsoventral polarity in equally cleaving gastropods and regulative abilities of their embryos, as studied by micromere deletions in *Lymnaea* and *Patella*. *Developmental Biology*, *192*, 75–85.

Aw, S., Adams, D. S., Qiu, D., & Levin, M. (2008). H,K-ATPase protein localization and Kir4.1 function reveal concordance of three axes during early determination of left-right asymmetry. *Mechanisms of Development*, *125*(3-4), 353–72.

Aw, S., & Levin, M. (2009). Is left-right asymmetry a form of planar cell polarity? *Development*, *136*(3), 355–366.

Babu, D., & Roy, S. (2013). Left-right asymmetry: cilia stir up new surprises in the node. *Open Biology*, *3*(5).

Balemans, W., & Van Hul, W. (2002). Extracellular regulation of BMP signalling in vertebrates: a cocktail of modulators. *Developmental Biology*, *250*(2), 231–50.

Ballantine, W. (1965). *The population dynamics of* Patella vulgata *and other limpets*. Ph.D Thesis, Queen Mary College, University of London.

Barr, J., M. (1973). The teratogenicity of cadmium chloride in two stocks of Wistar rats. *Teratology*, *7*(3), 237–42.

Barroso del Jesus, A., Lucena-Aguilar, G., Sanchez, L., Ligero, G., Gutierrez-Aranda, I., & Menendez, P. (2011). The Nodal inhibitor Lefty is negatively modulated by the microRNA miR-302 in human embryonic stem cells. *The FASEB Journal*, *25*(5), 1497–1508.

Basu, B., & Brueckner, M. (2008). Cilia: multifunctional organelles at the center of vertebrate left-right asymmetry. *Current Topics in Developmental Biology*, *85*, 151–74.

Beck, S., Le Good, J. A., Guzman, M., Ben Haim, N., Roy, K., Beermann, F., & Constam, D. B. (2002). Extraembryonic proteases regulate Nodal signalling during gastrulation. *Nature Cell Biology*, *4*(12), 981–5.

Beisson, J., & Jerka-Dziadosz, M. (1999). Polarities of the centriolar structure: morphogenetic consequences. *Biology of the Cell*, *91*(4-5), 367–378.

Ben-Haim, N., Lu, C., Guzman-Ayala, M., Pescatore, L., Mesnard, D., Bischofberger, M., Naef, F., Robertson, E. J., & Constam, D. B. (2006). The Nodal precursor acting via Activin receptors induces mesoderm by maintaining a source of its convertases and BMP4. *Developmental Cell*, *11*(3), 313–23.

Bentley, D. R. (2006). Whole-genome re-sequencing. *Current Opinion in Genetics and Development*, *16*(6), 545 – 552.

Bergmann, D. C., Lee, M., Robertson, B., Tsou, M. F., Rose, L. S., & Wood, W. B. (2003). Embryonic handedness choice in *C. elegans* involves the gα protein GPA-16. *Development*, *130*(23), 5731–40.

Berriman, M., Haas, B. J., LoVerde, P. T., Wilson, R. A., Dillon, G. P., Cerqueira, G. C., Mashiyama, S. T., Al-Lazikani, B., Andrade, L. F., Ashton, P. D., & Aslett, M. A. (2009). The genome of the blood fluke *Schistosoma mansoni*. *Nature*, *460*(7253), 352–358.

Bessodes, N., Haillot, E., Duboc, V., Rottinger, E., Lahaye, F., & Lepage, T. (2012). Reciprocal signaling between the ectoderm and a mesendodermal left-right organizer directs left-right determination in the sea urchin embryo. *PLoS Genetics*, *8*(12), e1003121.

Beyer, T., Danilchik, M., Thumberger, T., Vick, P., Tisler, M., Schneider, I., Bogusch, S., Andre, P., Ulmer, B., Walentek, P., Niesler, B., Blum, M., & Schweickert, A. (2012). Serotonin signaling is required for Wnt-dependent GRP specification and leftward flow in *Xenopus*. *Current Biology*, *22*(1), 33–39.

Birol, I., Jackman, S. D., Nielsen, C. B., Qian, J. Q., Varhol, R., Stazyk, G., Morin, R. D., Zhao, Y., Hirst, M., Schein, J. E., Horsman, D. E., Connors, J. M., Gascoyne, R. D., Marra, M. A., & Jones, S. J. (2009). *De novo* transcriptome assembly with ABySS. *Bioinformatics*, *25*(21), 2872–7.

Boorman, C. J., & Shimeld, S. M. (2002). *Pitx* homeobox genes in *Ciona* and amphioxus show left-right asymmetry is a conserved chordate character and define the ascidian adenohypophysis. *Evolution and Development*, *4*(5), 354–65.

Boring, L. (1989). Cell-cell interactions determine the dorsoventral axis in embryos of an equally cleaving opisthobranch mollusc. *Developmental Biology*, *136*, 239–53.

Borovina, A., Superina, S., Voskas, D., & Ciruna, B. (2010). Vangl2 directs the posterior tilting and asymmetric localization of motile primary cilia. *Nature Cell Biology*, *12*(4), 407–12.

Bowles, J., Schepers, G., & Koopman, P. (2000). Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators. *Developmental Biology*, *227*(2), 239–255.

Bowman, R., & Lewis, J. (1977). Annual fluctuations in the recruitment of *Patella vulgata* l. *Journal of the Marine Biological Association, UK.*, *57*, 793–815.

Boycott, A. E., Diver, C., Garstang, S. L., & Turner, F. M. (1930). The inheritance of sinistrality in *Limnaea peregra* (Mollusca: Pulmonata). *Philosophical Transactions of the Royal Society B: Biological Sciences*, *219*, 51–131.

Boyle, M. J., & Seaver, E. C. (2008). Developmental expression of *FoxA* and *Gata* genes during gut formation in the polychaete Annelid, *Capitella* sp. i. *Evolution and Development*, *10*(1), 89–105.

Branch, G. (1981). The biology of limpets: physical factors, energy flow, and ecological interactions. *Oceanographic Marine Biology Annual Review*, *19*(19), 235–380.

Brown, N., & Wolpert, L. (1990). The development of handedness in left/right asymmetry. *Development*, *109*(1), 1–9.

Brusca, R., & Brusca, G. (2002). *Invertebrates*. Sunderland, Mass.: Sinauer Associates, 2nd ed.

Burdine, R. D., & Schier, A. F. (2000). Conserved and divergent mechanisms in left/right axis formation. *Genes and Development*, *14*(7), 763–776.

Burgess, S. A., Walker, M. L., Sakakibara, H., Knight, P. J., & Oiwa, K. (2003). Dynein structure and power stroke. *Nature*, *421*(6924), 715–718.

Burglin, T. R., & Cassata, G. (2002). Loss and gain of domains during evolution of cut superclass homeobox genes. *The International Journal of Developmental Biology*, *46*(1), 115–23.

Camey, T., & Verdonk, N. (1969). The early development of the snail *Biomphalaria glabrata* (Say) and the origin of the head organs. *Netherlands Journal of Zoology*, *20*(1), 93–121.

Carlsson, P., & Mahlapuu, M. (2002). Forkhead transcription factors: Key players in development and metabolism. *Developmental Biology*, *250*(1), 1–23.

Cartwright, J. H. E., Piro, O., & Tuval, I. (2004). Fluid-dynamical basis of the embryonic development of left-right asymmetry in vertebrates. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(19), 7234–7239.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, *17*(4), 540–552.

Chen, C., & Shen, M. M. (2004). Two modes by which Lefty proteins inhibit Nodal signalling. *Current Biology*, *14*(7), 618–24.

Cheng, S. K., Olale, F., Bennett, J. T., Brivanlou, A. H., & Schier, A. F. (2003). EGF-CFC proteins are essential coreceptors for the TGF β signals Vg1 and GDF1. *Genes and Development*, *17*(1), 31–6.

Chevreux, B., Wetter, T., & Suhai, S. (1999). Genome sequence assembly using trace signals and additional sequence information. *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB)*, *99*, 45–56.

Choi, J.-H., Kijimoto, T., Snell-Rood, E., Tae, H., Yang, Y., Moczek, A., & Andrews, J. (2010). Gene discovery in the horned beetle *Onthophagus taurus*. *BMC Genomics*, *11*(1), 703.

Choi, W.-Y., Giraldez, A. J., & Schier, A. F. (2007). Target protectors reveal dampening and balancing of Nodal agonist and antagonist by miR-430. *Science*, *318*(5848), 271–274.

Clement, A. (1962). Development of *Ilyanassa* following the removal of the D macromere at successive cleavage stages. *Journal of Experimental Zoology*, *149*, 193–216.

Clements, M., Pernaute, B., Vella, F., & Rodriguez, T. (2011). Crosstalk between Nodal/Activin and MAPK p38 signaling is essential for anterior-posterior axis specification. *Current Biology*, *21*(15), 1289–1295.

Cohen Jr., M. M. (2010). Hedgehog signaling update. *American Journal of Medical Genetics Part A*, *152A*(8), 1875–1914.

Cohen Jr., M. M. (2012). Perspectives on asymmetry: The Erickson Lecture. *American Journal of Medical Genetics Part A*, (pp. 1–5).

Collignon, J., Varlet, I., & Robertson, E. (1996). Relationship between asymmetric Nodal expression and the direction of embryonic turning. *Nature*, *381*, 155 – 8.

Conesa, A., Gotz, S., Garcia-Gomez, J., Terol, J., Talon, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, *21*(18), 3674–3676.

Conlon, F. L., Lyons, K. M., Takaesu, N., Barth, K. S., Kispert, A., Herrmann, B., & Robertson, E. J. (1994). A primary requirement for Nodal in the formation and maintenance of the primitive streak in the mouse. *Development*, *120*(7), 1919–28.

Costello, D. P. (1945). Experimental studies of germinal localization in *Nereis*. i. the development of isolated blastomeres. *Journal of Experimental Zoology*, *100*(1), 19.

Dahal, G. R., Rawson, J., Gassaway, B., Kwok, B., Tong, Y., Ptacek, L. J., & Bates, E. (2012). An inwardly rectifying k^+ channel is required for patterning. *Development*, *139*(19), 3653–3664.

Damen, P., & Dictus, W. J. (1994). Cell lineage of the prototroch of *Patella vulgata* (Gastropoda, Mollusca). *Developmental Biology*, *162*(2), 364–83.

Damen, P., & Dictus, W. J. (1996). Organiser role of the stem cell of the mesoderm in prototroch patterning in *Patella vulgata* (Mollusca, Gastropoda). *Mechanisms of Development*, *56*(1-2), 41–60.

Damen, W., van Grunsven, L., & van Loon, A. (1994). Transcriptional regulation of *tubulin* gene expression in differentiating trochoblasts during early development of *Patella vulgata*. *Development*, *120*(10), 2835–2845.

Danilchik, M. V., Brown, E. E., & Riegert, K. (2006). Intrinsic chiral properties of the *Xenopus* egg cortex: an early indicator of left-right asymmetry? *Development*, *133*(22), 4517–26.

Davies, S. J., Shoemaker, C. B., & Pearce, E. J. (1998). A divergent member of the transforming growth factor β receptor family from *Schistosoma mansoni* is expressed on the parasite surface membrane. *Journal of Biological Chemistry*, *273*(18), 11234–11240.

de Rosa, R., Grenier, J. K., Andreeva, T., Cook, C. E., Adoutte, A., Akam, M., Carroll, S. B., & Balavoine, G. (1999). Hox genes in Brachiopods and Priapulids and protostome evolution. *Nature*, *399*, 772–776.

Dearden, P. K., Donly, C., & Grbic, M. (2002). Expression of pair-rule gene homologues in a chelicerate: early patterning of the two-spotted spider mite *Tetranychus urticae*. *Development*, *129*(23), 5461–72.

Deleury, E., Dubreuil, G., Elangovan, N., Wajnberg, E., Reichhart, J.-M., Gourbal, B., Duval, D., Baron, O. L., Gouzy, J., & Coustau, C. (2012). Specific versus non-specific immune responses in an invertebrate species evidenced by a comparative *de novo* sequencing study. *PLoS ONE*, *7*(3), e32512.

Der, J., Barker, M., Wickett, N., de Pamphilis, C., & Wolf, P. (2011). *De novo* characterization of the gametophyte transcriptome in bracken fern, *Pteridium aquilinum*. *BMC Genomics*, *12*(1), 99.

Derynck, R., & Miyazono, K. (2008). TGF β and the TGF β family. In D. R, & M. K (Eds.) *The TGF β family*, (pp. 29–43). New York: Cold Spring Harbor Laboratory Press.

Dettmer, U., Kuhn, P.-H., Abou-Ajram, C., Lichtenthaler, S. F., Kruger, M., Kremmer, E., Haass, C., & Haffner, C. (2010). Transmembrane protein 147 (TMEM147) is a novel component of the Nicalin-NOMO protein complex. *Journal of Biological Chemistry*, *285*(34), 26174–26181.

Dhert, P., & Sorgeloos, P. (1994). *Live feeds in aquaculture. In: K.P.P. Nambier and T. Singh, Editors, Aquaculture Towards the 21st Century INFOFISH-Aquatech Conference, Colombo, Sri Lanka*. Kuala Lumpur: INFOFISH.

Di Guglielmo, G. M., Le Roy, C., Goodfellow, A. F., & Wrana, J. L. (2003). Distinct endocytic pathways regulate TGF β receptor signalling and turnover. *Nature Cell Biology*, *5*(5), 410–21.

Dickmeis, T., Aanstad, P., Clark, M., Fischer, N., Herwig, R., Mourrain, P., Blader, P., Rosa, F., Lehrach, H., & Strahle, U. (2001). Identification of Nodal signalling targets by array analysis of induced complex probes. *Developmental Dynamics*, *222*(4), 571–80.

Dictus, W. J. A. G., & Damen, P. (1997). Cell-lineage and clonal-contribution map of the trochophore larva of *Patella vulgata* (Mollusca). *Mechanisms of Development*, *62*(2), 213–226.

Duboc, V., Rottinger, E., Lapraz, F., Besnardeau, L., & Lepage, T. (2005). Left-right asymmetry in the sea urchin embryo is regulated by Nodal signalling on the right side. *Developmental Cell*, *9*(1), 147–58.

Duncan, E. J., Benton, M. A., & Dearden, P. K. (2013). Canonical terminal patterning is an evolutionary novelty. *Developmental Biology*, *377*(1), 245–261.

Dunn, C., Hejnol, A., Matus, D., Pang, K., Browne, W., Smith, S., Seaver, E., Rouse, G., Obst, M., Edgecombe, G., Sorensen, M., Haddock, S., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R., Wheeler, W., Martindale, M., & Giribet, G. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, *452*, 745 – 749.

Ellertsdottir, E., Ganz, J., Darr, K., Loges, N., Biemar, F., Seifert, F., Ettl, A.-K., Kramer-Zucker, A. K., Nitschke, R., & Driever, W. (2006). A mutation in the zebrafish Na,K-ATPase subunit provides genetic evidence that the sodium potassium pump contributes to left-right asymmetry downstream or in parallel to Nodal flow. *Developmental Dynamics*, *235*(7), 1794–1808.

Esser, A., Smith, K., Weaver, J., & Levin, M. (2006). Mathematical model of morphogen electrophoresis through gap junctions. *Developmental Dynamics*, *235*, 2144 – 59.

Essner, J. J., Amack, J. D., Nyholm, M. K., Harris, E. B., & Yost, H. J. (2005). Kupffer's vesicle is a ciliated organ of asymmetry in the zebrafish embryo that initiates left-right development of the brain, heart and gut. *Development*, *132*(6), 1247–1260.

Feasey, N., Wansbrough-Jones, M., Mabey, D. C. W., & Solomon, A. W. (2010). Neglected tropical diseases. *British Medical Bulletin*, *93*(1), 179–200.

Field, K., Olsen, G., Lane, D., Giovannoni, S., Ghiselin, M., Raff, E., Pace, N., & Raff, R. (1988). Molecular phylogeny of the animal kingdom. *Science*, *239*, 748–753.

Field, S., Riley, K.-L., Grimes, D. T., Hilton, H., Simon, M., Powles-Glover, N., Siggers, P., Bogani, D., Greenfield, A., & Norris, D. P. (2011). Pkd1l1 establishes left-right asymmetry and physically interacts with Pkd2. *Development*, *138*(6), 1131–1142.

Flachsova, M., Sindelka, R., & Kubista, M. (2013). Single blastomere expression profiling of *Xenopus laevis* embryos of 8 to 32-cells reveals developmental asymmetry. *Scientific Reports*, *3*, 1–6.

Fleury, E., Fabioux, C., Lelong, C., Favrel, P., & Huvet, A. (2008). Characterization of a gonad-specific transforming growth factor β superfamily member differentially expressed during the reproductive cycle of the oyster *Crassostrea gigas*. *Gene*, *410*(1), 187–96.

Flot, J.-F., Hespeels, B., Li, X., Noel, B., Arkhipova, I., Danchin, E. G. J., Hejnol, A., Henrissat, B., Koszul, R., Aury, J.-M., Barbe, V., Barthelemy, R.-M., Bast, J., Bazykin, G. A., Chabrol, O., Couloux, A., Da Rocha, M., Da Silva, C., Gladyshev, E., Gouret, P., Hallatschek, O., Hecox-Lea, B., Labadie, K., Lejeune, B., Piskurek, O., Poulain, J., Rodriguez, F., Ryan, J. F., Vakhrusheva, O. A., Wajnberg, E., Wirth, B., Yushenova, I., Kellis, M., Kondrashov, A. S., Mark Welch, D. B., Pontarotti, P., Weissenbach, J., Wincker, P., Jaillon, O., & Van Doninck, K. (2013). Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga*. *Nature*, *advance online publication*.

Forrester, S. G., Warfel, P. W., & Pearce, E. J. (2004). Tegumental expression of a novel type ii receptor serine/threonine kinase (smrk2) in *Schistosoma mansoni*. *Molecular and Biochemical Parasitology*, *136*(2), 149–156.

Freeman, G., & Lundelius, J. W. (1982). The developmental genetics of dextrality and sinistrality in the gastropod *Lymnaea peregra*. *Development, Genes and Evolution*, *191*, 69–83.

Freeman, G., & Lundelius, J. W. (1992). Evolutionary implications of the mode of D quadrant specification in coelomates with spiral cleavage. *Journal of Evolutionary Biology*, *5*(2), 205–247.

Freitas, T. C., Jung, E., & Pearce, E. J. (2007). TGF β signalling controls embryo development in the parasitic flatworm *Schistosoma mansoni*. *PLoS Pathogens*, *3*(4), e52.

Fujinaga, M., & Baden, J. M. (1991). Evidence for an adrenergic mechanism in the control of body asymmetry. *Developmental Biology*, *143*(1), 203–5.

Fukumoto, T., Blakely, R., & Levin, M. (2005a). Serotonin transporter function is an early step in left-right patterning in chick and frog embryos. *Developmental Neuroscience*, *27*(6), 349–63.

Fukumoto, T., Kema, I., & Levin, M. (2005b). Serotonin signalling is a very early step in patterning of the left-right axis in chick and frog embryos. *Current Biology*, *15*, 794 – 803.

Gage, P. J., Suh, H., & Camper, S. A. (1999). The Bicoid-related *Pitx* gene family in development. *Mammalian Genome*, *10*, 197–200.

Garcia-Fernandez, J., & Benito-Gutierrez, E. (2009). It's a long way from amphioxus: descendants of the earliest chordate. *BioEssays*, *31*(6), 665–675.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D., & Bairoch, A. (2003). Expasy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research*, *31*(13), 3784–8.

Gavino, M. A., & Reddien, P. W. (2011). A BMP/ADMP regulatory circuit controls maintenance and regeneration of dorsal-ventral polarity in planarians. *Current Biology*, *21*(4), 294–9.

Gerhart, J., Ubbels, G., Black, S., Hara, K., & Kirschner, M. (1981). A reinvestigation of the role of the grey crescent in axis formation in *Xenopus laevis*. *Nature*, *292*(5823), 511–516.

Gilbert, J. (1963). Mictic female production in the rotifer *Brachionus calyciflorus*. *Journal of Experimental Zoology*, *153*, 113–124.

Gilbert, J. (1974). Dormancy in rotifers. *Transactions of the American Microscopical Society*, *93*, 490–513.

Gilbert, S. (2000). *Developmental Biology*. Sunderland (MA): Sinauer Associates, 6th ed.

Giribet, G. (2008). Assembling the lophotrochozoan (spiralian) tree of life. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*, 1513–1522.

Glinka, A., Delius, H., Blumenstock, C., & Niehrs, C. (1996). Combinatorial signalling by Xwnt-11 and Xnr3 in the organizer ephithelium. *Mechanisms of Development*, *60*(2), 221 – 231.

Golikov, A., & Starobogatov, Y. (1975). Systematics of prosobranch gastropods. *Malacologia*, *15*, 185–232.

Gomez, A., Serra, M., Carvalho, G. R., Lunt, D., & Baum, D. (2002). Speciation in ancient cryptic species complexes: Evidence from the molecular phylogeny of *Brachionus plicatilis* (rotifera). *Evolution*, *56*, 1431–1444.

Gotz, S., Arnold, R., Sebastian-Leon, P., Martin-Rodriguez, S., Tischler, P., Jehl, M., Dopazo, J., Rattei, T., & Conesa, A. (2011). B2G-FAR, a species-centered GO annotation repository. *Bioinformatics*, *27*(7), 919–924.

Gotz, S., Garcia-Gomez, J., Terol, J., Williams, T., Nagaraj, S., Nueda, M., Robles, M., Talon, M., Dopazo, J., & Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*, *36*, 3420 – 3435.

Gourronc, F., Ahmad, N., Nedza, N., Eggleston, T., & Rebagliati, M. (2007). Nodal activity around Kupffer's vesicle depends on the T-box transcription factors Notail and Spadetail and on Notch signalling. *Developmental Dynamics*, *236*(8), 2131–46.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., & Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652.

Grande, C. (2010). Left/right asymmetries in Spiralia. *Integrative and Comparative Biology*, *50*(5), 744–755.

Grande, C., & Patel, N. (2010). Lophotrochozoa get into the game: The Nodal pathway and left/right asymmetry in Bilateria. *Cold Spring Harbor Symposia on Quantitative Biology*, *74*, 281–287.

Grande, C., & Patel, N. H. (2009). Nodal signalling is involved in left-right asymmetry in snails. *Nature*, *457*(7232), 1007–11.

Granier, C., Gurchenkov, V., Perea-Gomez, A., Camus, A., Ott, S., Papanayotou, C., Iranzo, J., Moreau, A., Reid, J., Koentges, G., Saberan-Djoneidi, D., & Collignon, J. (2011). Nodal cis-regulatory elements reveal epiblast and primitive endoderm heterogeneity in the peri-implantation mouse embryo. *Developmental Biology*, *349*(2), 350 – 362.

Gray, P. C., & Vale, W. (2012). Cripto/grp78 modulation of the TGF β pathway in development and oncogenesis. *FEBS Letters*, *586*(14), 1836–1845.

Greenaway, J. C., Fantel, A. G., & Juchau, M. R. (1986). On the capacity of nitroheterocyclic compounds to elicit an unusual axial asymmetry in cultured rat embryos. *Toxicology and Applied Pharmacology*, *82*(2), 307–15.

Gros, J., Feistel, K., Viebahn, C., Blum, M., & Tabin, C. (2009). Cell movements at Hensen's node establish left/right asymmetric gene expression in the chick. *Science*, *324*, 941 – 4.

Grunert, S., & Johnston, D. S. (1996). RNA localization and the development of asymmetry during *Drosophila* oogenesis. *Current Opinion in Genetics and Development*, *6*(4), 395 – 402.

Guerrier, P. (1970). Characteristics of segmentation and determination of dorsoventral polarity in the development of Spiralia. 3. *Pholas dactylus* and *Spisula subtruncata*. *Journal of Embryology and Experimental Morphology*, *23*(3), 667–92.

Guzman-Ayala, M., Lee, K. L., Mavrakis, K. J., Goggolidou, P., Norris, D. P., & Episkopou, V. (2009). Graded Smad2/3 activation is converted directly into levels of target gene expression in embryonic stem cells. *PLoS ONE*, *4*(1), e4268.

Haffner, C., Frauli, M., Topp, S., Irmler, M., Hofmann, K., Regula, J. T., Bally-Cuif, L., & Haass, C. (2004). Nicalin and its binding partner Nomo are novel Nodal signaling antagonists. *EMBO Journal*, *23*(15), 3041–3050.

Hagiwara, A., Kotani, T., Snell, T. W., Assava-Aree, M., & Hirayama, K. (1995). Morphology, reproduction, genetics, and mating behavior of small, tropical marine *Brachionus* strains (Rotifera). *Journal of Experimental Marine Biology and Ecology*, *194*, 25 – 37.

Halanych, K. (2004). The new view of animal phylogeny. *Annual Review of Ecology, Evolution, and Systematics*, *35*, 229 – 256.

Halanych, K. M., Bacheller, J. D., Aguinaldo, A. M., Liva, S. M., Hillis, D. M., & Lake, J. A. (1995). Evidence from 18s ribosomal DNA that the lophophorates are protostome animals. *Science*, *267*(5204), 1641–3.

Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, *41*, 95–98.

Hamada, H., Meno, C., Watanabe, D., & Saijoh, Y. (2002). Establishment of vertebrate left-right asymmetry. *Nature Reviews Genetics*, *3*, 103 – 13.

Hamer, J. (2002). *The development and settlement of certain marine tubeworm (Serpulidae and Spirorbidae) larvae in response to biofilms*. Ph.D. thesis, University of Wales Bangor.

Hamer, J., Walker, G., & Latchford, J. (2001). Settlement of *Pomatoceros lamarkii* (Serpulidae) larvae on biofilmed surfaces and the effect of aerial drying. *Journal of Experimental Marine Biology and Ecology*, *260*, 113 – 132.

Hansen, K. B., Furukawa, H., & Traynelis, S. F. (2010). Control of assembly and function of glutamate receptors by the amino-terminal domain. *Molecular Pharmacology*.

Harms, P. W., & Chang, C. (2003). Tomoregulin-1 (TMEFF1) inhibits Nodal signalling through direct binding to the Nodal coreceptor Cripto. *Genes and Development*, *17*(21), 2624–9.

Harnad, S. (1977). *Lateralization in the Nervous System*. New York: Academic Press.

Hashimoto, M., Shinohara, K., Wang, J., Ikeuchi, S., Yoshiba, S., Meno, C., Nonaka, S., Takada, S., Hatta, K., Wynshaw-Boris, A., & Hamada, H. (2010). Planar polarization of node cells determines the rotational axis of node cilia. *Nature Cell Biology*, *12*(2), 170–6.

Haszprunar, G. (1988). On the origin and evolution of major gastropod groups, with special reference to the Streptoneura. *Journal of Molluscan Studies*, *54*(4), 367–441.

Hegedus, Z., Zakrzewska, A., Agoston, V., Ordas, A., Racz, P., Mink, M., Spaink, H., & Meijer, A. (2009). Deep sequencing of the zebrafish transcriptome response to mycobacterium infection. *Molecular Immunology*, *46*, 2918 – 2930.

Hejnol, A. (2010). A twist in time - the evolution of spiral cleavage in the light of animal phylogeny. *Integrative and Comparative Biology*, *50*(5), 695–706.

Heldin, C. H., & Moustakas, A. (2012). Role of Smads in TGF β signalling. *Cell and Tissue Research*, *347*(1), 21–36.

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, *89*(22), 10915–10919.

Henry, J. J., Collin, R., & Perry, K. J. (2010a). The slipper snail, *Crepidula*: An emerging lophotrochozoan model system. *The Biological Bulletin*, *218*(3), 211–229.

Henry, J. J., Perry, K. J., Fukui, L., & Alvi, N. (2010b). Differential localization of mRNAs during early development in the Mollusc, *Crepidula fornicata*. *Integrative and Comparative Biology*, *50*(5), 720–733.

Henry, J. Q., Okusu, A., & Martindale, M. Q. (2004). The cell lineage of the polyplacophoran, *Chaetopleura apiculata*, variation in the spiralian program and implications for molluscan evolution. *Developmental Biology*, *272*(1), 145–160.

Herpin, A., Lelong, C., Becker, T., Rosa, F., Favrel, P., & Cunningham, C. (2005). Structural and functional evidence for a singular repertoire of BMP receptor signal transducing proteins in the lophotrochozoan *Crassostrea gigas* suggests a shared ancestral BMP/activin pathway. *The FEBS journal*, *272*(13), 3424–40.

Herpin, A., Lelong, C., & Favrel, P. (2004). TGF β-related proteins: an ancestral and widespread superfamily of cytokines in metazoans. *Developmental and Comparative Immunology*, *28*(5), 461–85.

Hibino, T., Ishii, Y., Levin, M., & Nishino, A. (2006). Ion flow regulates left/right asymmetry in sea urchin development. *Development, Genes and Evolution*, *216*(5), 265–276.

Hong, S.-K., & Dawid, I. B. (2009). FGF-dependent left/right asymmetry patterning in zebrafish is mediated by Ier2 and Fibp1. *Proceedings of the National Academy of Sciences*, *106*(7), 2230–2235.

Honjo, T. (1996). The shortest path from the surface to the nucleus: RBP-J/Su(H) transcription factor. *Genes to Cells*, *1*(1), 1–9.

Huang, S., Ma, J., Liu, X., Zhang, Y., & Luo, L. (2011). Retinoic acid signalling sequentially controls visceral and heart laterality in zebrafish. *Journal of Biological Chemistry*, *286*(32), 28533–28543.

Huang, X., & Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome Research*, *9*(9), 868 – 877.

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, *17*(8), 754–5.

Huminiecki, L., Goldovsky, L., Freilich, S., Moustakas, A., Ouzounis, C., & Heldin, C. H. (2009). Emergence, development and diversification of the TGF β signalling pathway within the animal kingdom. *BMC Evolutionary Biology*, *9*, 28.

Huse, M., Chen, Y. G., Massagué, J., & Kuriyan, J. (1999). Crystal structure of the cytoplasmic domain of the type i TGF β receptor in complex with FKBP12. *Cell*, *96*(3), 425–36.

Huson, D. H., Richter, D. C., Rausch, C., Dezulian, T., Franz, M., & Rupp, R. (2007). Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics*, *8*, 460.

Inacio, J. M., Marques, S., Nakamura, T., Shinohara, K., Meno, C., Hamada, H., & Belo, J. A. (2013). The dynamic right-to-left translocation of Cerl2 is involved in the regulation and termination of Nodal activity in the mouse node. *PLoS ONE*, *8*(3), e60406.

Itoh, S., & ten Dijke, P. (2007). Negative regulation of TGF β receptor/Smad signal transduction. *Current Opinion in Cell Biology*, *19*(2), 176–84.

Jager, M., Queinnec, E., Houliston, E., & Manuel, M. (2006). Expansion of the SOX gene family predated the emergence of the Bilateria. *Molecular Phylogenetics and Evolution*, *39*(2), 468–77.

Jiang, S., Meadows, J., Anderson, S. A., & Mukkada, A. J. (2002). Antileishmanial activity of the antiulcer agent omeprazole. *Antimicrobial Agents Chemotherapy*, *46*(8), 2569–74.

Johnston, R. J., & Hobert, O. (2003). A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. *Nature*, *426*(6968), 845–849.

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer applications in the biosciences : CABIOS*, *8*(3), 275–282.

Kaestner, K. H., Knoechel, W., & Martinez, D. E. (2000). Unified nomenclature for the winged helix/forkhead transcription factors. *Genes and Development*, *14*(2), 142–146.

Kato, Y. (2011). The multiple roles of Notch signalling during left-right patterning. *Cellular and Molecular Life Sciences*, (pp. 1–13).

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066.

Katoh, K., & Standley, D. M. (2013). Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780.

Katoh, K., & Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics*, *9*, 286 – 298.

Kawakami, Y., Raya, A., Raya, R. M., Rodriguez-Esteban, C., & Belmonte, J. C. (2005). Retinoic acid signalling links left-right asymmetric patterning and bilaterally symmetric somitogenesis in the zebrafish embryo. *Nature*, *435*(7039), 165–71.

Keane, T., Creevey, C., Pentony, M., Naughton, T., & Mclnerney, J. (2006). Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evolutionary Biology*, *6*(1), 29.

Kennedy, M. P., Omran, H., Leigh, M. W., Dell, S., Morgan, L., Molina, P. L., Robinson, B. V., Minnix, S. L., Olbrich, H., Severin, T., Ahrens, P., Lange, L., Morillas, H. N., Noone, P. G., Zariwala, M. A., & Knowles, M. R. (2007). Congenital heart disease and other heterotaxic defects in a large cohort of patients with primary ciliary dyskinesia. *Circulation*, *115*(22), 2814–2821.

Kenny, N., & Shimeld, S. (2012). Additive multiple *k*-mer transcriptome of the keelworm *Pomatoceros lamarckii*; (Annelida; Serpulidae) reveals annelid trochophore transcription factor cassette. *Development Genes and Evolution*, *222*, 325–339.

Kenny, N. J., Quah, S., Holland, P. W. H., Tobe, S. S., & Hui, J. H. L. (2013). How are comparative genomics and the study of microrna changing our views on arthropod endocrinology and adaptations to the environment? *General and Comparative Endocrinology*, *188*(0), 16–22.

Kerner, P., Simionato, E., Le Gouar, M., & Vervoort, M. (2009). Orthologs of key vertebrate neural genes are expressed during neurogenesis in the Annelid *Platynereis dumerilii*. *Evolution and Development*, *11*(5), 513–524.

Kin, K., Kakoi, S., & Wada, H. (2009). A novel role for dpp in the shaping of bivalve shells revealed in a conserved molluscan developmental program. *Developmental Biology*, *329*(1), 152–66.

King, T., & Brown, N. A. (1999). Embryonic asymmetry: The left side gets all the best genes. *Current Biology*, *9*(1), R18 – R22.

Kingsley, E. P., Chan, X. Y., Duan, Y., & Lambert, J. D. (2007). Widespread RNA segregation in a spiralian embryo. *Evolution and Development*, *9*(6), 527–539.

Klar, A. J. (1994). A model for specification of the left-right axis in vertebrates. *Trends in Genetics*, *10*(11), 392–6.

Klar, A. J. (2008). Support for the selective chromatid segregation hypothesis advanced for the mechanism of left-right body axis development in mice. *Breast Disease*, *29*, 47–56.

Kocot, K. M., Cannon, J. T., Todt, C., Citarella, M. R., Kohn, A. B., Meyer, A., Santos, S. R., Schander, C., Moroz, L. L., Lieb, B., & Halanych, K. M. (2011). Phylogenomics reveals deep molluscan relationships. *Nature*, *477*, 452–456.

Kofron, M., Puck, H., Standley, H., Wylie, C., Old, R., Whitman, M., & Heasman, J. (2004). New roles for FoxH1 in patterning the early embryo. *Development*, *131*(20), 5065–5078.

Konno, A., Kaizu, M., Hotta, K., Horie, T., Sasakura, Y., Ikeo, K., & Inaba, K. (2010). Distribution and structural diversity of cilia in tadpole larvae of the ascidian *Ciona intestinalis*. *Developmental Biology*, *337*(1), 42 – 62.

Koopman, P., Schepers, G., Brenner, S., & Venkatesh, B. (2004). Origin and diversity of the SOX transcription factor gene family: genome-wide analysis in *Fugu rubripes*. *Gene*, *328*, 177–186.

Kozomara, A., & Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*, *39*, D152–D157.

Krebs, L. T., Iwai, N., Nonaka, S., Welsh, I. C., Lan, Y., Jiang, R., Saijoh, Y., O'Brien, T. P., Hamada, H., & Gridley, T. (2003). Notch signalling regulates left-right asymmetry determination by inducing Nodal expression. *Genes and Development*, *17*(10), 1207–12.

Kuhtreiber, W., Til, E., & Dongen, C. (1988). Monensin interferes with the determination of the mesodermal cell line in embryos of *Patella vulgata*. *Roux's archives of developmental biology*, *197*(1), 10–18.

Kumar, S., & Blaxter, M. (2010). Comparing *de novo* assemblers for 454 transcriptome data. *BMC Genomics*, *11*(1), 571.

Kuo, D.-H., & Weisblat, D. (2011). A new molecular logic for BMP-mediated dorsoventral patterning in the leech *Helobdella*. *Current Biology*, *21*(15), 1282–1288.

Kurita, Y., & Wada, H. (2011). Evidence that gastropod torsion is driven by asymmetric cell proliferation activated by TGFβ signalling. *Biology Letters*, *7*(5), 759–762.

Kuroda, R., Endo, B., Abe, M., & Shimizu, M. (2009). Chiral blastomere arrangement dictates zygotic left/right asymmetry pathway in snails. *Nature*, *462*(7274), 790–794.

Kurtz, S. (2011). Vmatch large scale sequence analysis software.
URL `http://www.vmatch.de/`

Lambert, J., & Nagy, L. (2001). MAPK signalling by the D quadrant embryonic organizer of the mollusc *Ilyanassa obsoleta*. *Development*, *128*(1), 45–56.

Lambert, J., & Nagy, L. (2003). The MAPK cascade in equally cleaving spiralian embryos. *Developmental Biology*, *263*(2), 231 – 241.

Lambrecht, N., Corbett, Z., Bayle, D., Karlish, S. J. D., & Sachs, G. (1998). Identification of the site of inhibition by omeprazole of a fusion protein of the H,K-ATPase using site-directed mutagenesis. *Journal of Biological Chemistry*, *273*(22), 13719–13728.

Lamonerie, T., Tremblay, J., Lanctot, C., Therrien, M., Gauthier, Y., & Drouin, J. (1996). *Ptx1*, a Bicoid-related homeo box transcription factor involved in transcription of the pro-opiomelanocortin gene. *Genes and Development*, *10*, 1284–1295.

Lane, M. C., & Sheets, M. D. (2006). Heading in a new direction: Implications of the revised fate map for understanding *Xenopus laevis* development. *Developmental Biology*, *296*(1), 12 – 28.

Langenbacher, A., & Chen, J. N. (2008). Calcium signalling: A common thread in vertebrate left/right axis development. *Developmental Dynamics*, *237*(12), 3491–3496.

Lapraz, F., Rottinger, E., Duboc, V., Range, R., Duloquin, L., Walton, K., Wu, S.-Y., Bradham, C., Loza, M. A., Hibino, T., Wilson, K., Poustka, A., McClay, D., Angerer, L., Gache, C., & Lepage, T. (2006). RTK and TGF β signaling pathways genes in the sea urchin genome. *Developmental Biology*, *300*(1), 132–152.

Lartillot, N., Le Gouar, M., & Adoutte, A. (2002a). Expression patterns of fork head and goosecoid homologues in the mollusc *Patella vulgata* supports the ancestry of the anterior mesendoderm across Bilateria. *Development, Genes and Evolution*, *212*(11), 551–561.

Lartillot, N., Lespinet, O., Vervoort, M., & Adoutte, A. (2002b). Expression pattern of brachyury in the mollusc *Patella vulgata* suggests a conserved role in the establishments of the ap axis in bilateria. *Development*, *129*(6), 1411–1421.

Le Good, J. A., Joubin, K., Giraldez, A. J., Ben-Haim, N., Beck, S., Chen, Y., Schier, A. F., & Constam, D. B. (2005). Nodal stability determines signalling range. *Current Biology*, *15*(1), 31–6.

Lesch, B. J., & Bargmann, C. I. (2010). The homeodomain protein *HMBX-1* maintains asymmetric gene expression in adult *C. elegans* olfactory neurons. *Genes and Development*, *24*(16), 1802–1815.

Levin, M. (2005). Left-right asymmetry in embryonic development: a comprehensive review. *Mechanisms of Development*, *122*(1), 3 – 25.

Levin, M., Johnson, R. L., Stern, C. D., Kuehn, M., & Tabin, C. (1995). A molecular pathway determining left-right asymmetry in chick embryogenesis. *Cell*, *82*(5), 803–14.

Levin, M., & Mercola, M. (1998). Gap junctions are involved in the early generation of left/right asymmetry. *Developmental Biology*, *203*(1), 90 – 105.

Levin, M., & Mercola, M. (1999). Gap junction-mediated transfer of left-right patterning signals in the early chick blastoderm is upstream of Shh asymmetry in the node. *Development*, *126*(21), 4703–14.

Levin, M., & Palmer, A. R. (2007). Left/right patterning from the inside out: Widespread evidence for intracellular control. *BioEssays*, *29*(3), 271–287.

Levin, M., Thorlin, T., Robinson, K. R., Nogi, T., & Mercola, M. (2002). Asymmetries in H+/K+-ATPase and cell membrane potentials comprise a very early step in left-right patterning. *Cell*, *111*(1), 77–89.

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., et al. (2010a). The sequence and *de novo* assembly of the giant panda genome. *Nature*, *463*(7279), 311–317.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., & Kristiansen, K. (2010b). *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Research*, *20*, 265 – 272.

Lindberg, D. (2008). Patellogastropoda, Neritimorpha, and Cocculinoidea. In W. Ponder, & D. Lindberg (Eds.) *Phylogeny and evolution of the Mollusca*, (pp. 271–296). California: University of California Press.

Liu, M., Davey, J. W., Banerjee, R., Han, J., Yang, F., Aboobaker, A., Blaxter, M. L., & Davison, A. (2013a). Fine mapping of the pond snail left-right asymmetry (chirality) locus using RAD-Seq and Fibre-FISH. *PLoS ONE*, *8*(8), e71067.

Liu, X., Sun, Y., Weinberg, R. A., & Lodish, H. F. (2001). Ski/Sno and TGF β signalling. *Cytokine and Growth Factor Reviews*, *12*(1), 1–8.

Liu, Y., Schroder, J., & Schmidt, B. (2013b). Musket: a multistage *k*-mer spectrum-based error corrector for Illumina sequence data. *Bioinformatics*, *29*(3), 308–315.

Lockyer, A., Spinks, J., Kane, R., Hoffmann, K., Fitzpatrick, J., Rollinson, D., Noble, L., & Jones, C. (2008). *Biomphalaria glabrata* transcriptome: cDNA microarray profiling identifies resistant- and susceptible-specific gene expression in haemocytes from snail strains exposed to *Schistosoma mansoni*. *BMC Genomics*, *9*(1), 634.

Lockyer, A. E., Spinks, J. N., Walker, A. J., Kane, R. A., Noble, L. R., Rollinson, D., Dias-Neto, E., & Jones, C. S. (2007). *Biomphalaria glabrata* transcriptome: Identification of cell-signalling, transcriptional control and immune-related genes from open reading frame expressed sequence tags (ORESTES). *Developmental and Comparative Immunology*, *31*(8), 763–782.

Long, S., Ahmad, N., & Rebagliati, M. (2003). The zebrafish Nodal-related gene *Southpaw* is required for visceral and diencephalic left-right asymmetry. *Development*, *130*(11), 2303–2316.

Lowe, L., Supp, D., Sampath, K., Yokoyama, T., Wright, C., Potter, S., Overbeek, P., & Kuehn, M. (1996). Conserved left-right asymmetry of Nodal expression and alterations in murine *situs inversus*. *Nature*, *381*, 158 – 61.

Lubzens, L. (1987). Raising rotifers for use in aquaculture. *Hydrobiologia*, *147*, 245–255.

Luo, Y.-J., & Su, Y.-H. (2012). Opposing Nodal and BMP signals regulate left-right asymmetry in the sea urchin larva. *PLoS Biol*, *10*(10), e1001402.

Mark Welch, D. B. (2001). Early contributions of molecular phylogenetics to understanding the evolution of rotifera. *Hydrobiologia*, *446*, 315–322.

Mark Welch, D. B., & Meselson., M. (1998). Measurements of the genome size of the monogonont rotifer *Brachionus plicatilis* and of the bdelloid rotifers *Philodina roseola* and *Habrotrocha constricta*. *Hydrobiologia*, *387*, 395–402.

Marszalek, J. R., Ruiz-Lozano, P., Roberts, E., Chien, K. R., & Goldstein, L. S. (1999). *Situs inversus* and embryonic ciliary morphogenesis defects in mouse mutants lacking the KIF3a subunit of kinesin-ii. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(9), 5043–8.

Martello, G., Zacchigna, L., Inui, M., Montagner, M., Adorno, M., Mamidi, A., Morsut, L., Soligo, S., Tran, U., Dupont, S., Cordenonsi, M., Wessely, O., & Piccolo, S. (2007). MicroRNA control of Nodal signalling. *Nature*, *449*(7159), 183–188.

Martin, J. A., & Wang, Z. (2011). Next-generation transcriptome assembly. *Nature Reviews Genetics*, *12*(10), 671–682.

Martindale, M. (1986). The organizing role of the D quadrant in an equal-clearing spiralian, *Lymnea stagnalis* as studied by UV-laser deletion of macromeres at intervals between third and fourth quartet formation. *International Journal of Invertebrate Reproduction and Development*, *9*(2).

Martindale, M. Q. (1985). The role of animal-vegetal interaction with respect to the determination of dorsoventral polarity in the equal-cleaving spiralian, *Lymnaea palustris*. *Development, Genes and Evolution*, *194*(5), 281.

Massagué, J. (1998). TGF-β signal transduction. *Annual Review of Biochemistry*, *67*, 753–91.

Massagué, J., Blain, S. W., & Lo, R. S. (2000). TGF β signalling in growth control, cancer, and heritable disorders. *Cell*, *103*(2), 295–309.

Massagué, J., Seoane, J., & Wotton, D. (2005). Smad transcription factors. *Genes and Development*, *19*(23), 2783–2810.

May-Simera, H. L., Kai, M., Hernandez, V., Osborn, D. P., Tada, M., & Beales, P. L. (2010). Bbs8, together with the planar cell polarity protein Vangl2, is required to establish left-right asymmetry in zebrafish. *Developmental Biology*, *345*(2), 215–25.

McDougall, C., Chen, W.-C., Shimeld, S., & Ferrier, D. (2006). The development of the larval nervous system, musculature and ciliary bands of *Pomatoceros lamarckii* (Annelida): heterochrony in polychaetes. *Frontiers in Zoology*, *3*(1), 16.

McDougall, C., Korchagina, N., Tobin, J., & Ferrier, D. (2011). Annelid Distal-less/Dlx duplications reveal varied post-duplication fates. *BMC Evolutionary Biology*, *11*(1), 241.

McGrath, J., & Brueckner, M. (2003). Cilia are at the heart of vertebrate left-right asymmetry. *Current Opinion in Genetics and Development*, *13*(4), 385–392.

McGrath, J., Somlo, S., Makova, S., Tian, X., & Brueckner, M. (2003). Two populations of node monocilia initiate left-right asymmetry in the mouse. *Cell*, *114*(1), 61–73.

Meno, C., Gritsman, K., Ohishi, S., Ohfuji, Y., Heckscher, E., Mochida, K., Shimono, A., Kondoh, H., Talbot, W. S., Robertson, E. J., Schier, A. F., & Hamada, H. (1999). Mouse Lefty2 and zebrafish Antivin are feedback inhibitors of Nodal signalling during vertebrate gastrulation. *Molecular Cell*, *4*(3), 287–98.

Meshcheryakov, V. N., & Beloussov, L. V. (1975). Asymmetrical rotations of blastomeres in early cleavage of Gastropoda. *Development, Genes and Evolution*, *177*(3), 193–203.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics*, *11*(1), 31–46.

Mikawa, T., Poh, A. M., Kelly, K. A., Ishii, Y., & Reese, D. E. (2004). Induction and patterning of the primitive streak, an organizing center of gastrulation in the amniote. *Developmental Dynamics*, *229*(3), 422–432.

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, *95*(6), 315 – 327.

Molina, M. D., de Croze, N., Haillot, E., & Lepage, T. (2013). Nodal: master and commander of the dorsal-ventral and left-right axes in the sea urchin embryo. *Current Opinion in Genetics and Development*, *23*(1), 445–453.

Molina, M. D., Neto, A., Maeso, I., Gomez-Skarmeta, J. L., Salo, E., & Cebria, F. (2011). *Noggin* and *Noggin-like* genes control dorsoventral axis regeneration in Planarians. *Current Biology*, *21*(4), 300–305.

Morgan, T. H. (1917). The theory of the gene. *The American Naturalist*, *51*(609), 513–544.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., & Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research*, *35*(Web Server issue), W182–5.

Moroz, L. (2012). Phylogenomics meets neuroscience: How many times might complex brains have evolved? *Acta Biologica Hungarica*, *63*, 3–19.

Morozova, O., Hirst, M., & Marraey, M. (2009). Applications of new sequencing technologies for transcriptome analysis. *Annual Reviews in Genomics and Human Genetics*, *10*, 13551.

Moustakas, A., & Heldin, C. H. (2009). The regulation of TGF β signal transduction. *Development*, *136*(22), 3699–714.

Mukherjee, T., Mandal, D., & Bhaduri, A. (2001). *Leishmania* plasma membrane Mg2+-ATPase is a H+/K+-antiporter involved in glucose symport. studies with sealed ghosts and vesicles of opposite polarity. *Journal of Biological Chemistry*, *276*(8), 5563–9.

Munson, K., Lambrecht, N., Shin, J. M., & Sachs, G. (2000). Analysis of the membrane domain of the gastric H(+)/K(+)-ATPase. *Journal of Experimental Biology*, *203*(Pt 1), 161–70.

Murcia, N. S., Richards, W. G., Yoder, B. K., Mucenski, M. L., Dunlap, J. R., & Woychik, R. P. (2000). The *Oak Ridge Polycystic Kidney (ORPK)* disease gene is required for left-right axis determination. *Development*, *127*(11), 2347–55.

Myers, E. W. (1995). Towards simplifying and accurately formulating fragment assembly. *Journal of Computational Biology*, *2*(1), 275–290.

Nagarajan, N., & Pop, M. (2013). Sequence assembly demystified. *Nature Reviews Genetics*, *14*(3), 157–167.

Nakano, T., & Ozawa, T. (2007). Worldwide phylogeography of limpets of the order Patellogastropoda: molecular, morphological and palaeontological evidence. *Journal of Molluscan Studies*, *73*(1), 79–99.

Nakano, T., & Sasaki, T. (2011). Recent advances in molecular phylogeny, systematics and evolution of Patellogastropod limpets. *Journal of Molluscan Studies*, *77*(3), 203–217.

NCBI (2011). The National Center for Biotechnology Information website. `http://www.ncbi.nlm.nih.gov/`.

Nederbragt, A. J., van Loon, A. E., & Dictus, W. J. A. G. (2002). Expression of *Patella vulgata* orthologs of *engrailed* and *dpp-BMP2/4* in adjacent domains during molluscan shell development suggests a conserved compartment boundary mechanism. *Developmental Biology*, *246*(2), 341 – 355.

Nelsen, E., Frankel, J., & Jenkins, L. (1989). Non-genic inheritance of cellular handedness. *Development*, *105*(3), 447–456.

Neugebauer, J. M., Amack, J. D., Peterson, A. G., Bisgrove, B. W., & Yost, H. J. (2009). FGF signalling during embryo development regulates cilia length in diverse epithelia. *Nature*, *458*(7238), 651–654.

Neville, A. (1976). *Animal Asymmetry*. London: Edward Arnold.

Nielsen, C. (2010). Some aspects of spiralian development. *Acta Zoologica*, *91*(1), 20–28.

Nishide, K., Mugitani, M., Kumano, G., & Nishida, H. (2012). Neurula rotation determines left-right asymmetry in ascidian tadpole larvae. *Development*, *139*(8), 1467–1475.

Nogi, T., Yuan, Y. E., Sorocco, D., Perez-Tomas, R., & Levin, M. (2005). Eye regeneration assay reveals an invariant functional left-right asymmetry in the early bilaterian, *Dugesia japonica*. *Laterality*, *10*(3), 193–205.

Nonaka, S., Shiratori, H., Saijoh, Y., & Hamada, H. (2002). Determination of left-right patterning of the mouse embryo by artificial Nodal flow. *Nature*, *418*(6893), 96–9.

Nonaka, S., Tanaka, Y., Okada, Y., Takeda, S., Harada, A., Kanai, Y., Kido, M., & Hirokawa, N. (1998). Randomization of left-right asymmetry due to loss of Nodal cilia generating leftward flow of extraembryonic fluid in mice lacking KIF3b motor protein. *Cell*, *95*(6), 829–37.

Nonaka, S., Yoshiba, S., Watanabe, D., Ikeuchi, S., Goto, T., Marshall, W. F., & Hamada, H. (2005). *De novo* formation of left/right asymmetry by posterior tilt of Nodal cilia. *PLoS Biology*, *3*(8), e268.

Norris, D. P., Brennan, J., Bikoff, E. K., & Robertson, E. J. (2002). The FoxH1-dependent autoregulatory enhancer controls the level of Nodal signals in the mouse embryo. *Development*, *129*(14), 3455–3468.

Norris, D. P., & Robertson, E. J. (1999). Asymmetric and node-specific Nodal expression patterns are controlled by two distinct cis-acting regulatory elements. *Genes and Development*, *13*(12), 1575–1588.

Nusslein-Volhard, C. (1991). Determination of the embryonic axes of *Drosophila*. *Development*, *113*, 1–10.

Oelgeschlager, M., Larrain, J., Geissert, D., & De Robertis, E. M. (2000). The evolutionarily conserved BMP-binding protein Twisted gastrulation promotes BMP signalling. *Nature*, *405*(6788), 757–763.

Okada, Y., Nonaka, S., Tanaka, Y., Saijoh, Y., Hamada, H., & Hirokawa, N. (1999). Abnormal Nodal flow precedes *situs inversus* in iv and inv mice. *Molecular Cell*, *4*(4), 459–68.

Okada, Y., Takeda, S., Tanaka, Y., Belmonte, J. C., & Hirokawa, N. (2005). Mechanism of Nodal flow: a conserved symmetry breaking event in left-right axis determination. *Cell*, *121*(4), 633–44.

Oki, S., Kitajima, K., & Meno, C. (2010). Dissecting the role of FGF signalling during gastrulation and left-right axis formation in mouse embryos using chemical inhibitors. *Developmental Dynamics*, *239*(6), 1768–1778.

Okumura, T., Utsuno, H., Kuroda, J., Gittenberger, E., Asami, T., & Matsuno, K. (2008). The development and evolution of left-right asymmetry in invertebrates: Lessons from *Drosophila* and snails. *Developmental Dynamics*, *237*(12), 3497–3515.

Oliverio, M., Digilio, M. C., Versacci, P., Dallapiccola, B., & Marino, B. (2010). Shells and heart: Are human laterality and chirality of snails controlled by the same maternal genes? *American Journal of Medical Genetics Part A*, *152A*(10), 2419–2425.

Onichtchouk, D., Chen, Y. G., Dosch, R., Gawantka, V., Delius, H., Massagué, J., & Niehrs, C. (1999). Silencing of TGF β signalling by the pseudoreceptor BAMBI. *Nature*, *401*(6752), 480–5.

Orton, J. H., Southward, A. J., & Dodd, J. M. (1956). Studies on the biology of limpets: The breeding of *Patella vulgata* L. in Britain. *Journal of the Marine Biological Association of the United Kingdom*, *35*(01), 149–176.

Osman, A., Niles, E. G., Verjovski-Almeida, S., & LoVerde, P. T. (2006). *Schistosoma mansoni* TGF β receptor ii: Role in host ligand-induced regulation of a schistosome target gene. *PLoS Pathogens*, *2*(6), e54.

Pagan-Westphal, S. M., & Tabin, C. J. (1998). The transfer of left-right positional information during chick embryogenesis. *Cell*, *93*(1), 25–35.

Palmer, A. R. (1996). From symmetry to asymmetry: phylogenetic patterns of asymmetry variation in animals and their evolutionary significance. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(25), 14279–86.

Palmer, A. R. (2004). Symmetry breaking and the evolution of development. *Science*, *306*(5697), 828–33.

Pan, Y., & Wang, B. (2007). A novel protein-processing domain in gli2 and gli3 differentially blocks complete protein degradation by the proteasome. *Journal of Biological Chemistry*, *282*(15), 10846–10852.

Pang, K., Ryan, J. F., Baxevanis, A. D., & Martindale, M. Q. (2011). Evolution of the TGF β signalling pathway and its potential role in the ctenophore, *Mnemiopsis leidyi*. *PLoS ONE*, *6*(9), e24152.

Papaioannou, V. E., & Silver, L. M. (1998). The T-box gene family. *Bioessays*, *20*(1), 9–19.

Paps, J., Holland, P. W., & Shimeld, S. M. (2012). A genome-wide view of transcription factor gene diversity in chordate evolution: less gene loss in amphioxus? *Briefings in Functional Genomics*, *11*(2), 177–186.

Paszkiewicz, K., & Studholme, D. J. (2010). *De novo* assembly of short sequence reads. *Briefings in Bioinformatics*, *11*(5), 457–472.

Patterson, G. I., & Padgett, R. W. (2000). TGF β-related pathways. roles in *Caenorhabditis elegans* development. *Trends in Genetics*, *16*(1), 27–33.

Paxton, H. (2009). Phylogeny of Eunicida (Annelida) based on morphology of jaws. *Proceedings of the 9th International Polychaete Conference*, *2*, 241–264.

Pearson, R. (2003). The determined embryo. In B. Hall, R. Pearson, & G. Mller (Eds.) *Environment, Development, and Evolution*, (p. 6769). Boston: MIT Press.

Pedersen, H., & Mygind, N. (1976). Absence of axonemal arms in nasal mucosa cilia in Kartagener's syndrome. *Nature*, *262*(5568), 494–495.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., & Parvizi, B. (2003). TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, *19*, 651 – 652.

Pevzner, P. A., Tang, H., & Waterman, M. S. (2001). An eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, *98*, 9748–9753.

Ponder, W. F., & Lindberg, D. R. (1997). Towards a phylogeny of gastropod molluscs: an analysis using morphological characters. *Zoological Journal of the Linnean Society*, *119*(2), 83–265.

Poole, R. J., & Hobert, O. (2006). Early embryonic programming of neuronal left/right asymmetry in *C. elegans*. *Current Biology*, *16*(23), 2279–92.

Pop, M., & Salzberg, S. L. (2008). Bioinformatics challenges of new sequencing technology. *Trends in Genetics*, *24*(3), 142 – 149.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., Bateman, A., & Finn, R. D. (2012). The Pfam protein families database. *Nucleic Acids Research*, *40*(D1), D290–D301.

Putnam, N., Butts, T., Ferrier, D., Furlong, R., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., Terry, A., Yu, J., & et al (2008). The amphioxus genome and the evolution of the chordate karyotype. *Nature*, *453*, 1064 – 1071.

Putnam, N. H., Srivastava, M., Hellsten, U., Dirks, B., Chapman, J., Salamov, A., Terry, A., Shapiro, H., Lindquist, E., Kapitonov, V. V., Jurka, J., Genikhovich, G., Grigoriev, I. V., Lucas, S. M., Steele, R. E., Finnerty, J. R., Technau, U., Martindale, M. Q., & Rokhsar, D. S. (2007). Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science*, *317*(5834), 86–94.

Qiu, D., Cheng, S.-M., Wozniak, L., McSweeney, M., Perrone, E., & Levin, M. (2005). Localization and loss-of-function implicates ciliary proteins in early, cytoplasmic roles in left-right asymmetry. *Developmental Dynamics*, *234*(1), 176–189.

Raghavan, N., & Knight, M. (2006). The snail (*Biomphalaria glabrata*) genome project. *Trends in Parasitology*, *22*, 148 – 151.

Ramsdell, A. F. (2005). Left-right asymmetry and congenital cardiac defects: getting to the heart of the matter in vertebrate left-right axis determination. *Developmental Biology*, *288*(1), 1–20.

Raya, A., & Belmonte, J. C. I. (2006). Left-right asymmetry in the vertebrate embryo: from early information to higher-level integration. *Nature Reviews Genetics*, *7*(4), 283–293.

Raya, A., Kawakami, Y., Rodriguez-Esteban, C., Buscher, D., Koth, C. M., Itoh, T., Morita, M., Raya, R. M., Dubova, I., Bessa, J. G., de la Pompa, J. L., & Izpisua Belmonte, J. C. (2003). Notch activity induces Nodal expression and mediates the establishment of left-right asymmetry in vertebrate embryos. *Genes and Development*, *17*(10), 1213–8.

Raya, A., Kawakami, Y., Rodriguez-Esteban, C., Ibanes, M., Rasskin-Gutman, D., Rodriguez-Leon, J., Buscher, D., Feijo, J. A., & Izpisua Belmonte, J. C. (2004). Notch activity acts as a sensor for extracellular calcium during vertebrate left-right determination. *Nature*, *427*(6970), 121–8.

Render, J. (1997). Cell fate maps in the *Ilyanassa obsoleta* embryo beyond the third division. *Developmental Biology*, *189*, 301 – 310.

Rentzsch, F., Anton, R., Saina, M., Hammerschmidt, M., Holstein, T. W., & Technau, U. (2006). Asymmetric expression of the bmp antagonists chordin and gremlin in the sea anemone *Nematostella vectensis*: Implications for the evolution of axial patterning. *Developmental Biology*, *296*(2), 375–387.

Ridgway, S. A., Reid, D. G., Taylor, J. D., Branch, G. M., & Hodgson, A. N. (1998). A cladistic phylogeny of the family Patellidae (Mollusca: Gastropoda). *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *353*(1375), 1645–1671.

Romer, A. (1949). *The Vertebrate Body*. Philadelphia: W.B. Saunders.

Ronquist, F., & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, *19*(12), 1572–4.

Rosenkranz, R., Borodina, T., Lehrach, H., & Himmelbauer, H. (2008). Characterizing the mouse ES cell transcriptome with Illumina sequencing. *Genomics*, *92*(4), 187–194.

Ross, S., & Hill, C. S. (2008). How the Smads regulate transcription. *The International Journal of Biochemistry and Cell Biology*, *40*(3), 383–408.

Rouse, G. (1998). The Annelida and their close relatives. In D. Anderson (Ed.) *Invertebrate Zoology*, (pp. 196–202). Oxford: Oxford University Press.

Ruppert, E., Fox, R., & Barnes, R. (2004). *Annelida in Invertebrate Zoology (7 ed)*. New York: Brooks and Cole.

Saijoh, Y., Adachi, H., Sakuma, R., Yeo, C.-Y., Yashiro, K., Watanabe, M., Hashiguchi, H., Mochida, K., Ohishi, S., Kawabata, M., Miyazono, K., Whitman, M., & Hamada, H. (2000). Left/right asymmetric expression of Lefty2 and Nodal is induced by a signalling pathway that includes the transcription factor FAST2. *Molecular Cell*, *5*(1), 35 – 47.

Saijoh, Y., Oki, S., Tanaka, C., Nakamura, T., Adachi, H., Yan, Y.-T., Shen, M. M., & Hamada, H. (2005). Two Nodal-responsive enhancers control left/right asymmetric expression of Nodal. *Developmental Dynamics*, *232*(4), 1031–1036.

Sakuta, H., Suzuki, R., Takahashi, H., Kato, A., Shintani, T., Iemura, S.-i., Yamamoto, T. S., Ueno, N., & Noda, M. (2001). Ventroptin: A BMP-4 antagonist expressed in a double-gradient pattern in the retina. *Science*, *293*(5527), 111–115.

Sampath, K., Cheng, A., Frisch, A., & Wright, C. (1997). Functional differences among *Xenopus* Nodal-related genes in left/right axis determination. *Development*, *124*(17), 3293–3302.

Sandquist, J., Kita, A. M., & Bement, W. M. (2011). And the dead shall rise: Actin and myosin return to the spindle. *Developmental Cell*, *21*(3), 410–419.

Sandtner, W., Egwolf, B., Khalili-Araghi, F., Sanchez-Rodrguez, J. E., Roux, B., Bezanilla, F., & Holmgren, M. (2011). Ouabain binding site in a functioning na+/k+ atpase. *Journal of Biological Chemistry*, *286*(44), 38177–38183.

Sarmah, B., Latimer, A. J., Appel, B., & Wente, S. R. (2005). Inositol polyphosphates regulate zebrafish left-right asymmetry. *Developmental Cell*, *9*(1), 133 – 145.

Sarmah, B., Winfrey, V. P., Olson, G. E., Appel, B., & Wente, S. R. (2007). A role for the inositol kinase ipk1 in ciliary beating and length maintenance. *Proceedings of the National Academy of Sciences*, *104*(50), 19843–19848.

Satir, P., & Christensen, S. T. (2007). Overview of structure and function of mammalian cilia. *Annual Review of Physiology*, *69*, 377–400.

Schier, A. F. (2003). Nodal signalling in vertebrate development. *Annual Review of Cell and Developmental Biology*, *19*, 589–621.

Schier, A. F. (2009). Nodal morphogens. *Cold Spring Harbor Perspectives in Biology*, *1*(5).

Schier, A. F., & Talbot, W. S. (2001). Nodal signalling and the zebrafish organizer. *International Journal of Developmental Biology*, *45*(1), 289–97.

Schilling, T., Concordet, J., & Ingham, P. (1999). Regulation of left/right asymmetries in the zebrafish by Shh and BMP4. *Developmental Biology*, *210*(2), 277 – 287.

Schilthuizen, M., & Davison, A. (2005). The convoluted evolution of snail chirality. *Naturwissenschaften*, *92*, 504–515.

Schmerer, M., Savage, R. M., & Shankland, M. (2009). Pax β: a novel family of lophotrochozoan Pax genes. *Evolution and Development*, *11*(6), 689–696.

Schneider, I., Houston, D. W., Rebagliati, M. R., & Slusarski, D. C. (2008). Calcium fluxes in dorsal forerunner cells antagonize β-catenin and alter left-right patterning. *Development*, *135*(1), 75–84.

Schochet, J. (1973). Opercular regulation in the polychaete *Hydroides dianthus*. i. Opercular ontogeny, distribution and flux. *The Biological Bulletin*, *144*(2), 400–420.

Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: Robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, *28*(8), 1086–1092.

Schuster, S. (2008). Next-generation sequencing transforms today's biology. *Nature Methods*, *5*, 16 – 18.

Scott, I. C., Blitz, I. L., Pappano, W. N., Imamura, Y., Clark, T. G., Steiglitz, B. M., Thomas, C. L., Maas, S. A., Takahara, K., Cho, K. W. Y., & Greenspan, D. S. (1999). Mammalian BMP-1/Tolloid-related metalloproteinases, including novel family member mammalian Tolloid-like 2, have differential enzymatic activities and distributions of expression relevant to patterning and skeletogenesis. *Developmental Biology*, *213*(2), 283–300.

Scott, I. C., Blitz, I. L., Pappano, W. N., Maas, S. A., Cho, K. W. Y., & Greenspan, D. S. (2001). Homologues of twisted gastrulation are extracellular cofactors in antagonism of bmp signalling. *Nature*, *410*(6827), 475–478.

Segers, H. (2002). The nomenclature of the Rotifera: Annotated checklist of valid family- and genus-group names. *Journal of Natural History*, *36*, 631–640.

Segrove, F. (1941). The development of the Serpulid *Pomatoceros triqueter* L. *Quarterly Journal of Microscopical Science*, *82*, 467 – 540.

Shankland, M., & Seaver, E. C. (2000). Evolution of the bilaterian body plan: What have we learned from annelids? *Proceedings of the National Academy of Sciences*, *97*(9), 4434–4437.

Shearer, T., & Snell, T. (2007). Transfection of siRNA into *Brachionus plicatilis*. *Hydrobiologia*, *593*, 141–145.

Shen, M. M. (2007). Nodal signalling: developmental roles and regulation. *Development*, *134*(6), 1023–34.

Shen, M. M., & Schier, A. F. (2000). The EGF CFC gene family in vertebrate development. *Trends in Genetics*, *16*(7), 303–9.

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D., & Church, G. M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, *309*(5741), 1728–1732.

Shenefelt, R. E. (1972). Morphogenesis of malformations in hamsters caused by retinoic acid: relation to dose and stage at treatment. *Teratology*, *5*(1), 103–18.

Shi, Y., & Massagué, J. (2003). Mechanisms of TGF β signalling from cell membrane to the nucleus. *Cell*, *113*(6), 685–700.

Shibazaki, Y., Shimizu, M., & Kuroda, R. (2004). Body handedness is directed by genetically determined cytoskeletal dynamics in the early embryo. *Current Biology*, *14*(16), 1462 – 1467.

Shimeld, S. M., Boyle, M. J., Brunet, T., Luke, G. N., & Seaver, E. C. (2010a). Clustered Fox genes in lophotrochozoans and the evolution of the bilaterian Fox gene cluster. *Developmental Biology*, *340*(2), 234–48.

Shimeld, S. M., Degnan, B., & Luke, G. N. (2010b). Evolutionary genomics of the Fox genes: origin of gene families and the ancestry of gene clusters. *Genomics*, *95*(5), 256–60.

Shimeld, S. M., & Levin, M. (2006). Evidence for the regulation of left-right asymmetry in *Ciona intestinalis* by ion flux. *Developmental Dynamics*, *235*(6), 1543–1553.

Shimeld, S. M., Purkiss, A. G., Dirks, R. P., Bateman, O. A., Slingsby, C., & Lubsen, N. H. (2005). Urochordate $\beta\gamma$-crystallin and the evolutionary origin of the vertebrate eye lens. *Current Biology*, *15*(18), 1684–9.

Shu, X., Huang, J., Dong, Y., Choi, J., Langenbacher, A., & Chen, J.-N. (2007). Na,K-ATPase α2 and Ncx4a regulate zebrafish left-right patterning. *Development*, *134*(10), 1921–1930.

Simakov, O., Marletaz, F., Cho, S.-J., Edsinger-Gonzales, E., Havlak, P., Hellsten, U., Kuo, D.-H., Larsson, T., Lv, J., Arendt, D., Savage, R., Osoegawa, K., de Jong, P., Grimwood, J., Chapman, J. A., Shapiro, H., Aerts, A., Otillar, R. P., Terry, A. Y., Boore, J. L., Grigoriev, I. V., Lindberg, D. R., Seaver, E. C., Weisblat, D. A., Putnam, N. H., & Rokhsar, D. S. (2013). Insights into bilaterian evolution from three spiralian genomes. *Nature*, *493*(7433), 526–531.

Simonet, G., Claeys, I., & Broeck, J. (2002). Structural and functional properties of a novel serine protease inhibiting peptide family in arthropods. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, *132*(1), 247 – 255.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., & Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Research*, *19*(6), 1117–23.

Small, K., Brudno, M., Hill, M., & Sidow, A. (2007). A haplome alignment and reference sequence of the highly polymorphic *Ciona savignyi* genome. *Genome Biology*, *8*(3).

Smith, F. G. W. (1935). The development of *Patella vulgata*. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *225*(520), 95–125.

Smith, S. A., Wilson, N. G., Goetz, F. E., Feehery, C., Andrade, S. C. S., Rouse, G. W., Giribet, G., & Dunn, C. W. (2011). Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature*, *480*, 364–367.

Smith, W. C., McKendry, R., Ribisi, S., & Harland, R. M. (1995). A Nodal-related gene defines a physical and functional domain within the Spemann organizer. *Cell*, *82*(1), 37 – 46.

Sodergren, E., Weinstock, G. M., Davidson, E. H., Cameron, R. A., Gibbs, R. A., Angerer, R. C., Angerer, L. M., Arnone, M. I., Burgess, D. R., Burke, R. D., et al. (2006). The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science*, *314*(5801), 941–952.

Song, H., Hu, J., Chen, W., Elliott, G., Andre, P., Gao, B., & Yang, Y. (2010). Planar cell polarity breaks bilateral symmetry by controlling ciliary positioning. *Nature*, *466*(7304), 378–382.

Spéder, P., Adam, G., & Noselli, S. (2006). Type ID unconventional myosin controls left/right asymmetry in *Drosophila*. *Nature*, *440*(7085), 803–807.

Spéder, P., Petzoldt, A., Suzanne, M., & Noselli, S. (2007). Strategies to establish left/right asymmetry in vertebrates and invertebrates. *Current Opinion in Genetics and Development*, *17*(4), 351–358.

Srivastava, M., Begovic, E., Chapman, J., Putnam, N., Hellsten, U., Kawashima, T., Kuo, A., Mitros, T., Salamov, A., & Carpenter, M. (2008). The *Trichoplax* genome and the nature of placozoans. *Nature*, *454*, 955 – 960.

Srivastava, M., Simakov, O., Chapman, J., Fahey, B., Gauthier, M. E. A., Mitros, T., Richards, G. S., Conaco, C., Dacre, M., Hellsten, U., Larroux, C., Putnam, N. H., Stanke, M., Adamska, M., Darling, A., Degnan, S. M., Oakley, T. H., & et al (2010). The *Amphimedon queenslandica* genome and the evolution of animal complexity. *Nature*, *466*(7307), 720–726. 10.1038/nature09201.

Stechmann, A., & Schlegel, M. (1999). Analysis of the complete mitochondrial DNA sequence of the brachiopod *Terebratulina retusa* places Brachiopoda within the protostomes. *Proceedings of the Royal Society of London B: Biological Sciences*, *266*, 2043–2052.

Stelzer, C., & Snell, T. (2003). Induction of sexual reproduction in *Brachionus plicatilis* (Monogononta, Rotifera) by a density-dependent chemical cue. *Limnology and Oceanography*, *48*, 939–943.

Struck, T., Schult, N., Kusen, T., Hickman, E., Bleidorn, C., McHugh, D., & Halanych, K. (2007). Annelid phylogeny and the status of Sipuncula and Echiura. *BMC Evolutionary Biology*, *7*, 57.

Struck, T. H., Paul, C., Hill, N., Hartmann, S., Hosel, C., Kube, M., Lieb, B., Meyer, A., Tiedemann, R., Purschke, G., & Bleidorn, C. (2011). Phylogenomic analyses unravel annelid evolution. *Nature*, *471*(7336), 95–98.

Sturtevant, A. H. (1923). Inheritance of direction of coiling in *Limnaea*. *Science*, *58*(1501), 269–70.

Suatoni, E., Vicario, S., Rice, S., Snell, T., & Caccone, A. (2006). An analysis of species boundaries and biogeographic patterns in a cryptic species complex: the rotifer *Brachionus plicatilis*. *Molecular Phylogenetics and Evolution*, *41*, 86–98.

Suga, H., Ono, K., & Miyata, T. (1999). Multiple TGF β receptor related genes in sponge and ancient gene duplications before the parazoan-eumetazoan split. *FEBS Letters*, *453*(3), 346–50.

Suga, K., Mark Welch, D., Tanaka, Y., Sakakura, Y., & Hagiwara, A. (2007). Analysis of expressed sequence tags of the cyclically parthenogenetic rotifer *Brachionus plicatilis*. *PLoS ONE*, *2*, 671.

Sun, P. D., & Davies, D. R. (1995). The cystine-knot growth-factor superfamily. *Annual Review of Biophysics and Biomolecular Structure*, *24*, 269–91.

Supp, D., Brueckner, M., Kuehn, M., Witte, D., Lowe, L., McGrath, J., Corrales, J., & Potter, S. (1999). Targeted deletion of the ATP binding domain of left-right dynein confirms its role in specifying development of left-right asymmetries. *Development*, *126*(23), 5495–5504.

Supp, D. M., Witte, D. P., Potter, S. S., & Brueckner, M. (1997). Mutation of an axonemal dynein affects left-right asymmetry in *inversus viscerum* mice. *Nature*, *389*(6654), 963–6.

Surget-Groba, Y., & Montoya-Burgos, J. I. (2010). Optimization of *de novo* transcriptome assembly from next-generation sequencing data. *Genome Research*, *20*(10), 1432–1440.

Tabin, C. (2005). Do we know anything about how left/right asymmetry is first established in the vertebrate embryo? *Journal of Molecular Histology*, *36*, 317–323.

Tabin, C. (2006). The key to left-right asymmetry. *Cell*, *127*(1), 27 – 32.

Tabin, C. J., & Vogan, K. J. (2003). A two-cilia model for vertebrate left-right axis specification. *Genes and Development*, *17*(1), 1–6.

Tagawa, K., Humphreys, T., & Satoh, N. (2000). *T-brain* expression in the apical organ of hemichordate tornaria larvae suggests its evolutionary link to the vertebrate forebrain. *Journal of Experimental Zoology*, *288*(1), 23–31.

Takahashi, T., & Holland, P. W. H. (2004). Amphioxus and ascidian *DMBX* homeobox genes give clues to the vertebrate origins of midbrain development. *Development*, *131*(14), 3285–3294.

Takahashi, T., McDougall, C., Troscianko, J., Chen, W. C., Jayaraman-Nagarajan, A., Shimeld, S. M., & Ferrier, D. E. (2009). An EST screen from the annelid *Pomatoceros lamarckii* reveals patterns of gene loss and gain in animals. *BMC Evolutionary Biology*, *9*, 240.

Takatori, N., Butts, T., Candiani, S., Pestarino, M., Ferrier, D., Saiga, H., & Holland, P. (2008). Comprehensive survey and classification of homeobox genes in the genome of amphioxus, *Branchiostoma floridae*. *Development, Genes and Evolution*, *218*(11), 579–590.

Takeda, S., Yonekawa, Y., Tanaka, Y., Okada, Y., Nonaka, S., & Hirokawa, N. (1999). Left-right asymmetry and kinesin superfamily protein KIF3a: new insights in determination of laterality and mesoderm induction by KIF3a-/- mice analysis. *Journal of Cell Biology*, *145*(4), 825–36.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, *28*(10), 2731–2739.

Tanaka, C., Sakuma, R., Nakamura, T., Hamada, H., & Saijoh, Y. (2007). Long-range action of Nodal requires interaction with GDF1. *Genes and Development*, *21*(24), 3272–82.

Tanaka, Y., Okada, Y., & Hirokawa, N. (2005). FGF-induced vesicular release of sonic hedgehog and retinoic acid in leftward Nodal flow is critical for left-right determination. *Nature*, *435*, 172 – 177.

Taniguchi, Y., Furukawa, T., Tun, T., Han, H., & Honjo, T. (1998). Lim protein KyoT2 negatively regulates transcription by association with the RBP-J DNA-binding protein. *Molecular and Cellular Biology*, *18*(1), 644–654.

ten Dijke, P., & Arthur, H. M. (2007). Extracellular control of TGF β signalling in vascular development and disease. *Nauret Review of Molecular and Cell Biology*, *8*(11), 857–869. 10.1038/nrm2262.

The Honeybee Genome Sequencing Consortium (2006). Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature*, *443*(7114), 931–949.

Thompson, H., Shaw, M. K., Dawe, H. R., & Shimeld, S. M. (2012). The formation and positioning of cilia in *Ciona intestinalis* embryos in relation to the generation and evolution of chordate left/right asymmetry. *Developmental Biology*, *364*(2), 214 – 223.

Timmons, L., & Fire, A. (1998). Specific interference by ingested dsRNA. *Nature*, *395*, 854–56.

Toyoizumi, R., Kobayashi, T., Kikukawa, A., Oba, J., & Takeuchi, S. (1997). Adrenergic neurotransmitters and calcium ionophore-induced *situs inversus viscerum* in *Xenopus laevis* embryos. *Development, Growth and Differentiation*, *39*(4), 505–14.

Uehara, M., Yashiro, K., Takaoka, K., Yamamoto, M., & Hamada, H. (2009). Removal of maternal retinoic acid by embryonic CYP26 is required for correct Nodal expression during early embryonic patterning. *Genes and Development*, *23*(14), 1689–1698.

University of Wisconsin, T. (2005). Amplify3.

Vallier, L., Reynolds, D., & Pedersen, R. A. (2004). Nodal inhibits differentiation of human embryonic stem cells along the neuroectodermal default pathway. *Developmental Biology*, *275*(2), 403–21.

van den Biggelaar, J., & Guerrier, P. (1979). Dorsoventral polarity and mesentoblast determination as concomitant results of cellular interactions in the mollusk *Patella vulgata*. *Developmental Biology*, *68*, 462–71.

van den Biggelaar, J. A. M. (1977). Development of dorsoventral polarity and mesentoblast determination in *Patella vulgata*. *Journal of Morphology*, *154*(1), 157–186.

Van der Zee, M., da Fonseca, R. N., & Roth, S. (2008). TGF β signalling in *Tribolium*, vertebrate-like components in a beetle. *Development Genes and Evolution*, *218*, 203–213.

Vandenberg, L. N., & Levin, M. (2010). Far from solved: a perspective on what we know about early mechanisms of left-right asymmetry. *Developmental Dynamics*, *239*(12), 3131–46.

Verdonk, N., & van den Biggelaar, J. (1983). Early development and the formation of the germ layers. In N. Verdonk, J. van den Biggelaar, & A. Tompa (Eds.) *The Mollusca*, vol. Vol. 3, (p. 91122). New York: Academic Press.

Vignaud, T., Blanchoin, L., & Thiery, M. (2012). Directed cytoskeleton self-organization. *Trends in Cell Biology*, *22*(12), 671–682.

Vincent, S. D., Norris, D. P., Good, J. A. L., Constam, D. B., & Robertson, E. J. (2004). Asymmetric Nodal expression in the mouse is governed by the combinatorial activities of two distinct regulatory elements. *Mechanisms of Development*, *121*(11), 1403 – 1415.

Vinther, J., Sperling, E. A., Briggs, D. E. G., & Peterson, K. J. (2011). A molecular palaeobiological hypothesis for the origin of aplacophoran molluscs and their derivation from chiton-like ancestors. *Proceedings of the Royal Society B: Biological Sciences*, *279*, 1259–1268.

Wagner, G. (1996). Homologues, natural kinds and the evolution of modularity. *American Zoologist*, *36*(1), 36–43.

Wallace, R. L. (2002). Rotifers: exquisite metazoans. *Integrative and Comparative Biology.*, *42*, 660 – 667.

Wang, G., Cadwallader, A. B., Jang, D. S., Tsang, M., Yost, H. J., & Amack, J. D. (2011). The Rho kinase Rock2b establishes anteroposterior asymmetry of the ciliated Kupffer's vesicle in zebrafish. *Development*, *138*(1), 45–54.

Wang, X., Meng, X., Song, B., Qiu, X., & Liu, H. (2010a). Snps in the *myostatin* gene of the mollusk *Chlamys farreri*: association with growth traits. *Comparative biochemistry and physiology. Part B, Biochemistry and molecular biology*, *155*(3), 327–30.

Wang, X.-W., Luan, J.-B., Li, J.-M., Bao, Y.-Y., Zhang, C.-X., & Liu, S.-S. (2010b). *De novo* characterization of a whitefly transcriptome and analysis of its gene expression during development. *BMC Genomics*, *11*(1), 400.

Wanninger, A., Ruthensteiner, B., & Haszprunar, G. (2000). Torsion in *Patella caerulea* (mollusca, patellogastropoda): ontogenetic process, timing, and mechanisms. *Invertebrate Biology*, *119*(2), 177–187.

Wanninger, A., Ruthensteiner, B., Lobenwein, S., Salvenmoser, W., Dictus, W. J. A. G., & Haszprunar, G. (1999). Development of the musculature in the limpet *Patella* (Mollusca, Patellogastropoda). *Development, Genes and Evolution*, *209*, 226–238.

Warner, J., Lyons, D., & McClay, D. (2012). Left-right asymmetry in the sea urchin embryo: BMP and the asymmetrical origins of the adult. *PLoS Biology*, *10*(10), e1001404.

Werner, G., Gemmell, P., Grosser, S., Hamer, R., & Shimeld, S. (2012). Analysis of a deep transcriptome from the mantle tissue of *Patella vulgata* Linnaeus (Mollusca: Gastropoda: Patellidae) reveals candidate biomineralising genes. *Marine Biotechnology*, (pp. 1–14).

Whelan, S., & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution*, *18*(5), 691–699.

Whiteford, N., Haslam, N., Weber, G., Prügel-Bennett, A., Essex, J. W., Roach, P. L., Bradley, M., & Neylon, C. (2005). An analysis of the feasibility of short read sequencing. *Nucleic Acids Research*, *33*(19), e171.

Wilson, E. (1893). The mosaic theory of development. In *Biological Lectures delivered at the Marine Biological Laboratory*, (pp. 1–14). Boston: Ginn and Company.

Wilson, E. B. (1904a). Experimental studies in germinal localization. *Journal of Experimental Zoology*, *1*(2), 197–268.

Wilson, E. B. (1904b). Mosaic development in the Annelid egg. *Science*, *20*(518), 748–50.

Wood, W. B. (1991). Evidence from reversal of handedness in *C. elegans* embryos for early cell interactions determining cell fates. *Nature*, *349*(6309), 536–538.

Xu, J., Van Keymeulen, A., Wakida, N. M., Carlton, P., Berns, M. W., & Bourne, H. R. (2007). Polarity reveals intrinsic cell chirality. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(22), 9296–300.

Yamamoto, M., Mine, N., Mochida, K., Sakai, Y., Saijoh, Y., Meno, C., & Hamada, H. (2003). Nodal signalling induces the midline barrier by activating Nodal expression in the lateral plate. *Development*, *130*(9), 1795–804.

Yamamoto, Y., & Oelgeschlager, M. (2004). Regulation of bone morphogenetic proteins in early embryonic development. *Naturwissenschaften*, *91*(11), 519–34.

Yasui, K., Zhang, S., Uemura, M., & Saiga, H. (2000). Left-right asymmetric expression of *BbPtx*, a *Ptx*-related gene, in a lancelet species and the developmental left-sidedness in deuterostomes. *Development*, *127*(1), 187–195.

Yoshida, K., & Saiga, H. (2008). Left-right asymmetric expression of Pitx is regulated by the asymmetric Nodal signaling through an intronic enhancer in *Ciona intestinalis*. *Development, Genes and Evolution*, *218*(7), 353–360.

Yost, H. J. (1991). Development of the left-right axis in amphibians. *Ciba Foundation Symposia*, *162*, 165–76; discussion 176–81.

Yost, H. J. (1992). Regulation of vertebrate left-right asymmetries by extracellular matrix. *Nature*, *357*(6374), 158–61.

Yost, H. J. (1998). Left-right development in *Xenopus* and zebrafish. *Seminars in Cell and Developmental Biology*, *9*(1), 61 – 66.

Zerbino, D. R. (2010). Using the Velvet *de novo* assembler for short-read sequencing technologies. *Current Protocols in Bioinformatics*, *Chapter 11*, Unit 11.

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–9.

Zerbino, D. R., McEwen, G. K., Margulies, E. H., & Birney, E. (2009). Pebble and rock band: heuristic resolution of repeats and scaffolding in the Velvet short-read *de novo* assembler. *PLoS ONE*, *4*(12), e8407.

Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., Yang, P., Zhang, L., Wang, X., Qi, H., Xiong, Z., Que, H., Xie, Y., Holland, P. W. H., Paps, J., Zhu, Y., Wu, F., Chen, Y., Wang, J., Peng, C., Meng, J., Yang, L., & Wang, J. (2012a). The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*, *490*(7418), 49–54.

Zhang, L. N., Su, S. W., Guo, F., Guo, H. C., Shi, X. L., Li, W. Y., Liu, X., & Wang, Y. L. (2012b). Serotonin-mediated modulation of Na+/K+ pump current in rat hippocampal CA1 pyramidal neurons. *BMC Neuroscience*, *13*(1), 10.

Zhang, W., Chen, J., Yang, Y., Tang, Y., Shang, J., & Shen, B. (2011). A practical comparison of *de novo* genome assembly software tools for next-generation sequencing technologies. *PLoS ONE*, *6*(3), e17915.

Zhang, X. M., Ramalho-Santos, M., & McMahon, A. P. (2001). Smoothened mutants reveal redundant roles for Shh and Ihh signaling including regulation of L/R asymmetry by the mouse node. *Cell*, *105*(6), 781 – 792.

Zhang, Y., & Levin, M. (2009). Left-right asymmetry in the chick embryo requires core planar cell polarity protein Vangl2. *Genesis*, *47*(11), 719–28.

Zhong, Y. F., Butts, T., & Holland, P. W. (2008). HomeoDB: a database of homeobox gene diversity. *Evolution and Development*, *10*(5), 516–8.

Zhong, Y. F., & Holland, P. W. (2011). HomeoDB2: functional expansion of a comparative homeobox gene database for evolutionary developmental biology. *Evolution and Development*, *13*(6), 567–568.

Zhou, X., Sasaki, H., Lowe, L., Hogan, B. L., & Kuehn, M. R. (1993). Nodal is a novel TGF β-like gene expressed in the mouse node during gastrulation. *Nature*, *361*(6412), 543–7.

Appendix A: Abbreviations

Non S.I. Abbreviations used in text:

ALK = activin-like kinase
amp = ampicillin
AP = alkaline phosphatase
A/P = anterior/posterior
ASE = asymmetric element
ATP = adenosine triphosphate
BAMBI = BMP and activin membrane-bound inhibitor
BLAST = basic local alignment search tool
BMP = bone morphogenic protein
bp = base pairs
BSA = bovine serum albumin
cDNA = complementary deoxyribonucleic acid
contig = contiguous
CPU = central processing unit
DAPI = 4,6-diamidino-2-phenylindole
dATP = deoxyadenosinetriphosphate
ddNTP = dideoxy nucleoside triphosphate
DEPC = diethylpyrocarbonate
dH_2O = distilled water
DIG = digoxigenin
DNA = deoxyribonucleic acid
dNTP = deoxynucleotidetriphosphate
DTT = dithiothreitol
D/V = dorsal/ventral
FSW = filtered sea water
GABA = γ-aminobutyric acid
Gb = gigabase (1 000 000 000 base pairs)
hpf = hours post fertilization
IPTG =isopropyl-beta-D-thiogalactopyranoside
kb = kilobase (1000 base pairs)
KEGG = Kyoto encyclopedia of genes and genomes
LB = Luria broth
L/R = left/right
MAFFT = multiple alignment by fast Fourier transform
Mb = megabase (1,000,000 base pairs)
mRNA = messenger ribonucleic acid
miRNA = micro ribonucleic acid
NBT = nitro blue tetrazolium chloride
NCBI =National Centre for Biotechnology Information
NDE = node-specific enhancer
NOMO = *Nodal* modulator
OLC = overlap-layout-consensus (assembly algorithm)
PBS = phosphate buffered saline
PCR = polymerase chain reaction
PEE = proximal epiblast enhancer
pf = post fertilization
PTw = phosphate buffered Tween 20
PBTw = phosphate buffered Tween 20, with 0.1% (w/v) BSA
PPi = inorganic pyrophosphate
RA = retinoic acid
RACE = rapid amplification of chromosome ends

RAM = random access memory
RNA = ribonucleic acid
RNase = ribonucleic acid digesting enzyme
RNAi = ribonucleic acid interference
RPM = rotations per minute
SD = standard deviation
SOC = super optimal broth with catabolite repression
SSC = sodium chloride and sodium citrate
TGF = transforming growth factor
U = unit of enzyme, equivalent to 1/60 μkatal
v/v = volume per volume
w/v = weight per volume
X-gal = bromo-chloro-indolyl-galactopyranoside

Appendix B: Solutions Used

Constituents of solutions used in experimentation:

- *1 X Pomatoceros Wash:* 50% formamide, 1 x SSC, 0.1% Tween 80
- *2 X Pomatoceros Wash:* 50% formamide, 2 x SSC, 0.1% Tween 80
- *4 X Pomatoceros Wash:* 50% formamide, 4 x SSC, 0.1% Tween 80
- *Antibody solution:* Block solution with anti DIG-AP antibody added in a 1:3000 ratio. Placed on ice for 2 hours or longer, rocked gently for pre-absorption.
- *APT:* 100mM NaCl, 100 mM Tris HCl (pH 9.5), 50 mM MgCl2 and 0.1% Tween 80 in DEPC treated water.
- *Ciona (Patella vulgata) wash:* 50% formamide, 5x SSC, 0.1% Tween 80 in DEPC treated water.
- *DEPC water:* milliQ filtered ddH_2O with 0.1% v/v DEPC (Diethylpyrocarbonate), shaken thoroughly, left overnight at RT in a fume hood and autoclaved for in excess of 20 min.
- *Hybridisation solution:* The following made up in DEPC water. 50% formamide, 5 x SSC,100 μg/ml yeast RNA, 50 μg/ml heparin, 0.1% Tween 80.
- *MAB (maleic acid buffer):* 0.1 M maleic acid, 0.15 M NaCl, made in DEPC water, pH 7.5.
- *Magnesium-free AP solution:* 100 mM NaCl, 100 mM Tris HCl (pH 9.5) and 0.1% Tween 80 in DEPC treated water.
- *MOPS buffer:* $MgSO_4$ 10 mM, MOPS (3-(N-morpholino)propanesulfonic acid) 0.5 M, NaCl 2.5 M, pH 7.5
- *Patella Block solution:* 20% heat treated sheep serum in PBT.
- *PBS (Phosphate buffered saline solution):* 0.137 M NaCl, 0.0027 M KCl, 0.0101 M Na_2HPO_4, 0.0018 M KH_2PO_4 (pH 7.5), treated with DEPC and autoclaved.
- *PBT/ PTw:* PBS and 0.1%Tween 80.
- *PHB (2x):* 0.1 M EDTA, 0.02 M Tris (pH 8), 0.4% SDS in DEPC treated water.
- *Pomatoceros Hybridisation Solution:* 5 x SSC, 5 mM EDTA, 50% formamide, 100 μg/ml heparin, 0.1% Tween 20, 1 X Denhardts, 0.1 mg/ml total yeast RNA.
- *Staining solution:* APT to which 20 μl/ml of premixed NBT-BCIP stock is added.

Appendix C: Alignments

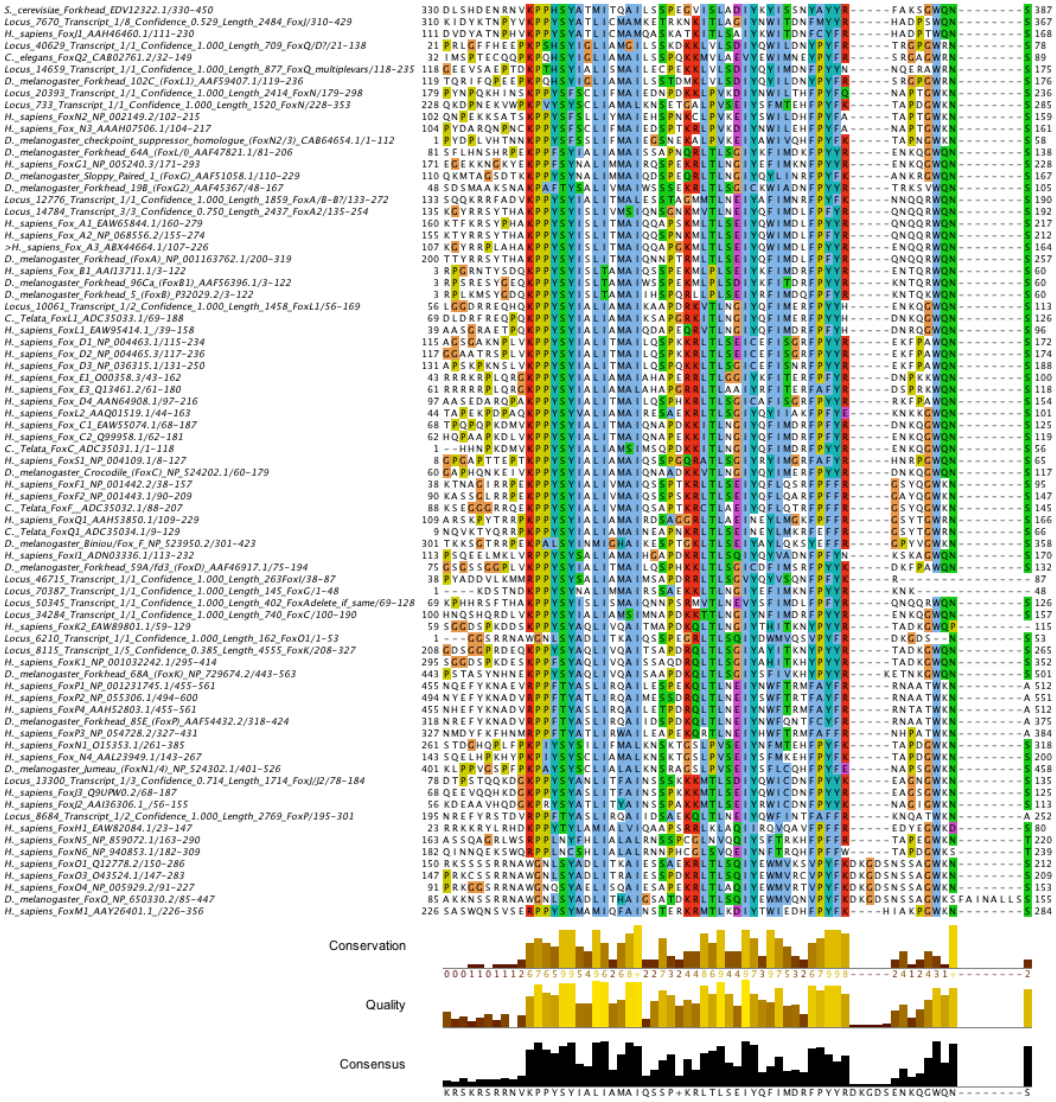


Figure 1: *Fox gene alignment used in Fig. 3.5. Alignment generated by MAFFT under the L-INS-i model as described in Methods, and displayed using Jalview with ClustalX colouring.*

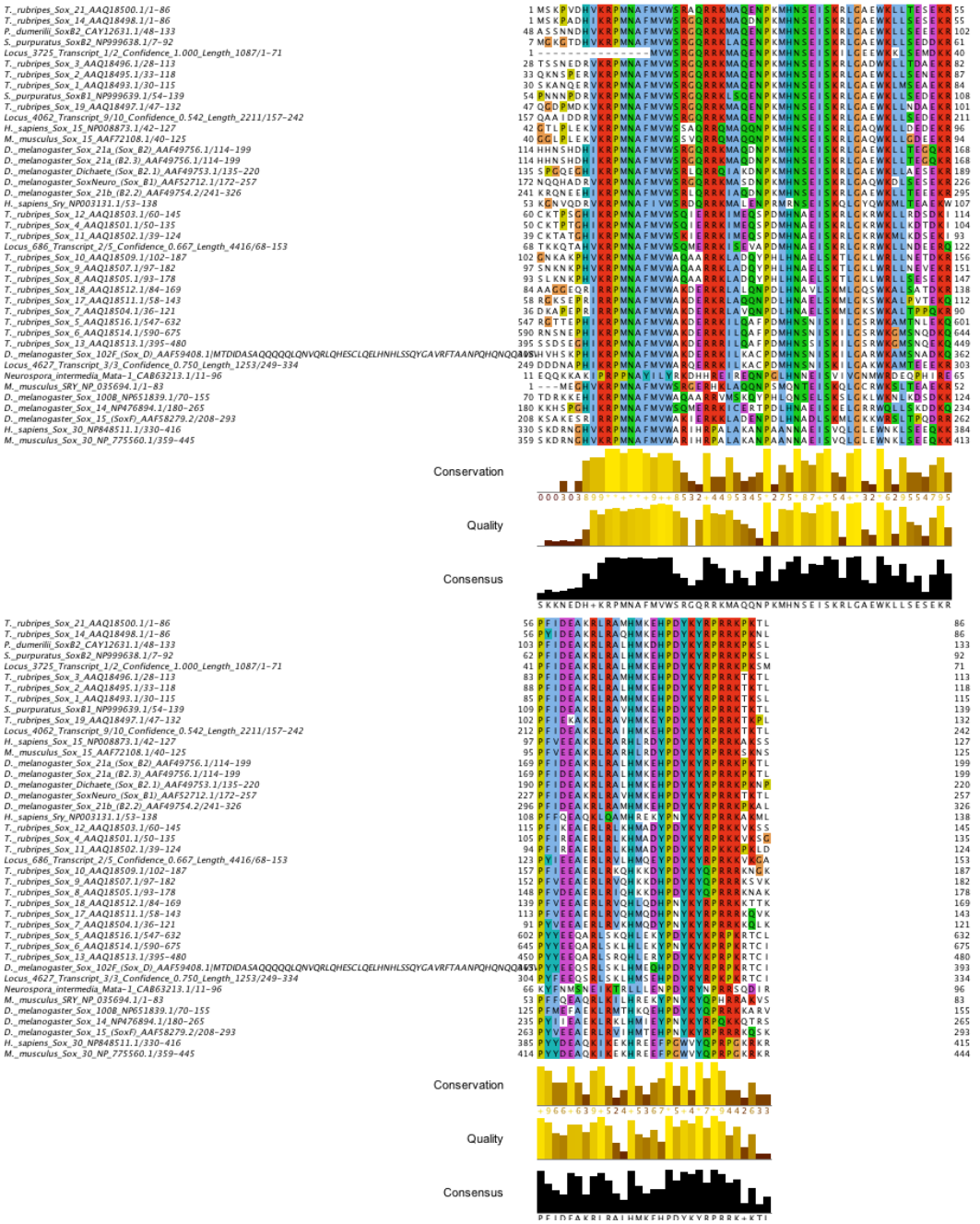


Figure 2: *Sox gene alignment used in Fig 3.6. Alignment generated by MAFFT under the L-INS-i model as described in Methods, and displayed using Jalview with ClustalX colouring.*

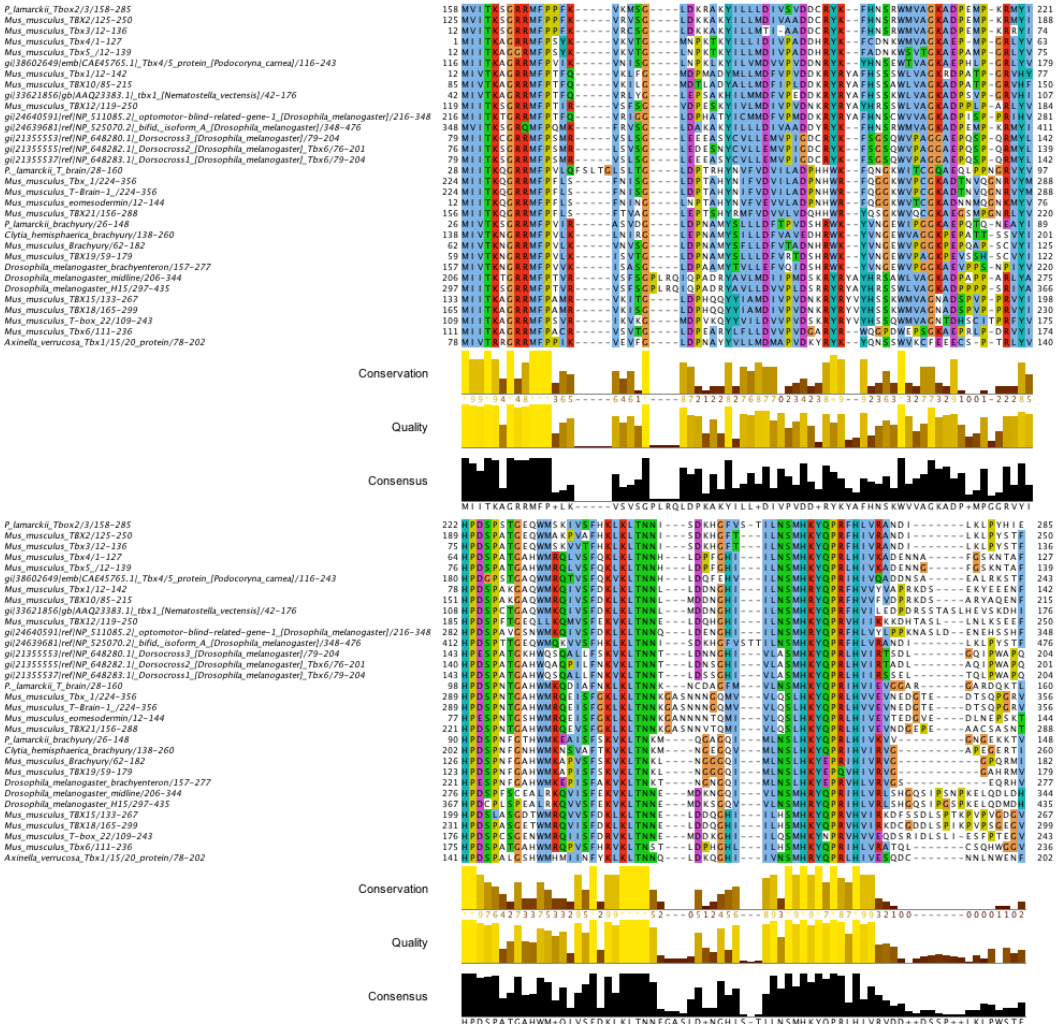


Figure 3: *T-box gene alignment used in Fig 3.7. Alignment generated by MAFFT under the L-INS-i model as described in Methods, and displayed using Jalview with ClustalX colouring.*

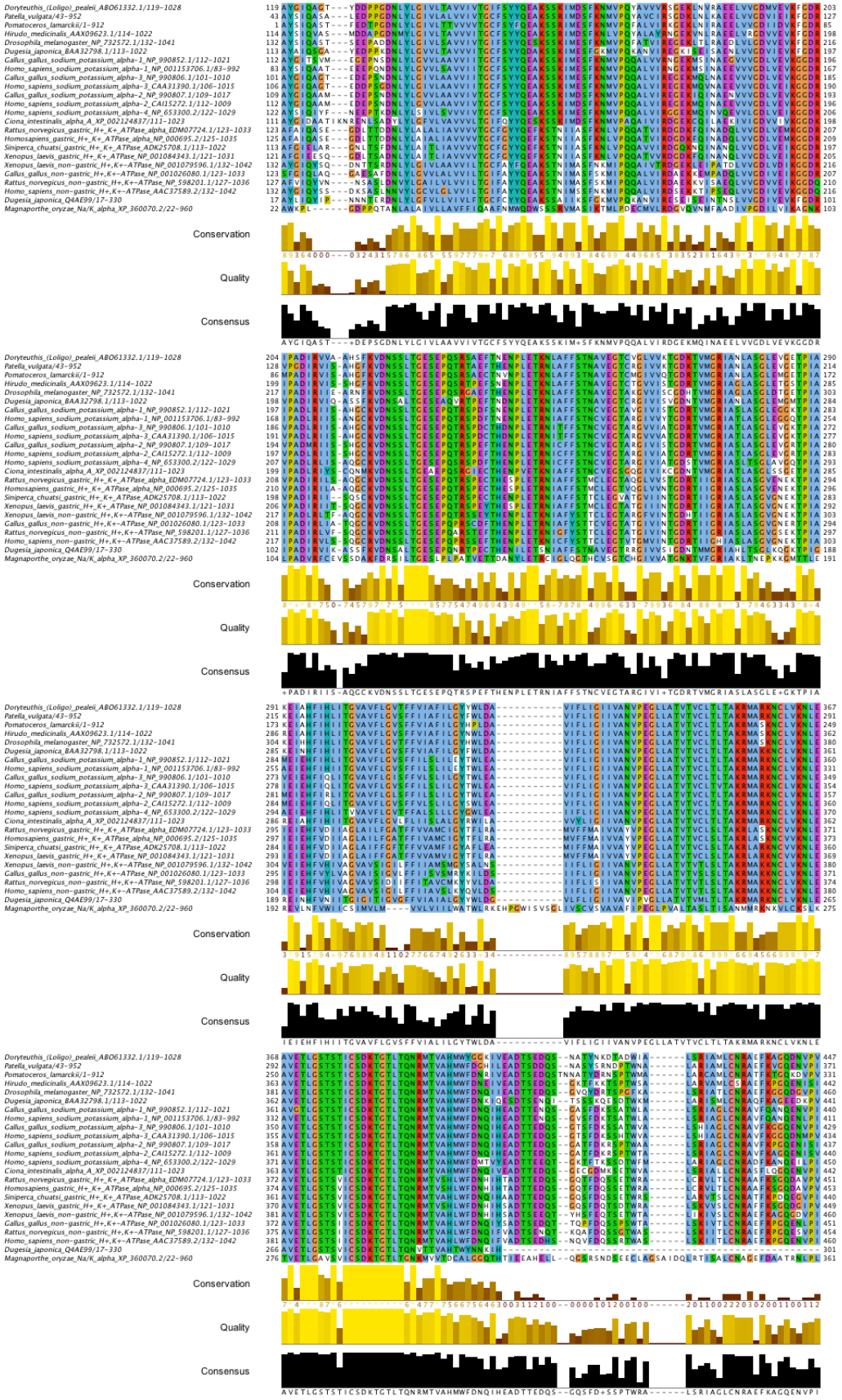


Figure 4: ATPase α *gene alignment used in Fig 4.1, Page 1, aligned in MAAFT under the G-INS-i model as described in Methods, and displayed using Jalview with ClustalX colouring.*

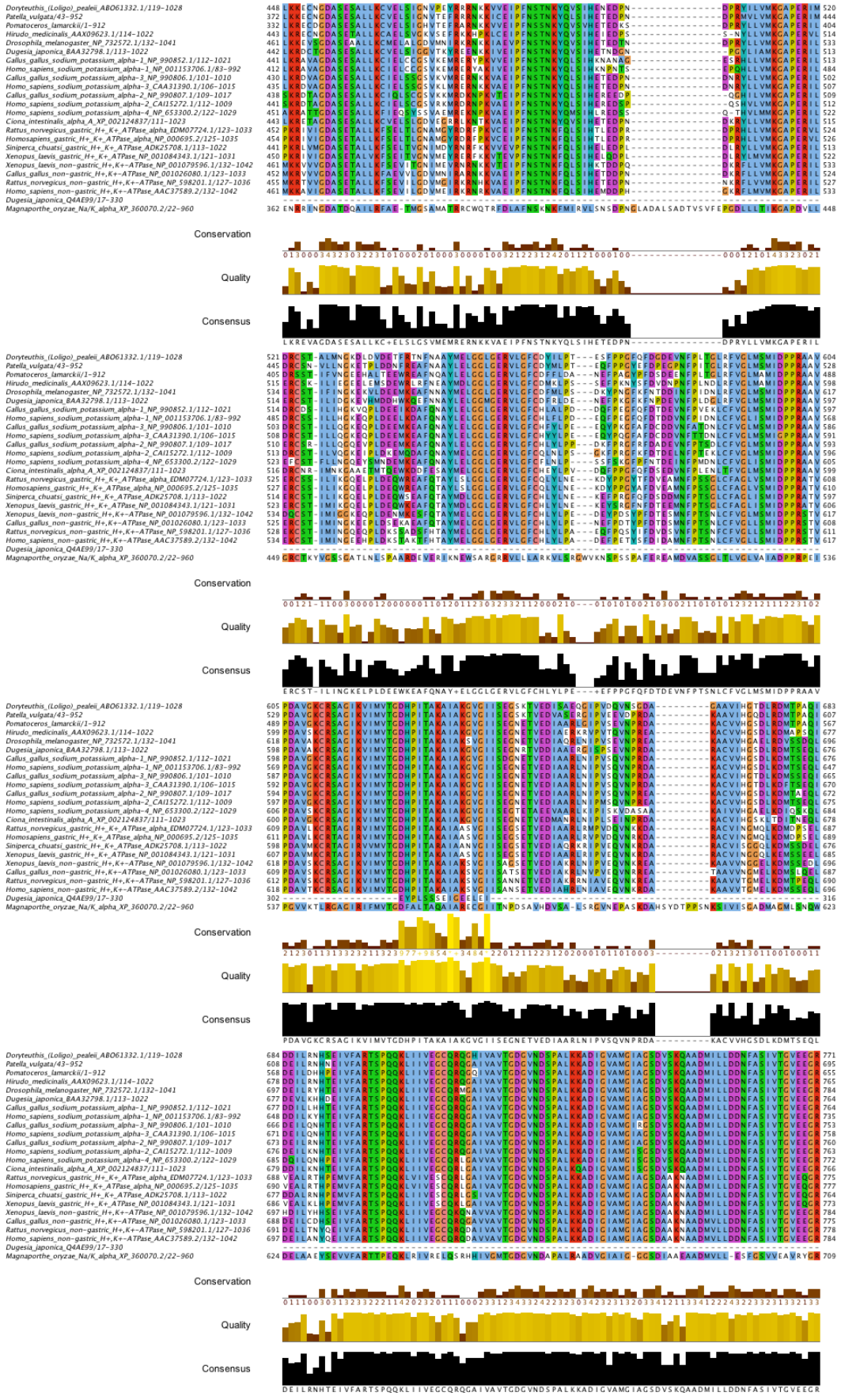


Figure 5: *ATPase α gene alignment continued from previous page.*

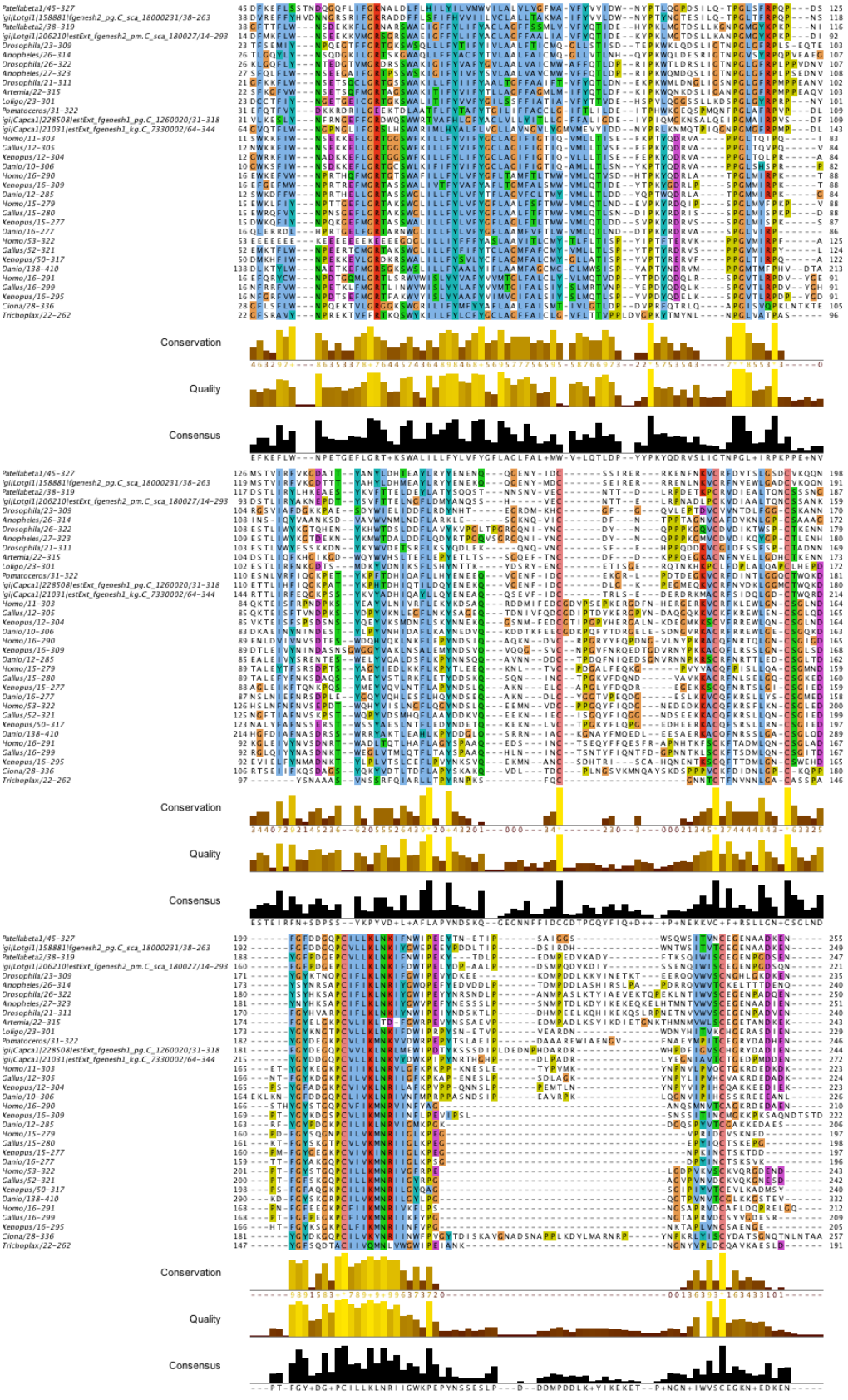


Figure 6: ATPase β *gene alignment used in Fig 4.2 , Page 1, aligned in MAAFT under the G-INS-i model as described in Methods, and displayed using Jalview with ClustalX colouring.*

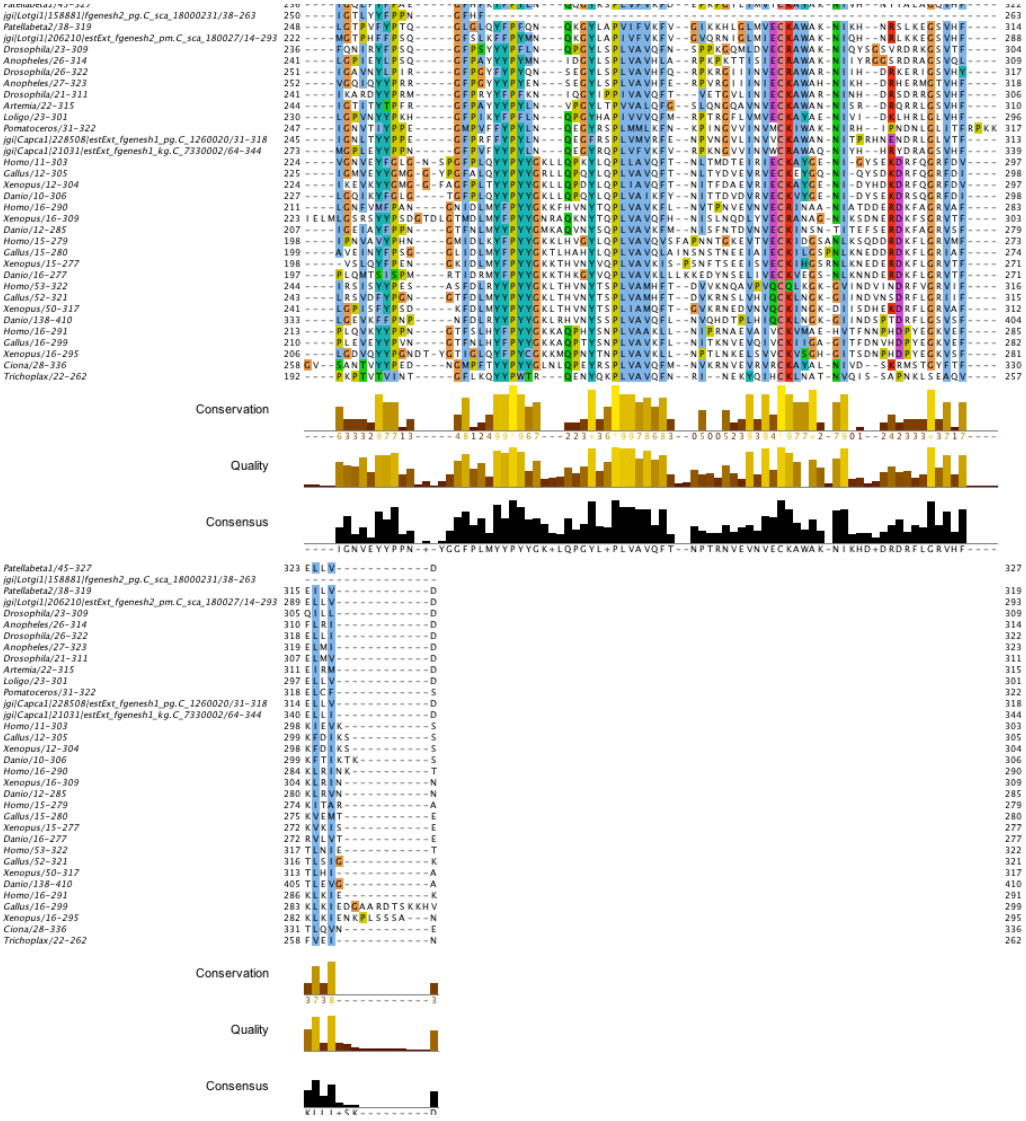


Figure 7: ATPase β *gene alignment used in Fig 4.2, continued from previous page.*

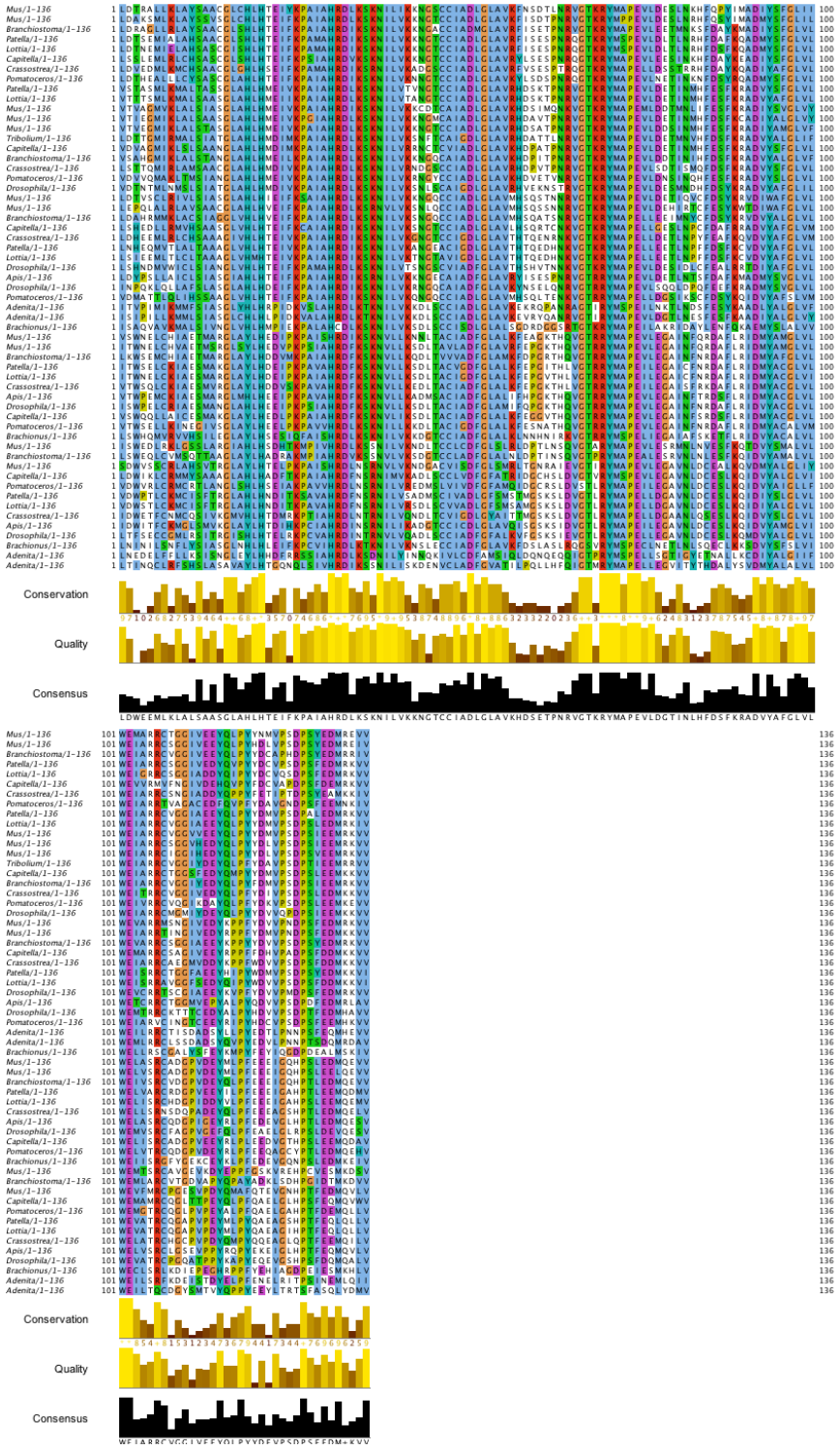


Figure 8: *ALK Family gene alignment used in Fig 5.9. Sequences aligned in MAAFT under the G-INS-i model as described in Methods, and displayed using Jalview with ClustalX colouring.*

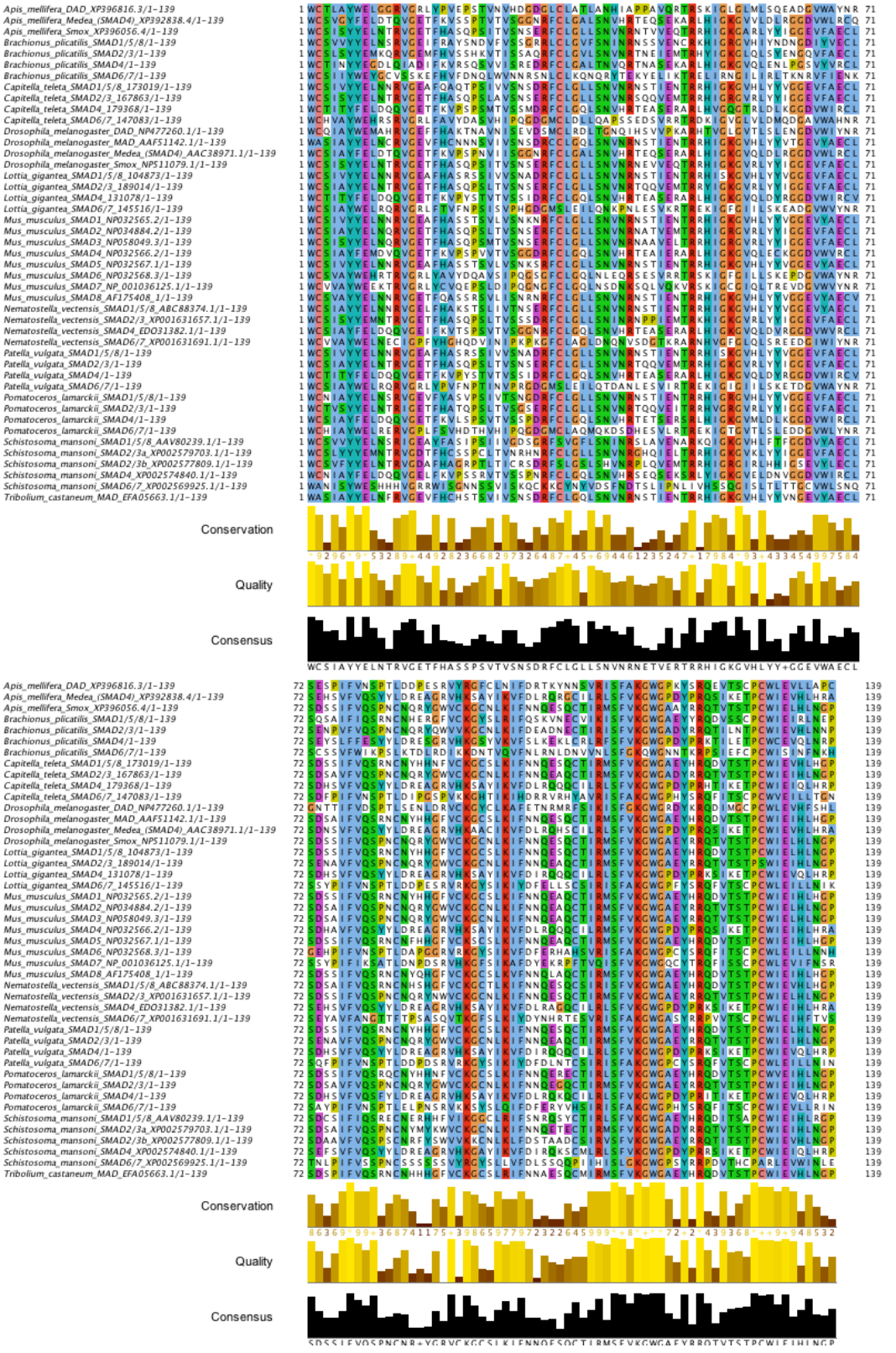


Figure 9: *SMAD Family gene alignment used in Fig 5.10. Sequences aligned in MAAFT under the G-INS-i model as described in Methods, and displayed using Jalview with ClustalX colouring.*

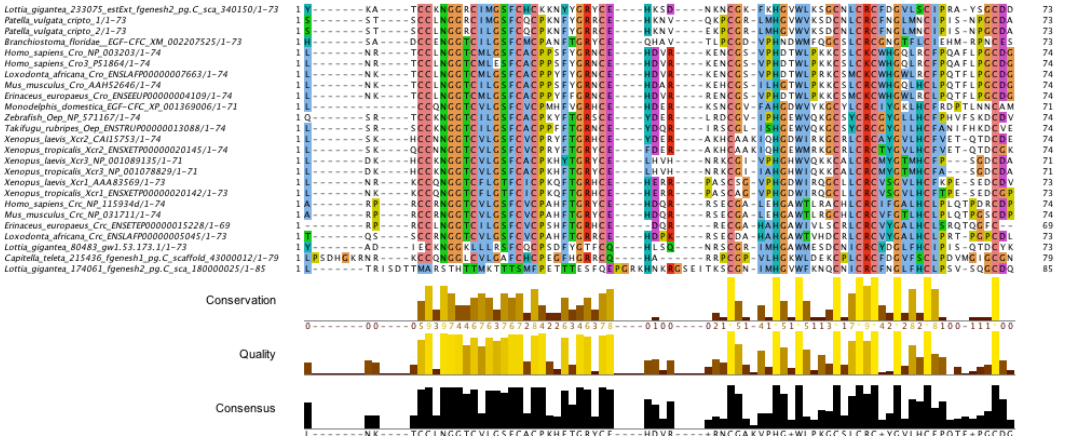


Figure 10: Cripto *gene alignment used in Fig 5.11. Sequences aligned in MAAFT under the G-INS-i model as described in Methods, and displayed using Jalview with ClustalX colouring.*

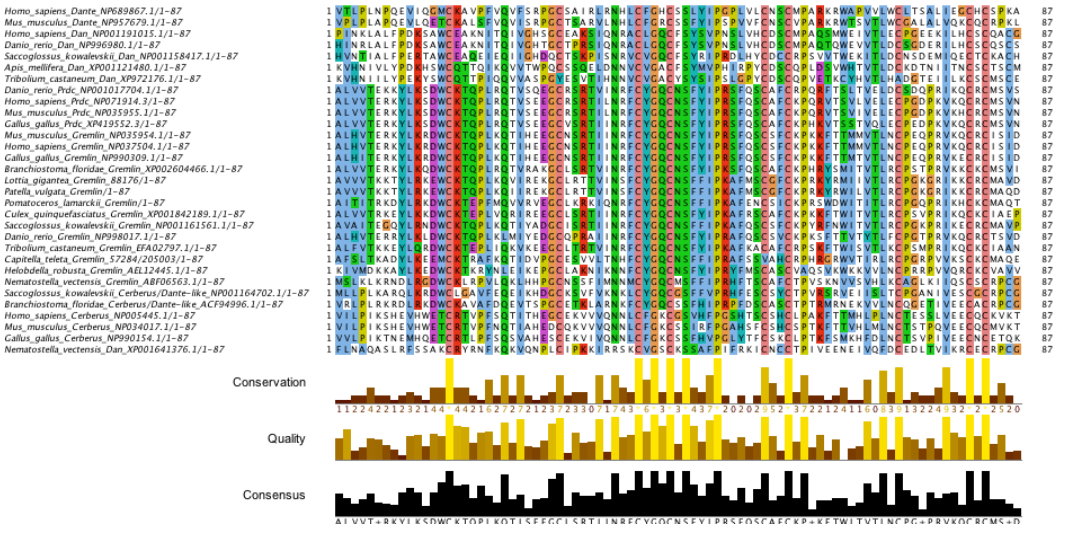


Figure 11: *DAN Family gene alignment used in Fig 5.12. Sequences aligned in MAAFT under the G-INS-i model as described in Methods, and displayed using Jalview with ClustalX colouring.*

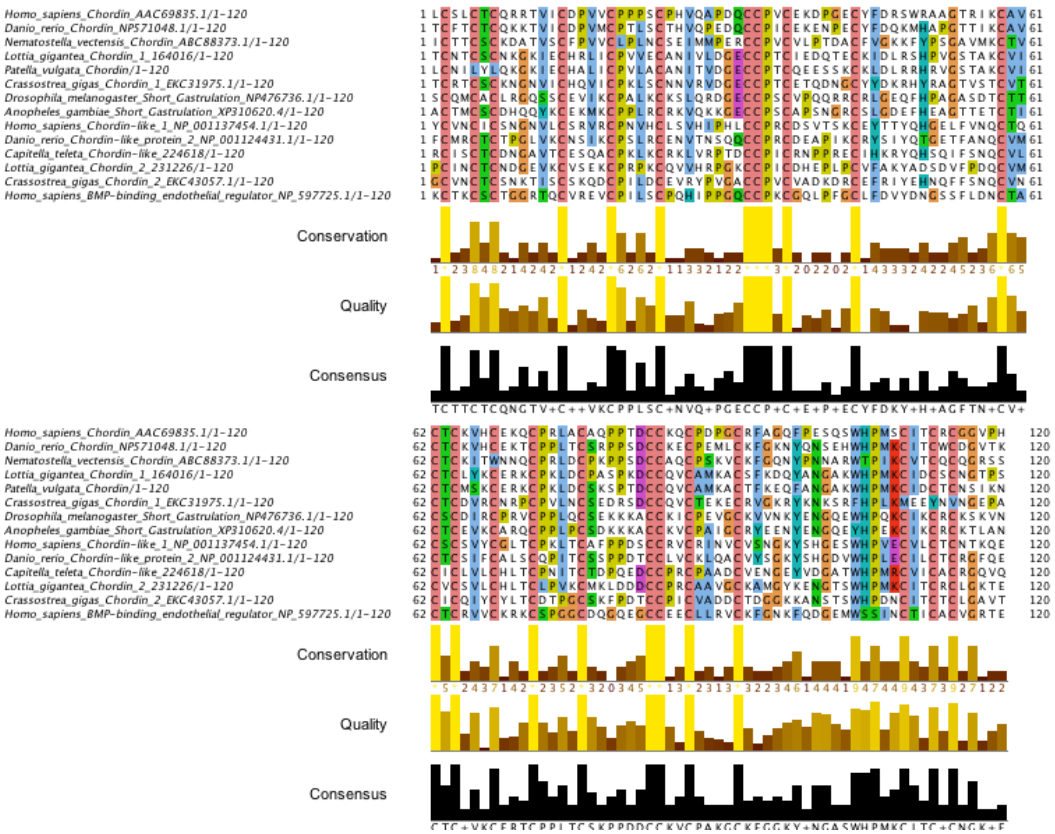


Figure 12: *Chordin gene alignment used in Fig 5.13. Sequences aligned in MAAFT under the G-INS-i model as described in Methods, and displayed using Jalview with ClustalX colouring.*

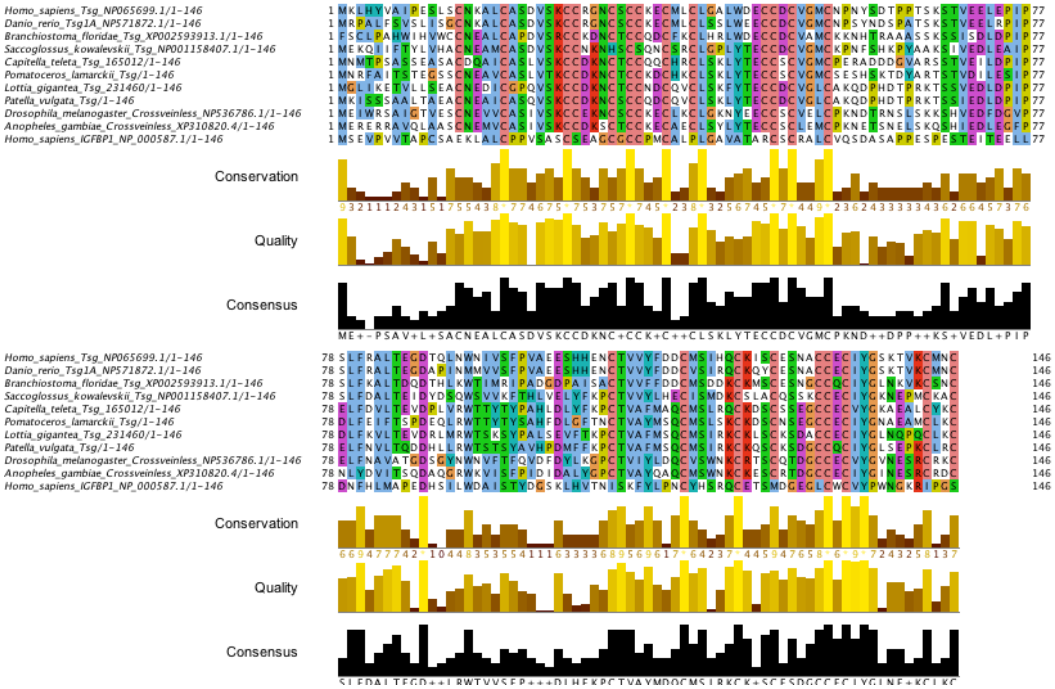


Figure 13: Tsg *gene alignment used in Fig 5.13. Sequences aligned in MAAFT under the G-INS-i model as described in Methods, and displayed using Jalview with ClustalX colouring.*

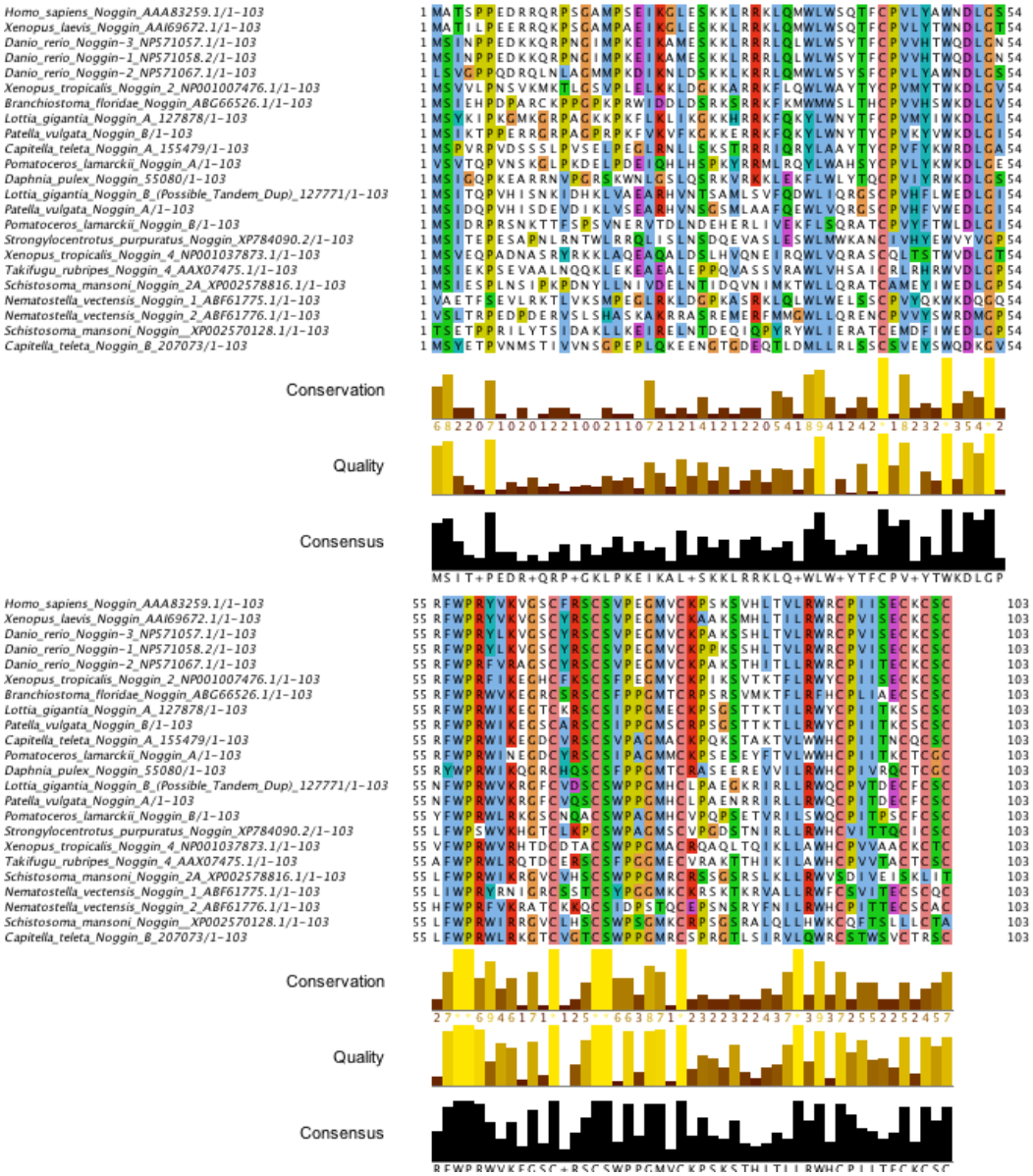


Figure 14: *Noggin gene alignment used in Fig 5.13. Sequences aligned in MAAFT under the G-INS-i model as described in Methods, and displayed using Jalview with ClustalX colouring.*

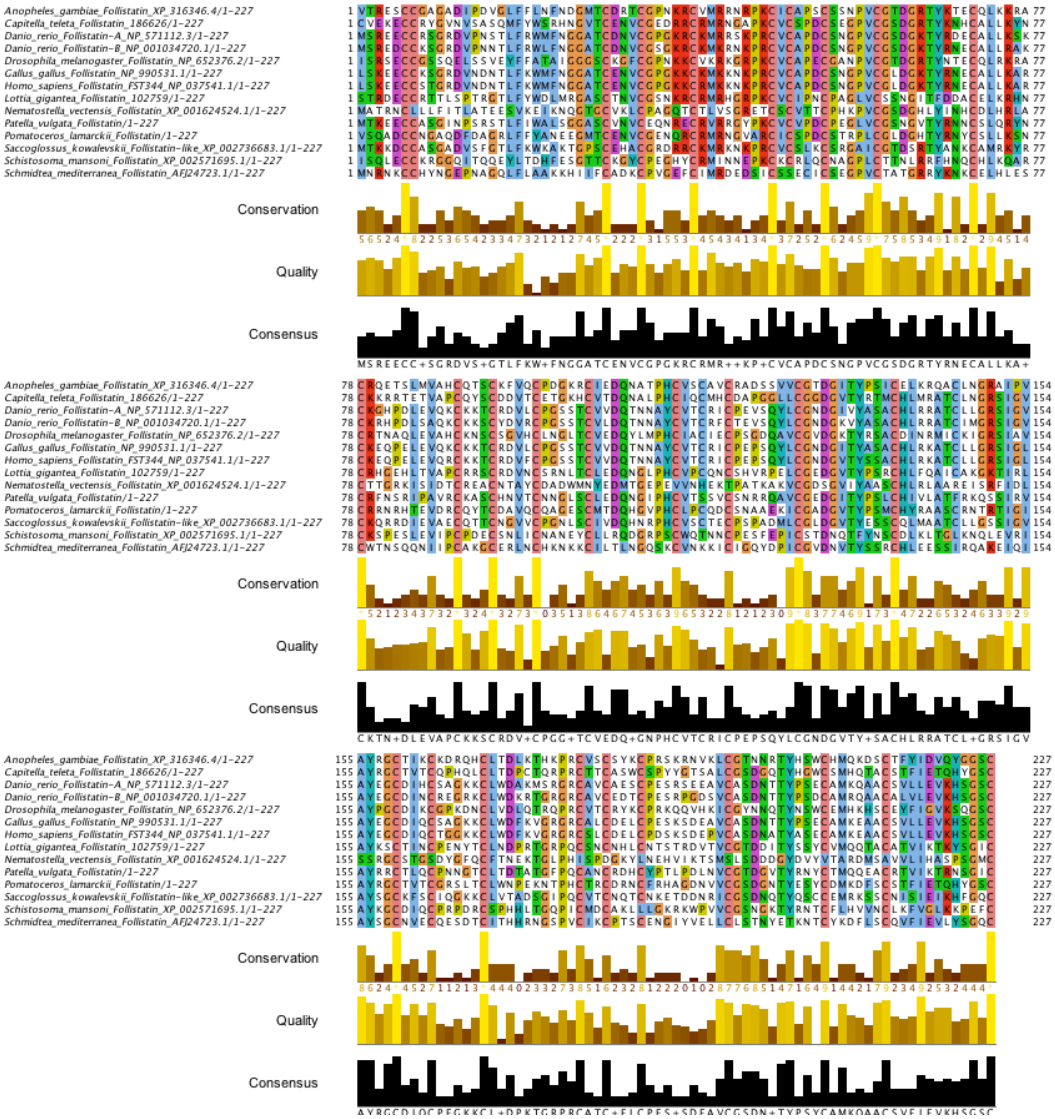


Figure 15: Follistatin *gene alignment used in Fig 5.14. Sequences aligned in MAAFT under the E-INS-i model as described in Methods, and displayed using Jalview with ClustalX colouring.*

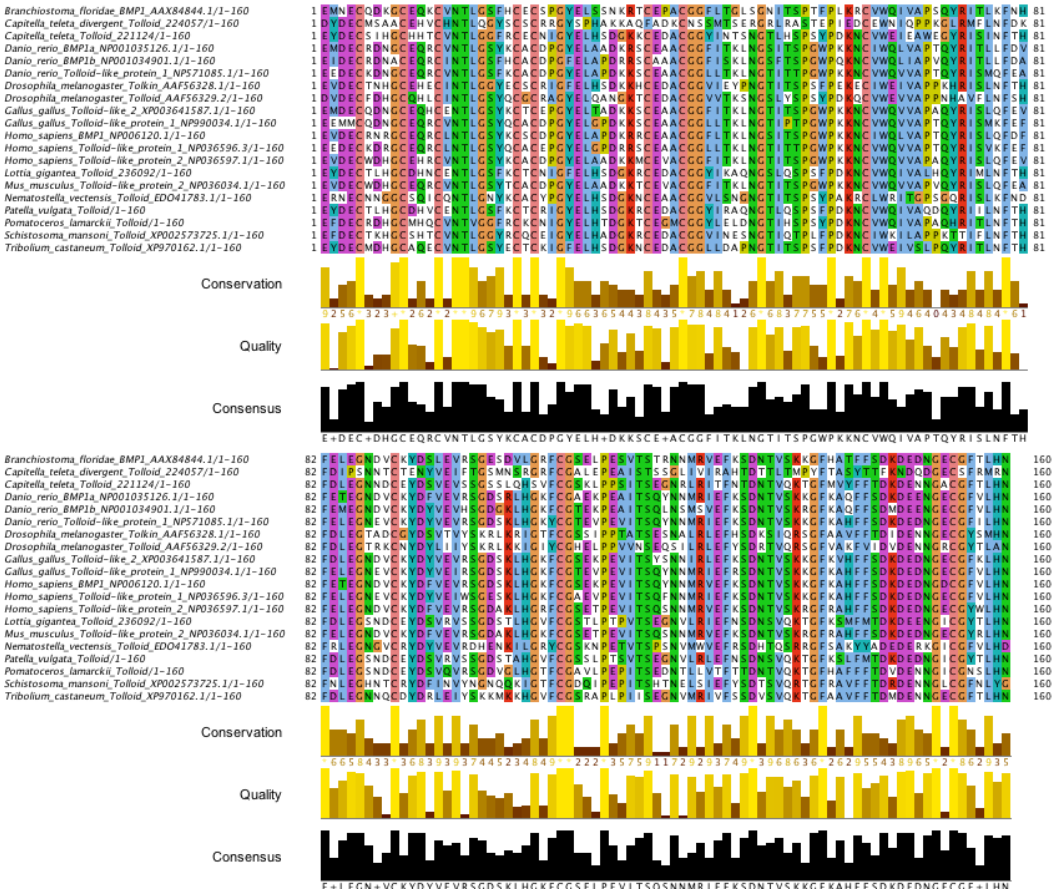


Figure 16: Tolloid *gene alignment used in Fig 5.14. Sequences aligned in MAAFT under the G-INS-i model as described in Methods, and displayed using Jalview with ClustalX colouring.*

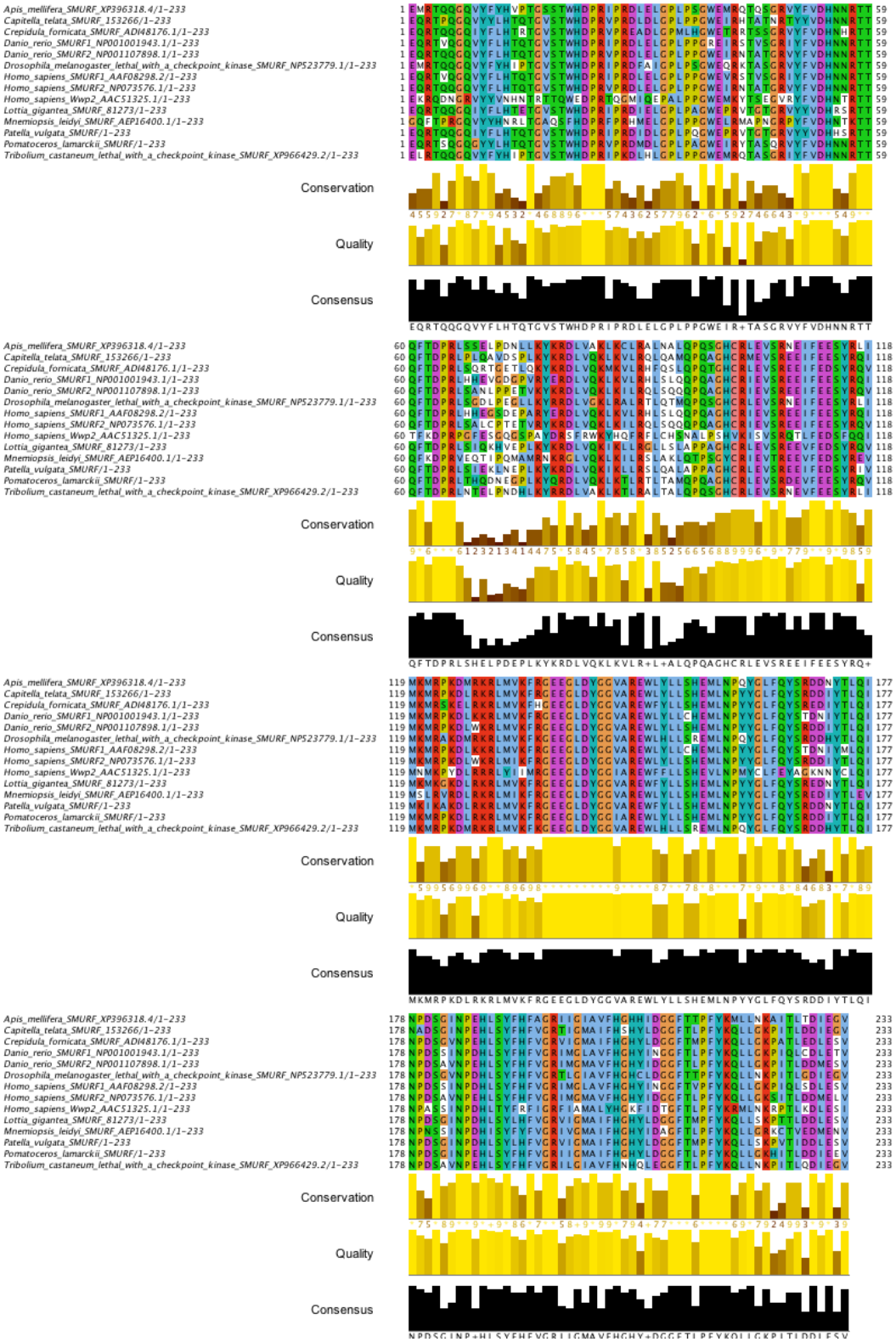


Figure 17: SMURF *gene alignment used in Fig 5.14. Sequences aligned in MAAFT under the G-INS-i model as described in Methods, and displayed using Jalview with ClustalX colouring.*

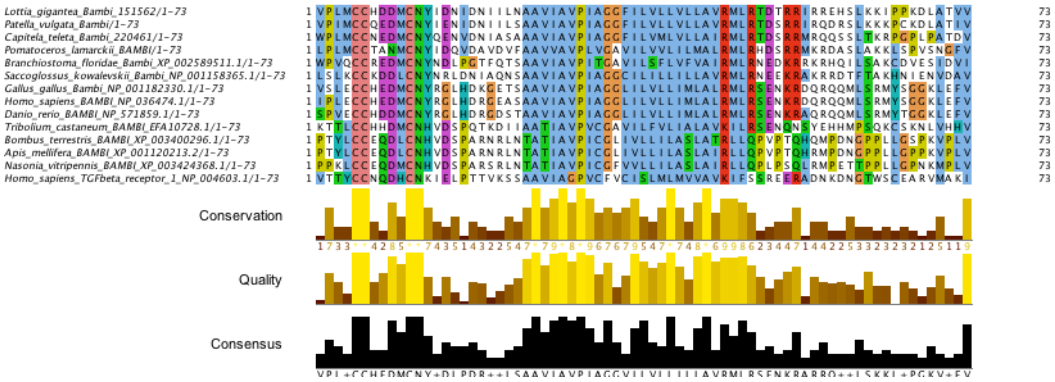


Figure 18: BAMBI *gene alignment used in Fig 5.15. Sequences aligned in MAAFT under the L-INS-i model as described in Methods, and displayed using Jalview with ClustalX colouring.*

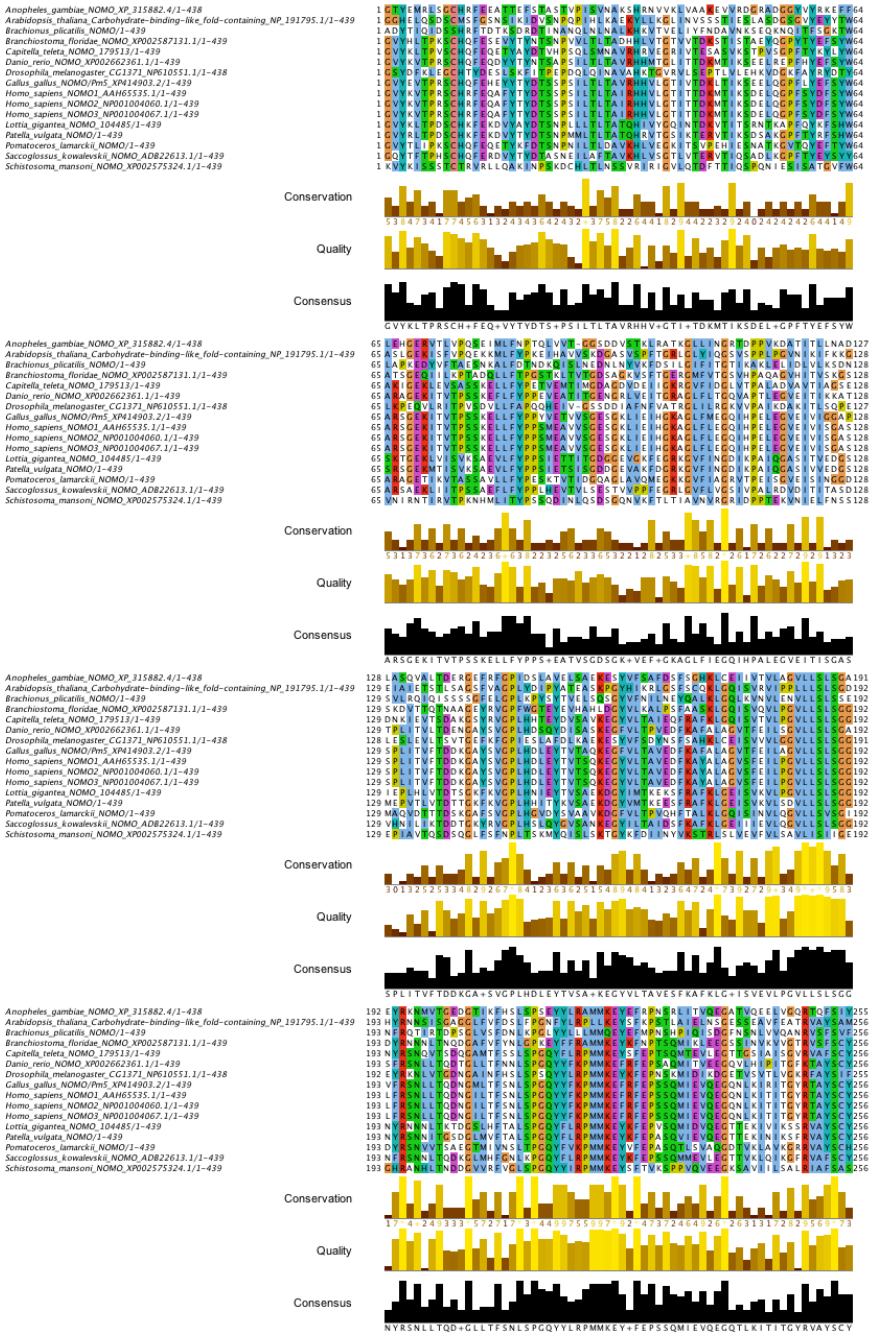


Figure 19: NOMO *gene alignment used in Fig 5.15, Pg. 1. Sequences aligned in MAAFT under the G-INS-i model as described in Methods, and displayed using Jalview with ClustalX colouring.*

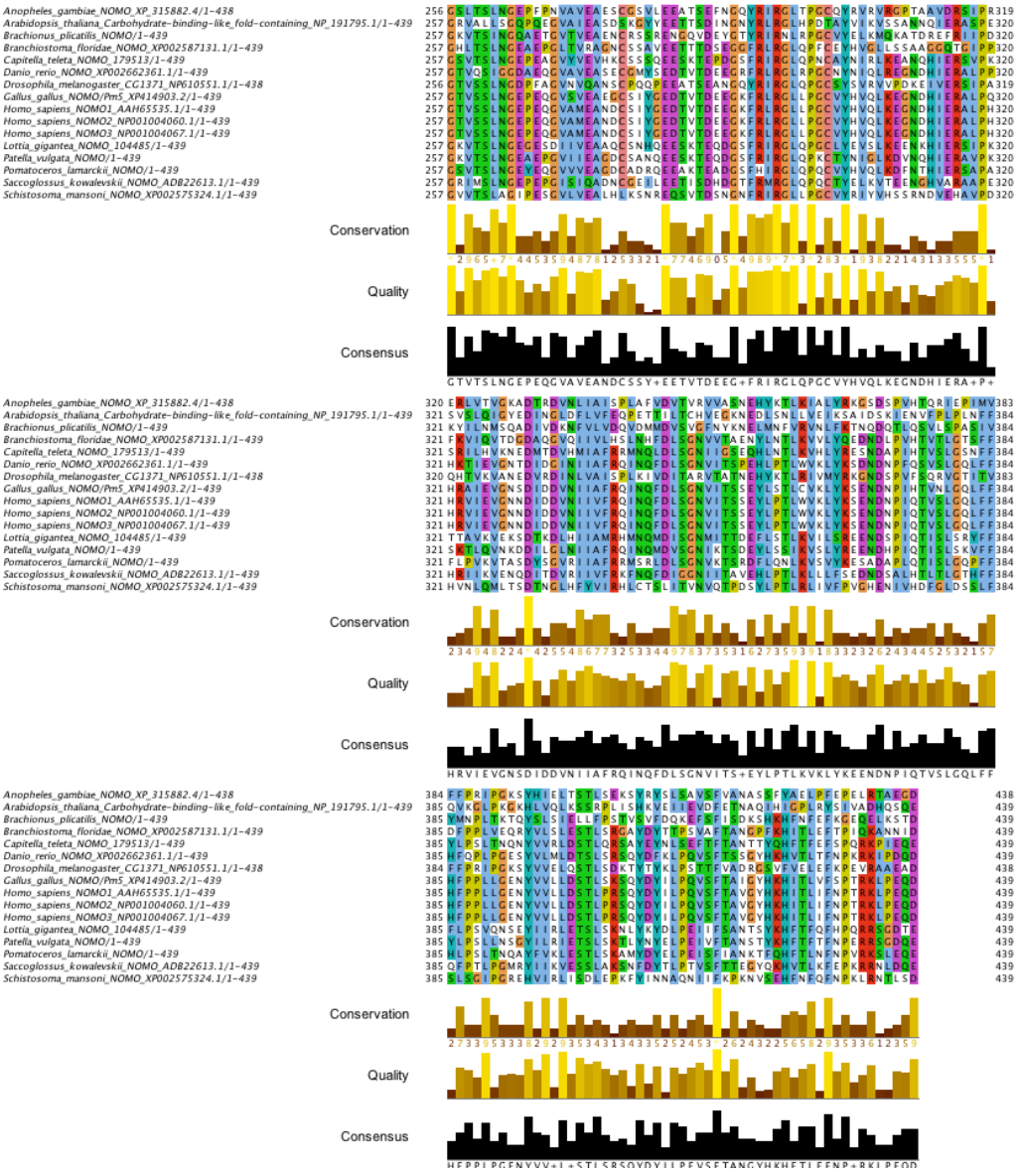


Figure 20: NOMO *gene alignment used in Fig 5.15, Pg. 2, continued from previous page.*

Appendix D: Papers

This appendix contains copies of the two papers published in the course of my doctoral studies. For more information in this regard, as well as a treatment of pending publications and their current status, please refer to Chapter 6

Dev Genes Evol (2012) 222:325–339
DOI 10.1007/s00427-012-0416-6

ORIGINAL ARTICLE

Additive multiple *k*-mer transcriptome of the keelworm *Pomatoceros lamarckii* (Annelida; Serpulidae) reveals annelid trochophore transcription factor cassette

Nathan J Kenny · Sebastian M Shimeld

Received: 20 April 2012 / Accepted: 14 September 2012 / Published online: 9 October 2012

Abstract Recent advances in both next-generation sequencing and assembly programmes have made the low-cost construction of transcriptome datasets for non-model species feasible, capable of yielding a raft of information even from less well-transcribed genes. Here we present the results of assemblies performed on a 51-bp paired end Illumina dataset derived from a mixed larval sample of the annelid *Pomatoceros lamarckii* at 24, 48 and 72 h post-fertilization. We used Oases to assemble 36.5 million paired end reads with *k*-mer sizes from 21 to 29, followed by amalgamation of assemblies, redundancy removal with Vmatch and TGICL and removal of contigs less than 500 bp in length. This resulted in a final assembly of 50,151 contigs, with a mean length of 1,221 bp and covering 61.3 Mbp. A total of 34,846 (69.4 %) of these returned a BlastX hit above a cutoff of $1.0e^{-3}$, and 17,967 (35.8 %) were assigned at least one GO annotation using Blast2GO. We used the assembly to identify genes belonging to the homeobox superclass and the Fox, Sox and Tbx classes, recovering 37, 16, four and three genes, respectively. This included orthologues of genes previously unidentified in lophotrochozoans and protostomes. Our study illustrates the utility of such transcriptomic assembly methods as a gene discovery tool and greatly expands our knowledge of transcription factor genes in annelids in general and in this species in particular.

Communicated by D. Weisblat

Electronic supplementary material The online version of this article (doi:10.1007/s00427-012-0416-6) contains supplementary material, which is available to authorized users.

N. J. Kenny (✉) · S. M. Shimeld (✉)
Department of Zoology, University of Oxford,
Tinbergen Building, South Parks Road,
OX1 3PS, Oxford, UK
e-mail: nathan.kenny@zoo.ox.ac.uk
e-mail: sebastian.shimeld@zoo.ox.ac.uk

Keywords Transcriptome · *Pomatoceros lamarckii* · Annelid · Hox · Sox · Fox · T-box

Abbreviations

bp	Base pair
Fox	Forkhead box
Hox	Homeobox
Sox	Sry-related HMG box

Introduction

Annelids are one of the most speciose phyla within the Metazoa and perform a wide range of ecologically vital roles (Brusca and Brusca 2002). The Annelida, and particularly the polychaetes that make up the major proportion of their diversity, are noted for a Baupläne proposed to be less derived than that of related phyla, such as the molluscs (Brusca and Brusca 2002; Tessmar-Raible et al. 2007), and their well-conserved gene complements (Brusca and Brusca 2002; Tessmar-Raible and Arendt 2003; Denes et al. 2007). To date, however, genomic sequencing efforts within this phylum have been largely confined to the leech *Helobdella robusta* and the polychaetes *Capitella teleta* and *Platynereis dumerilii* (Raible et al. 2005; JGI genome website). *Pomatoceros lamarckii* (Quatrefages) is a serpulid annelid found along the coasts of Northern Europe. It constructs calcareous tubes in which it resides and filter-feeds on plankton using a tentacle crown. Its common English name is keelworm, reflecting its propensity to settle on the underside of ships, and it also commonly fouls cultured shellfish. *P. lamarckii* exhibits typical indirect spiralian development via a trochophore stage (McDougall et al. 2006), and its ease of collection, reliable embryogenesis and placement in annelid phylogeny (Struck et al. 2011) make it a useful non-model organism for a range of investigations. Furthermore,

its equal cleavage (Segrove 1941), unlike that seen in annelid models such as *P. dumerilii* (Fischer et al. 2010), provides a useful outgroup to aid in inference of ancestral traits.

Recent phylogenomic analysis has led to an altered understanding of annelid relationships, with the three traditionally described classes of annelid, the Polychaeta, Oligochaeta and Hirudinea, replaced by two novel groupings, the Errantia and the Sedentaria (Struck et al. 2011). *P. lamarckii* is basal within the Sedentaria, acting as a sister group for comparison to more derived members of this clade (such as *C. teleta* and *H. robusta*). It also retains more primitive characters than some other sedentarian species and is an excellent outgroup for comparison to the Errantia (such as *P. dumerilii*).

Current transcriptome data for *P. lamarckii* are restricted to a small-scale EST screen based on cloned cDNAs, which identified few developmentally relevant genes, such as those encoding transcription factors (Takahashi et al. 2009). Furthermore, until recently, the construction of next-generation transcriptomes for non-model organisms has been limited to the 454 platform due to difficulties in assembling shorter reads from other platforms into contigs (Emrich et al. 2007; Vera et al. 2008; Schuster 2008). However, the advent of longer Illumina platform read lengths, together with paired-end reads and Illumina's high read number, has rendered this platform a potent means for undertaking next-generation sequencing, even when de novo assembly is the only option.

A number of de novo transcriptome builds have now been performed using Illumina data on organisms as diverse as pulmonate snails (Feldmeyer et al. 2011) and whiteflies (Wang et al. 2010), and adaptations of assembly algorithms specifically tailored for transcriptome data, such as Oases (Schulz et al. 2012) and Trinity (Grabherr et al. 2011), have been developed. However, only recently have investigations systematically attempted to ascertain the most efficient means of gaining maximum benefit from transcriptomic or genomic sequencing attempts, and no clear consensus has emerged yet in this regard (Kumar and Blaxter 2010; Feldmeyer et al. 2011; Zhao et al. 2011). It has been suggested that the most powerful potential means of extracting data from reads is through a meta-analysis, where data from multiple assemblies are combined and redundant sequences are removed (Feldmeyer et al. 2011; Martin et al. 2010; Robertson et al. 2010). Along with other methods, an 'additive multiple *k*-mer' method of meta-analysis has been suggested and has been found to be successful at providing longer contig sizes and greater recovery of data (Surget-Groba and Montoya-Burgos 2010; Martin and Wang 2011).

Here we report the acquisition of an Illumina GAII dataset of 36.5 million paired end 51-bp reads from a mixed larval sample of *P. lamarckii* at 24, 48 and 72 h post-fertilization. An additive multiple-*k* assembly of Oases-derived assembly data was performed, combining the power of Oases in retaining splice-form data at a given *k*-mer size, with the benefits of multiple-*k* assembly. We then examined the coverage in this transcriptome of members of the homeodomain-containing superclass and forkhead box (Fox), Sry-related HMG box (Sox) and T-box (Tbx) classes. These groups of transcription factor are well described in the Ecdysozoa and deuterostomes (Zhong et al. 2008; Zhong and Holland 2011; Kaestner et al. 2000; Jager et al. 2006; Shimeld et al. 2010a; Paps et al. 2012), thus facilitating comparison, and perform a variety of roles in patterning and differentiating tissue in developing embryos (Carlsson and Mahlapuu 2002; Papaioannou and Silver 1998; Bowles et al. 2000). Together they represent a major proportion of the larval transcription factor complement and as such allow us to estimate the utility of this method for developmental gene discovery, as well as providing a resource for future experimentation.

Materials and methods

Animal culture and collection

P. lamarckii were gathered from Tinside, Plymouth, UK, and transported to the Department of Zoology, University of Oxford, where they were kept in a recirculation aquarium system at 12 °C. Gametes were collected and fertilized as described previously (Takahashi et al. 2009).

mRNA extraction and sequencing

The *P. lamarckii* embryonic RNA sample used in this analysis was from the same stock as used in an earlier EST screen (Takahashi et al. 2009), however; only RNA from individuals 24 to 72 h post-fertilization (h.p.f.) was utilized. Samples were prepared using an Illumina mRNA-Seq kit incorporating poly-(A) selection and sequenced on a single lane by the High-Throughput Genomics unit at the Wellcome Trust Centre for Human Genetics, Oxford, UK, on an Illumina GAIIx platform. Raw reads (51 bp, paired end) have been uploaded to the SRA (accession number SRA055301).

Transcriptome assembly

The quality of reads derived from paired end sequencing was assessed using the FastQC programme before assembly (Andrews 2011). Transcriptomes were assembled using Velvet (version 1.1.04) (Zerbino and Birney 2008; Zerbino 2010) and Oases (version 0.1.8) (Schulz et al. 2012) on the ORAC server at the Oxford University Supercomputer Facility. An initial Velvet assembly with all default prior settings and a *k*-mer length of 31 was used as a template for

mapping reads using Bowtie 0.12.7 (Langmead et al. 2009) in order to determine average library fragment length and standard deviation for future assembly work. Oases assemblies of *k*-mer lengths of 21, 23, 25, 27 and 29 were used for the final additive multiple-*k* analysis. Settings used for Oases were: -ins_length 320 -ins_length_sd 24.37 -min_trans_lgth 100 -cov_cutoff 3 -min_pair_count 4. Transcriptome assembly metrics were ascertained using a custom Perl script.

Removal of redundancy

Redundancy was removed from concatenated Oases transcriptome builds using Vmatch (version 2.1.6) (Kurtz 2011) with a search length of 100 bp and a similarity cutoff of 100 %. The TGICL clustering tool (Pertea et al. 2003) was used to assess further redundancy or overlap within the dataset, using the CAP3 programme (Huang and Madan 1999), with 100 % similarity cutoffs for identifying and removing redundant contigs. For some subsequent analyses, a 500-bp cutoff was imposed, using a custom Perl script which is available upon request.

Functional annotation and KEGG pathway assignment

Our final dataset was automatically searched for homologues and annotated according to gene ontology (GO—http://www.geneontology.org) terms using Blast2GO 2.5.0 web start against the nr database (Conesa et al. 2005; Gotz et al. 2008). GO term distribution within the *Drosophila melanogaster* genome was downloaded from B2GO-FAR (Gotz et al. 2011) and quantified (as with transcriptome data) using the Combined Graph function of Blast2GO. For assignment to KEGG pathways, we uploaded our final dataset to the KAAS-KEGG Automatic Annotation Server (http://www.genome.jp/tools/kaas/) for processing using the single-directional best hit option, the default (60) BLAST bit score threshold and the hsa, mmu, xla, dre, cin, spu, dme, ame, cel, smm, nve and tad datasets.

Identification of homeobox, Fox, Sox and T-box genes

Homeobox, Fox, Sox and T-box genes were identified by local BlastX search with known gene sequences downloaded from NCBI and HomeoDB (Zhong et al. 2008; Zhong and Holland 2011) against the complete dataset prior to the imposition of a 500-bp cutoff. These included but were not limited to all lophotrochozoan homeodomain-containing Fox, Sox and T-box sequences on the NCBI database, the *Homo sapiens* and *Branchiostoma floridae* HomeoDB dataset, the *Takifugu rubripes* Sox dataset and the *Mus musculus* and *D. melanogaster* Fox and T-box complement.

P. lamarckii orthologues were identified by phylogenetic analysis, with bootstrap values greater than 80 used as direct evidence for orthology or embedding within clades of previously defined well-annotated examples from other species considered as evidence when bootstrap values alone were insufficient. Where possible, confirmation of identity was made using diagnostic domains or residues, especially for homeodomain-containing genes, as can be seen in Table 3 and File 3 of the "Electronic supplementary material". Additional focused phylogenetic analyses were undertaken in cases where initial evidence was uncertain, especially in the categorization of homeodomain-containing genes (data not shown).

Phylogenetic tree construction

Sequences were aligned using MAFFT 6 (http://mafft.cbrc.jp/alignment/server/index.html) (Katoh et al. 2002) using the L-INS-i strategy unless otherwise stated. Hmbox/Hnf alignments were placed into Jalview for visualization and display (Clamp et al. 2004; Waterhouse et al. 2009). Where noted, Gblocks was used to identify regions of conserved homology in alignments for phylogenetic analysis (Castresana 2000). Alignments were then saved and exported to MEGA 5, where maximum likelihood phylogenetic trees were constructed using the Jones–Taylor–Thornton model, 500/1,000 bootstrap replicates as indicated, and all other default prior settings (Tamura et al. 2011).

Results and discussion

Sequencing results and quality control

A total of 36,458,077 paired-end fragments, from a library with an average insert size of 320 bp (standard deviation 24.37 bp), were sequenced, for a total of approximately 73 million 51 bp reads. The mean GC content of the reads (43.33 %) is similar to an earlier value from *P. lamarckii* derived from EST data (42.42 %; Takahashi et al. 2009). The quality of sequencing was assessed via the Illumina quality score, provided through FastQC, which found the median per base sequence quality to be at least 39 throughout all 51 bp (Fig. 1a of the "Electronic supplementary material"), which corresponds to a probable per base call error rate of less than 1 in 10,000. A number of sequences were overrepresented in the first 11 positions of our reads (Fig. 1b of the "Electronic supplementary material"). These do not correspond to adaptors used in sequencing and are likely to reflect bias in Illumina hexamer binding as reported previously (Hansen et al. 2010).

Transcriptome assembly

Oases was run sequentially for *k*-mer sizes from 21 to 29 bp, the assembly statistics for which can be seen in Table 1. In

Table 1 Oases assembly metrics by *k*-mer size. Summary of metrics relating to the Oases assemblies, at *k*-mer size 21–29, used in the construction of the additive multiple-*k* dataset, along with the concatenated result of all *k*-mers joined together before the removal of redundancy

	k-mer size					Multi-*k* before removal of redundancy	Multi-k after removal of redundancy and TGICL clustering	Final multi-k assembly (500 bpmin)
	21	23	25	27	29			
Number of contigs	109,759	102,559	97,789	91,542	84,940	486,589	172,275	50,151
Max contig length (bp)	8,884	9,325	8,307	8,309	8,236	9,325	9,325	9,325
Mean contig length (bp)	448.68	427.5	402.38	374.54	347.08	403.23	467.03	1,221.54
Median contig length (bp)	209	194	183	174	165	185	207	985
N50 contig length (bp)	902	877	830	757	676	822	966	1,419
# contigs in N50	14,952	13,360	12,540	11,679	10,998	63,154	23,320	14,219
# contigs > 1 kb	13,015	11,270	9,770	8,178	6,615	48,848	24,562	24,562
# bases, total	49,246,609	43,843,965	39,348,264	34,286,127	29,481,062	196,206,027	80,457,457	61,261,605
# bases in contigs > 1 kb	22,786,871	19,965,326	17,152,846	14,094,850	11,137,993	85,137,886	42,926,898	42,926,898
GC content, %	41.79	41.86	42.04	42.18	42.33	42.01	42.60	43.87

general, lower *k*-mer sizes resulted in longer contig assemblies, and the rise in total contig number at a lower *k*-mer size is unsurprising as this is correlated with increased efficiency in recovering lowly expressed transcripts (Zerbino and Birney 2008; Surget-Groba and Montoya-Burgos 2010).

After removal of redundancy with Vmatch, the additive multiple-*k* assembly was trimmed to contigs of 500 bp in length and longer for further analysis, although sequences of 100–500 bp were subjected to Blast search to identify any remaining transcription factors. The 500-bp cutoff was chosen as a tradeoff between the loss of data incurred by excluding short contigs versus the increasing computational requirements of subsequent analysis, and the practical consideration that sequences of 500 bp or longer allow more effective expression analysis by in situ hybridization. The full dataset (100 bp and greater) is available from the authors upon request. Figure 1a shows the distribution of contig length in the final additive multiple-*k* assembly, both before (inset) and after trimming to 500 bp and above.

Statistics on final transcriptome assembly can be seen in the last column of Table 1. The GC content of the final assembled transcriptome closely mirrors that of reads at 43.87 %. The number of reads greater than 1 kb in length (24,562) was almost twofold that of the next highest individual *k*-mer size assembly, although little can be inferred from the higher N50 due to trimming to a minimum length of 500 bp.

BlastX was used to compare the contig set against the human proteome. A total of 27,693 contigs had hits to the human proteome with $e<1.0e^{-9}$; however, this value fell to 6,540 hits within the human proteome when multiple best hits from different *P. lamarckii* contigs to the same human target were counted only the first time they occurred. This suggests remaining redundancy in the *P. lamarckii* contig set, likely due to polymorphism present within our dataset, and splice variation reported by the Oases assembly. This could also reflect assembly errors, however genuine redundancy is expected. We specifically selected the Oases assembler as it should maintain multiple splice variants for single genes, and single errors in reads should be excluded from the dataset as a result of the coverage cutoff function, which discards *k*-mers which occur less than a set threshold number of times before performing contig assembly, discarding likely sources of sequencing error.

There was also further variation at the level of single nucleotides (data not shown). These too were expected as the original mRNA sample was derived from multiple individuals from an outbred population, and they likely reflect allelic variants. Sedentary or sessile broadcast marine spawners with planktonic larvae may have very large effective population sizes and maintain high levels of genetic polymorphism, as seen for example with *B. floridae* and

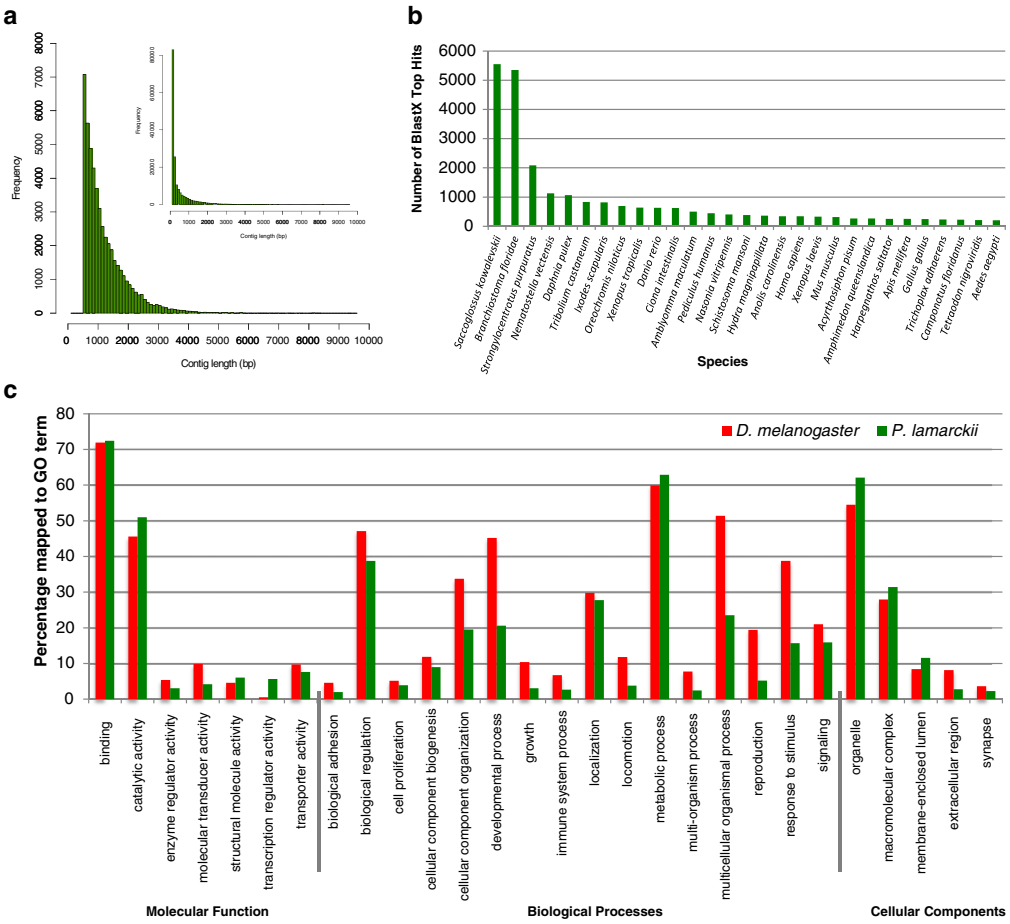


Fig. 1 **a** Contig build statistics. Histograms showing distribution of contig numbers according to length for additive multiple-*k* build. The *inset* shows the distribution of contig lengths for contigs 100 bp and greater in size. The *larger graph* shows the contig length distribution only for those contigs 500 bp and greater in size. **b** Distribution of the species source of the best hit found for each contig by BlastX against the NCBI nr database using Blast2GO (Conesa et al. 2005). **c** GO assignation distribution—*P. lamarckii* transcriptome vs. *D. melanogaster* proteome. *P. lamarckii* transcriptome (*green*) and *D. melanogaster* (*red*) assignation distribution at the second level of GO mapping is presented as the percentage of each dataset that falls within a given term. The *P. lamarckii* dataset was made up of 17,967 contigs assigned GO terms; the *D. melanogaster* dataset (downloaded from B2GO-FAR (Gotz et al. 2011))

Ciona savignyi (Small et al. 2007; Putnam et al. 2008). The multiple variants in our dataset represent true variation in the population rather than individual sequencing errors, as the coverage cutoff utilized (threefold recovery of a *k*-mer sequence), and the expected error rate (less than 1 in 10,000) suggests that the likelihood of the same error being found concurrently at a single site in three *k*-mers and incorporated into contig assembly is very low. Our dataset will thus provide a valuable resource for population genetic studies of *P. lamarckii*.

Multiple-*k* assembly—advantages and disadvantages

The majority of the benefits gained from the multiple-*k* approach came in terms of consistently high contig length rather than in the number of genes identified. For example,

of the 60 transcription factor genes identified in our analyses, only three (*Irx*, *Pou2* and *Lmx*) were completely absent in the 29-mer assembly when assessed by Blastn, with at least fragments of these genes present at all other *k*-mer sizes.

Contig length for a given gene, however, varied considerably between individual *k*-mer assemblies, with a slight trend to best recovery at lower *k*-mer sizes but better contiguity occasionally seen for longer *k*-mers. For example, 29 of the 60 transcription factor genes identified by this study were at their longest in the 21mer assembly, with five of them exhibiting best recovery in the 29mer dataset (data not shown). Contig length recovery also varied markedly between *k*-mers. NK 2.1a, for example, was recovered up to 1,902 bp at 29mer and only 899 bp for the 21mer assembly. Without the benefit of an additive multi-*k* approach, much useful data would therefore have been discarded in the course of assembly.

One caveat must be applied to the multiple-*k* approach. It can quickly and efficiently produce a large amount of contig sequence. However, while we were able to search for specific gene families within our larger (<100 bp) dataset, more global-level analysis tool performance has lagged behind our ability to produce data. We suspected that trimming to 500 bp and above, while facilitating Blast2GO analysis and functional annotation (as can be seen in the next section), removed from consideration a number of potentially interesting lowly expressed genes whose transcripts had not assembled to large contig sizes. We hence subsequently revisited shorter contig data, identifying, for example, that 11 homeodomain genes were initially found within our dataset from contigs less than 500 bp in size. For this reason, we would suggest that when low-expressed gene discovery is the primary goal of transcriptome assembly, smaller contigs are retained for investigation and that cutoffs such as 500 bp used here and commonly seen in the literature be imposed only when computational resources restrict global analyses.

Functional annotation and analysis

Blast2GO v. 2.5.0 (Conesa et al. 2005; Gotz et al. 2008) was used to perform functional annotation, firstly by performing BlastX against the NCBI nr protein database, then GO-mapping and annotation. Of 50,151 contigs, 34,846 (69.4 %) returned a BlastX hit above the cutoff score ($1e^{-3}$). A total of 17,967 (35.8 %) were assigned to a wide variety of GO groups at all levels, the details of which can be found in File 2 of the "Electronic supplementary material".

Top BlastX hit distribution by species can be seen in Fig. 1b. These were distributed among a range of species, with invertebrate deuterostomes being the most common top hits, which may reflect a shared tendency to less-derived gene complements in these species, which are also noted as slow-evolving (Putnam et al. 2008; Arendt et al. 2008). This result may also reflect the present paucity of lophotrochozoan sequences within the nr database.

The distribution of sequence GO assignations into second-level functional categories, divided between the three top-level divisions (biological process, cellular component and molecular function) of GO, can be seen in Fig. 1c. These are presented alongside the results for the proteome of *D. melanogaster*, as downloaded from B2GO-FAR (Gotz et al. 2011), used as a very-well-annotated protostome dataset for comparison.

Differences between the GO annotations of a stage-specific developmental transcriptome and a complete proteome are to be expected and could derive from species differences, from gene expression differences associated with the developmental stage chosen, as well as from the impact of the transcriptome assembly process. In many GO categories, contigs from our transcriptome were slightly underrepresented compared to the *D. melanogaster* sample, which is perhaps the result of limited temporal sampling (three larval stages in our dataset as opposed to the complete *D. melanogaster* proteome). Significant underrepresentation (assessed by Fisher's exact test with multiple testing correction of FDR (Benjamini and Hochberg 1995)) in the GO categories "reproduction", "multi-organism process" and "locomotion" (p values of 0, $1.40e^{-154}$ and $3.00e^{-208}$, respectively) was unsurprising, but underrepresentation in terms such as "growth" "developmental process" "cellular component organization" and "cell proliferation" (p values $8.10e^{-223}$, 0, 0 and $7.80e^{-21}$) was not expected and may be a result of genes assigned to these areas being weakly transcribed or being shorter than 500 bp when assembled and hence excluded in this analysis.

The most notable overrepresentation in the *P. lamarckii* dataset is in the proportion of genes with the GO term "transcription regulator activity" (p value $3.20e^{-04}$). This assignation is eleven times more common in our dataset than in the *D. melanogaster* proteome as a whole and may indicate that a large amount of differentially activated transcriptional activity is occurring in larvae at these stages of their development—a finding that would concur with what we understand of their biology. Other second-level GO terms also overrepresented in our dataset relative to the *D. melanogaster* proteome include the "structural molecule activity" and "catalytic activity" (p values $9.50e^{-7}$ and $5.30e^{-05}$), perhaps an indication of the processes of differentiation occurring within the developing larvae.

KEGG pathway analysis

Using the KAAS-KEGG Automatic Annotation Server (Moriya et al. 2007), we mapped our contigs to functional

Table 2 Coverage of a range of key conserved metabolic pathways within transcriptome. Coverage of a selection of key conserved metabolic and cell signalling pathways as determined by the KAAS-KEGG Automatic Annotation Server. Full pathways can be seen in Fig. 2 of the "Electronic supplementary material"

KO pathway ID	KO pathway name	Number of KEGG enzyme genes in pathway[a]	KEGG-KASS mapped *P. lamarckii* homologues[b]	Pathway reactions covered[c] (%)
00010	Glycolysis/gluconeogenesis	45	26	73
00020	Citrate cycle TCA cycle	22	22	85
00071	Fatty acid metabolism	37	24	79
00230	Purine metabolism	106	72	64
00240	Pyrimidine metabolism	63	52	70
04310	Wnt signalling pathway	97[d]	47	58
04330	Notch signalling pathway	28[d]	18	66

[a] From www.genome.jp/dbget-bin/get_linkdb?pathway + map00010 - final number (00010) represents KO pathway ID

[b] From automatic Pathway Mapping step

[c] Percentage of all KEGG REACTION or KEGG Map (Wnt and Notch) interactions mediated by at least one mapped *P. lamarckii* gene

[d] Total gene number used when enzyme number was not suitable

groups. Contigs from our transcriptome were assigned to 278 out of the 329 separate Kegg orthology reference hierarchy pathways, comprising 109 metabolic pathways (out of 152 paths catalogued on the KO server), 22 genetic information processing pathways (from 23), 20 environmental information processing pathways (from 22), 17 cellular process pathways (out of 19), 57 organismal signal pathways (from 57) and 53 disease pathways (out of 56). Full KO assignment is available from the authors on request. The majority of the pathways not represented in our dataset are those found in plants or prokaryotes only. Excellent coverage was found for most key conserved pathways as can be seen in Table 2 and in Fig. 2 of the "Electronic supplementary material", which shows the representation of our transcripts in these six key pathways.

Homeobox-containing contigs

Homeobox genes perform a variety of crucial roles in developmental patterning across the Metazoa, and members of this superclass are extremely well described in several deuterostome and ecdysozoan species (Zhong et al. 2008; Zhong and Holland 2011; Hui et al. 2012). From our transcriptome dataset, we were able to identify 28 contigs whose sequence entirely spanned the homeodomain and a further nine possessing sufficient sequence for satisfactory phylogenetic analysis. This represents a considerable increase on the four *P. lamarckii* homeodomain genes identified in previous work (Takahashi et al. 2009; McDougall et al. 2011).

To identify these genes, we conducted molecular phylogenetic analyses using the entire homeodomain complements of *H. sapiens* and *T. castaneum* as well-characterized and annotated gene sets, and the resultant phylogenetic trees upon which our assignation is based can be seen in Fig. 3 of the

Table 3 *P. lamarckii* homeodomain containing genes. Homologues of homeodomain containing proteins identified within our transcriptome dataset. Italicized entries in columns represent subclass names. Phylogenetic trees showing the position of these genes relative to *H. sapiens*, *T. castaneum* and annelid homeodomain containing proteins can be seen in Fig. 3 of the "Electronic supplementary material", and further details relating to their assignation, along with sequences, can be found in File 3 of the "Electronic supplementary material"

Whole homeodomain recovered		Partial homeodomain
ANTP/HoxL	Prd	Prd
Cad[a]	Dmbx[a]	Msx[b]
Mnx[b]	Gsc[a]	Nk 6 b[b]
Evx[a]	Hbn[a]	Abox-like[b]
ANTP/NKL	Otp[a]	Cut
Bsx[c]	Otx[a]	Cmp HD1[a]
Dlx a[c]	Paired-like[a]	Tale
Hhex[c]	Pitx[a]	Pknox[a]
NK 2.1 a[a]	Vsx[a]	Tgif[a]
Nk 2.1 b[b]	Lim	Pou
Nk 2.2[c]	Lhx 1/5[a]	Pou2[b]
Cut	Lhx 2/9[b]	Pou 3[a]
Cmp HD2[a]	Lhx 6/8[b]	Lim
Cux[a]	Ist/Tup[a]	Lmx[a]
Onecut[a]	Six	
Tale	Six 3/6[a]	
Irx[a]	Hnf	
Pbx[a]	Hmbox-like[a]	

[a] Identity confirmed by diagnostic conserved domain or specific amino acid signature

[b] Identity unable to be confirmed as transcriptome sequence does not cover area

[c] Diagnostic conserved domain or specific amino acid signature not present

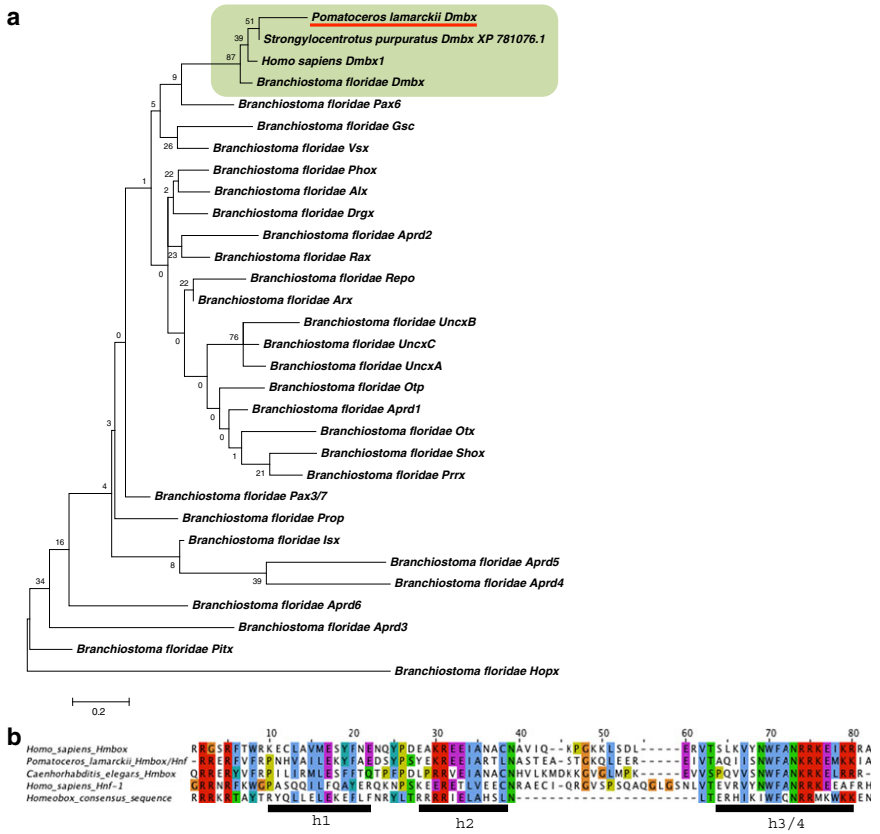


Fig. 2 Confirmation of the identity of Dmbx and Hmbox/Hnf homeodomain proteins. **a** Unrooted maximum likelihood phylogeny constructed from *B. floridae* paired genes and *H. sapiens* Dmbx data as downloaded from HomeoDB (Zhong et al. 2008; Zhong and Holland 2011) along with *S. purpuratus* Dmbx sequence and *P. lamarckii* homologue identified in the current study. Alignment was curated using Gblocks, resulting in a final 53-aa sequence (File 3 of the "Electronic supplementary material") used for analysis in Mega5 using the Jones–Taylor–Thornton model with all other default settings. Dmbx sequences are *boxed in green*; the sequence from *P. lamarckii* is *underlined in red*. Bootstrap proportions (from 1,000 replicates) are found at the base of each node. The *scale bar* represents 0.2 substitutions per amino acid at the given scale along branches on the tree. **b** Hnf family homeodomains, aligned in MAFFT and visualized in Jalview. Hnf sequences (with the exception of the *P. lamarckii* sequence, derived from the present study) obtained from HomeoDB. Consensus homeodomain sequence downloaded from Pfam (Finn et al. 2010) with helix positions as previously noted (Gehring 1992)

"Electronic supplementary material". These genes can be seen in Table 3, along with the results of attempts to confirm identity using diagnostic amino acid sequences and motifs.

We also noted the presence of a number of other contigs within our dataset whose sequence encoded a partial homeodomain which was insufficient for firm annotation. These can be seen putatively identified in File 3 of the "Electronic supplementary material", along with all the sequences used in phylogenetic analysis, which were drawn from HomeoDB (Zhong et al. 2008; Zhong and Holland 2011) along with previously published annelid sequences from Genbank.

We note a paucity of HoxL (Hox-like or Hox-linked) class genes within our dataset and suspect that this is a result of the timings of samples taken to create our RNA pool. Other subclasses, particularly the NKL (NK-like or NK-linked) and PRD (Paired) classes, are well represented. Several homeobox genes not previously sequenced and described in lophotrochozoan species were found in the course of our analysis, with homologues of the *Dmbx* and Hnf (*Hmbox*) families of particular interest.

Dmbx has, to date, only been described in deuterostomes and is possibly related to the cnidarian *Manacle* gene

(Takahashi and Holland 2004). In chordates, it is believed that *Dmbx* plays an important role in establishing the midbrain/hindbrain boundary, and the role of this gene in the Annelida would be a good candidate for future investigation. To confirm the identity of *Dmbx*, a focused molecular phylogenetic analysis was performed on PRD class genes, the results of which can be seen in Fig. 2a.

The transcriptional repressor *Hmbox* was traditionally believed to be chordate specific (Takatori et al. 2008) and has only recently been noted in an ecdysozoan (Lesch and Bargmann 2010). *Hmbox* identity is confirmed here by examination of clearly conserved features of this class, including an extended homeodomain courtesy of additional amino acids between the h2 and h3 domains (Fig. 2b), underlining the conservation of this gene in the Bilateria. The existence of *Cmp* homologue sequence supports the hypothesis that this gene is shared ancestrally within the Bilateria, with *Satb* (which is found only in vertebrates but appears similar to *Cmp*) likely representing a divergent vertebrate-specific form of this gene (Burglin and Cassata 2002).

Some features of homeobox genes generally observed in ecdysozoan and deuterostome datasets do not seem as prevalent in the sequences obtained by this transcriptomic analysis. Tinman (Tn) domains, which are generally observed in NKL family homeobox genes, were not observed in a homologous position in many of the *P. lamarckii* NKL genes identified. These may lie in regions as yet unsequenced, but their presence could not be used as a diagnostic feature to confirm the identity of a portion of our data. This is also the case for the octapeptide domains generally found in PRD-class homeobox genes, although the latter have other characteristic regions used to confirm identity.

The recovery of sequence for *Dlxa* and *Dlxb*, identical to that previously identified (McDougall et al. 2011), gives confidence in the validity of our approach (File 3 of the "Electronic supplementary material"). We also recovered a sequence similar to the *Pax beta* sequences first identified in the leech *H. robusta* (Schmerer et al. 2009). However, this does not span the homeodomain and is thus not incorporated into our phylogeny, but it can be seen in detail in File 3 of the "Electronic supplementary material".

Fox, Sox and T-box genes

Fox proteins are members of the helix-turn-helix group of proteins, with an 80- to 100-amino-acid "forkhead box" motif that defines the class. They are separated into 23 families (*FoxA–FoxS*) and perform a variety of roles in the specification of tissues during embryonic development, as well as in regulating a number of metabolic functions. Sox genes belong to the HMG box superclass and regulate many diverse elements of development, growth and differentiation. They are found throughout the Metazoa and their phylogeny is generally well understood. The T-box genes are less numerous than those families detailed earlier but nonetheless are vital for a range of developmental roles. They are defined by the presence of a roughly 200-amino-acid-long region known as the T-box domain, first identified in *Brachyury*.

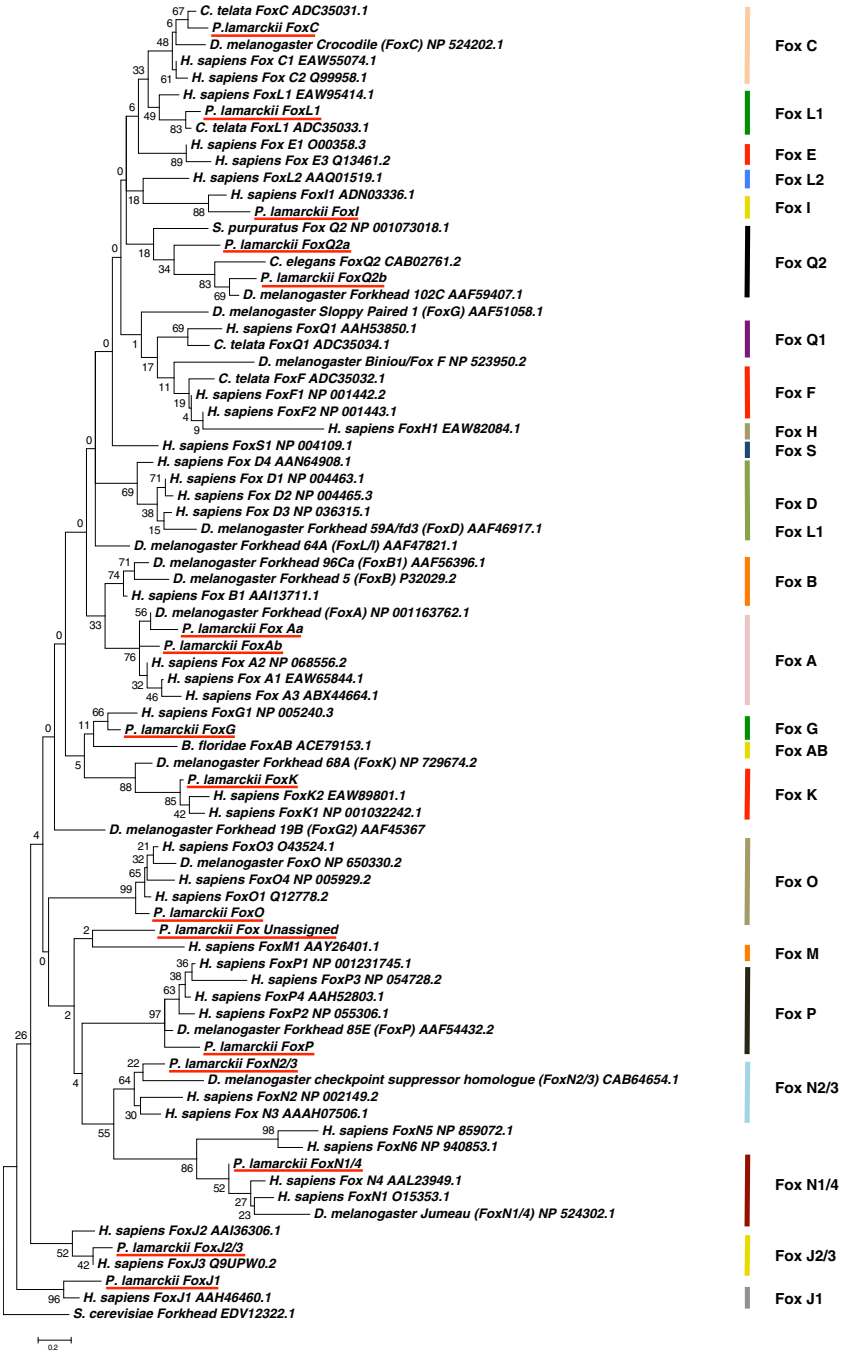
Only two genes from these classes had previously been observed in *P. lamarckii* (FoxN1/4 and SoxC) (Takahashi et al. 2009). The names of the genes firmly identified within our dataset can be seen in Table 4, while full sequences are available in File 3 of the "Electronic supplementary material".

Fox genes

The Fox gene complement is particularly well represented in our dataset and can be seen in Fig. 3, alongside the Fox genes of *H. sapiens*, *D. melanogaster* and those known from *C. teleta* (Shimeld et al. 2010a). We note the presence of *FoxI*, *FoxJ1*, *FoxJ2*, *FoxN1/4*, *FoxN2/3*, *FoxO* and *FoxQ2* transcripts, which are previously uncatalogued in the Lophotrochozoa. The presence of two *FoxA* and FoxQ2 sequences may represent recent duplications within these families in this species, while *FoxB1*, *D*, *H*, *L2* and *M* genes are absent from our analysis and from annelid datasets as represented on Genbank.

Table 4 *P. lamarckii* Fox, Sox and T-box genes identified in transcriptome homologues of genes of the forkhead box, SRY-related HMG box and T-box protein families identified within our transcriptome dataset. Figs. 3, 4 and 5 show their phylogenetic position relative to other genes of known identity and full sequence, along with other sequences used to determine phylogenetic relationships, can be found in File 3 of the "Electronic supplementary material"

Transcription factor		
Fox family	T-box family	Sox family
FoxAa	T-Brain	Sox B1
FoxAb	Brachyury	SoxB2
FoxC	Tbx 2/3	Sox C
FoxG		Sox D
FoxI		
FoxJ1		
FoxJ2/3		
FoxK		
FoxL1		
FoxN1/4		
FoxN2/3		
FoxO		
FoxP		
FoxQ2a		
FoxQ2b		
Fox (unknown family)		


C. telata FoxC ADC35031.1
P.lamarckii FoxC
D. melanogaster Crocodile (FoxC) NP 524202.1
H. sapiens Fox C1 EAW55074.1
H. sapiens Fox C2 Q99958.1
H. sapiens FoxL1 EAW95414.1
P. lamarckii FoxL1
C. telata FoxL1 ADC35033.1
H. sapiens Fox E1 O00358.3
H. sapiens Fox E3 Q13461.2
H. sapiens FoxL2 AAQ01519.1
H. sapiens FoxI1 ADN03336.1
P. lamarckii FoxI
S. purpuratus Fox Q2 NP 001073018.1
P. lamarckii FoxQ2a
C. elegans FoxQ2 CAB02761.2
P. lamarckii FoxQ2b
D. melanogaster Forkhead 102C AAF59407.1
D. melanogaster Sloppy Paired 1 (FoxG) AAF51058.1
H. sapiens FoxQ1 AAH53850.1
C. telata FoxQ1 ADC35034.1
D. melanogaster Biniou/Fox F NP 523950.2
C. telata FoxF ADC35032.1
H. sapiens FoxF1 NP 001442.2
H. sapiens FoxF2 NP 001443.1
H. sapiens FoxH1 EAW82084.1
H. sapiens FoxS1 NP 004109.1
H. sapiens Fox D4 AAN64908.1
H. sapiens Fox D1 NP 004463.1
H. sapiens Fox D2 NP 004465.3
H. sapiens Fox D3 NP 036315.1
D. melanogaster Forkhead 59A/fd3 (FoxD) AAF46917.1
D. melanogaster Forkhead 64A (FoxL/l) AAF47821.1
D. melanogaster Forkhead 96Ca (FoxB1) AAF56396.1
D. melanogaster Forkhead 5 (FoxB) P32029.2
H. sapiens Fox B1 AAI13711.1
D. melanogaster Forkhead (FoxA) NP 001163762.1
P. lamarckii Fox Aa
P. lamarckii FoxAb
H. sapiens Fox A2 NP 068556.2
H. sapiens Fox A1 EAW65844.1
H. sapiens Fox A3 ABX44664.1
H. sapiens FoxG1 NP 005240.3
P. lamarckii FoxG
B. floridae FoxAB ACE79153.1
D. melanogaster Forkhead 68A (FoxK) NP 729674.2
P. lamarckii FoxK
H. sapiens FoxK2 EAW89801.1
H. sapiens FoxK1 NP 001032242.1
D. melanogaster Forkhead 19B (FoxG2) AAF45367
H. sapiens FoxO3 O43524.1
D. melanogaster FoxO NP 650330.2
H. sapiens FoxO4 NP 005929.2
H. sapiens FoxO1 Q12778.2
P. lamarckii FoxO
P. lamarckii Fox Unassigned
H. sapiens FoxM1 AAY26401.1
H. sapiens FoxP1 NP 001231745.1
H. sapiens FoxP3 NP 054728.2
H. sapiens FoxP4 AAH52803.1
H. sapiens FoxP2 NP 055306.1
D. melanogaster Forkhead 85E (FoxP) AAF54432.2
P. lamarckii FoxP
P. lamarckii FoxN2/3
D. melanogaster checkpoint suppressor homologue (FoxN2/3) CAB64654.1
H. sapiens FoxN2 NP 002149.2
H. sapiens Fox N3 AAAH07506.1
H. sapiens FoxN5 NP 859072.1
H. sapiens FoxN6 NP 940853.1
P. lamarckii FoxN1/4
H. sapiens Fox N4 AAL23949.1
H. sapiens FoxN1 O15353.1
D. melanogaster Jumeau (FoxN1/4) NP 524302.1
H. sapiens FoxJ2 AAI36306.1
P. lamarckii FoxJ2/3
H. sapiens FoxJ3 Q9UPW0.2
P. lamarckii FoxJ1
H. sapiens FoxJ1 AAH46460.1
S. cerevisiae Forkhead EDV12322.1
0.2
Fox C
Fox L1
Fox E
Fox L2
Fox I
Fox Q2
Fox Q1
Fox F
Fox H
Fox S
Fox D
Fox L1
Fox B
Fox A
Fox G
Fox AB
Fox K
Fox O
Fox M
Fox P
Fox N2/3
Fox N1/4
Fox J2/3
Fox J1

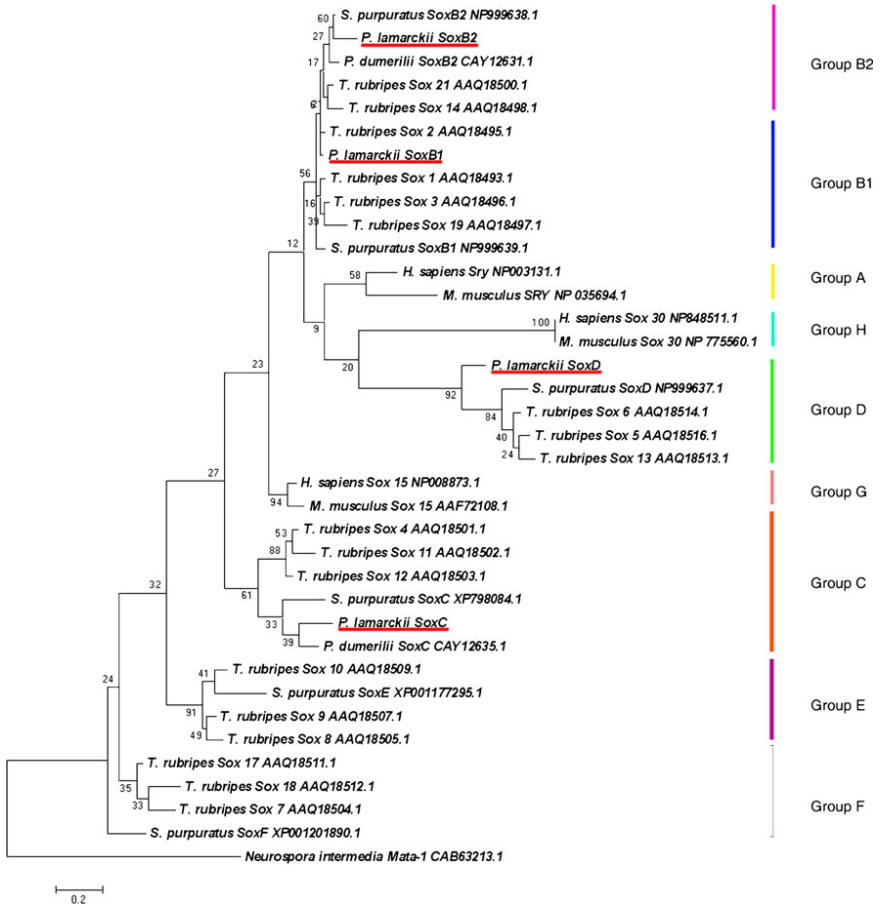


Fig. 4 Molecular phylogenetic tree of the Sox gene class. Phylogeny of *P. lamarckii*, *T. rubripes*, *P. dumerilii* and *S. purpuratus* Sox genes, computed from area surrounding the HMG domain (73-amino-acid alignment). All annotated *P. lamarckii* Sox genes are found in well-supported monophyletic groups with genes of known annotation, with the exception of the SoxB1 homologue. This maximum likelihood tree was constructed utilizing Mega5 with the Jones–Taylor–Thornton model and all other default priors, along with 500 bootstrap replicates, the proportions of which can be seen at the base of each node. *P. lamarckii* Sox genes are *underlined in red*. File 3 of the "Electronic supplementary material" contains data on all sequences used in analysis. Tree is rooted with *Neurospora intermedia* Mata-1 (CAB63213.1). *Scale bar* represents 0.2 substitutions per amino acid at the given branch length. Previously defined Sox gene families (Jager et al. 2006) are shown on the *right* of the figure

◀ **Fig. 3** Molecular phylogenetic tree of the Fox gene class. Maximum likelihood Fox gene phylogeny computed from aligned forkhead box domains (45 aa in length after gap removal) using Mega5 under the Jones–Taylor–Thornton model with all other default settings. Sequences from *P. lamarckii* are *underlined in red*. Bootstrap proportions (from 500 replicates) are found at the foot of each node. Sequences can be found in File 3 of the "Electronic supplementary material". Tree rooted with *Saccharomyces cerevisiae* forkhead (EDV12322.1). The *scale bar* represents 0.2 substitutions per amino acid at the given distance along branches on tree. Previously described Fox families, as defined by Shimeld et al. (2010b), are shown on the *right* of the figure

Sox genes

Our search for *Sox* genes uncovered four of the six known invertebrate Sox families, as can be seen in Fig. 4, using the well-understood *T. rubripes* Sox genes (Koopman et al. 2004), and those of *P. dumerilii* (Kerner et al. 2009) and the purple sea urchin, *Strongylocentrotus purpuratus* (Sodergren et al. 2006), to build the phylogeny. The chordate-specific families (groups A, G and H) were not

present, and our transcriptome has therefore uncovered two thirds of the expected *Sox* complement, with only groups E and F absent from our dataset. A monophyletic *SoxB1* group could not be recovered. Confirming our assignment, however, the *SoxB1* sequence can be recognized as the *SoxB1* orthologue by Blast and possesses the characteristic arginine amino acid residue at position 2 and threonine at position 78, while the *SoxB2* sequence possesses the diagnostic proline at this latter position.

T-Box genes

Three *T-box* class genes were uncovered, of the approximately seven/eight broad families of *T-box* genes identified by Papaioannou and Silver (1998), and can be seen in Fig. 5, alongside the *D. melanogaster* and *M. musculus* complements, and those identified in a number of Cnidarian and Poriferan species. The presence of *Brachyury* was expected, given its conserved role in the Bilateria, including other lophotrochozoans (Arendt et al. 2001; Lartillot et al. 2002). The expression of *Tbx2/3* has also been previously examined in the Annelida, indicating a role in nervous system differentiation (Winchell et al. 2010).

An orthologue of *T-brain*, which has been described extensively only in deuterostomes, was also found. It is represented outside the Deuterostoma in the NCBI database by a single sequence (*Hydroides elegans*, accession number ACA48210) (Arenas-Mena 2008). The expression of this gene in the hemichordate apical organ and in the vertebrate forebrain makes it an interesting target for future consideration with regards to conservation of expression and role (Tagawa et al. 2000).

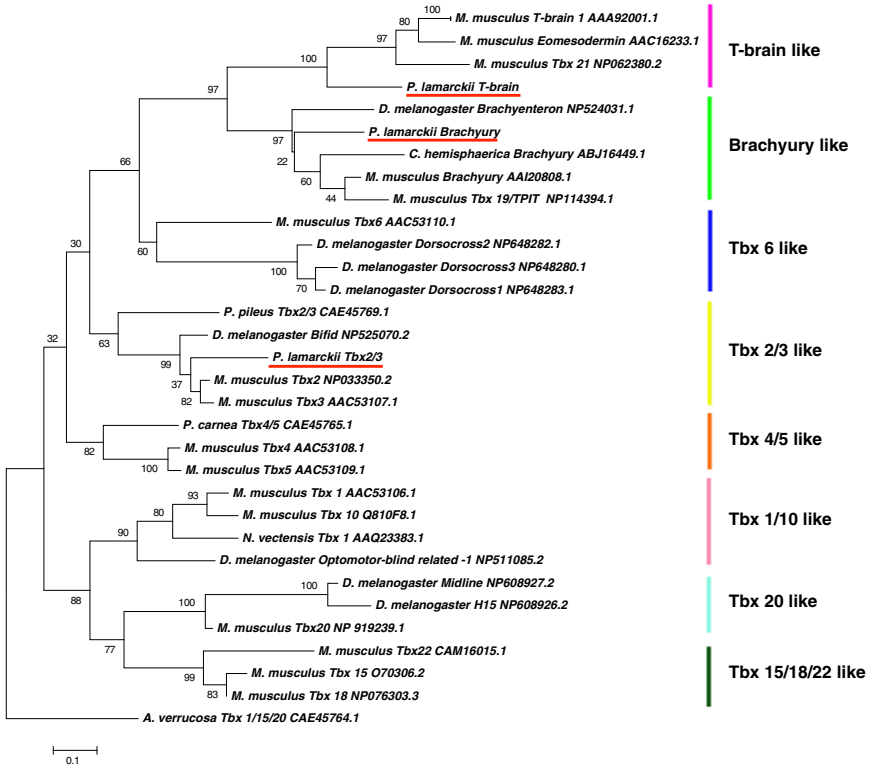


Fig. 5 Molecular phylogeny of the T-box gene class. Maximum likelihood phylogeny of T-box genes from *M. musculus*, *D. melanogaster*, *P. lamarckii* and a number of Cnidarian and Poriferan species derived from T-domain alignment. *P. lamarckii* sequences are *underlined in red*. Phylogeny computed in Mega5 using the Jones–Taylor–Thornton model and all other default priors. All sequences can be found in File 3 of the "Electronic supplementary material". Tree rooted with *Axinella verrucosa* Tbx1/15/20 protein (CAE45764.1). The *scale bar* represents 0.1 substitutions per amino acid at the given distance along branches. Previously defined T-box gene families (Papaioannou and Silver 1998) are shown on the *right* of the figure

Assessing the representation of the full transcriptome of *P. lamarckii* in our dataset

No firm conclusion can be drawn as to the coverage provided by this transcriptome data until the advent of a complete *P. lamarckii* genome as it can be difficult to estimate the coverage provided by transcriptomic studies due to variation in gene expression levels and temporal expression profiles. However, by determining the proportion of genes found in comparison with known 'housekeeping' datasets, we can gain an understanding of how representative our transcriptome is. KEGG mapping (Moriya et al. 2007), for example, has revealed approximately 70 % coverage across a range of key conserved metabolic pathways, with coverage in some cases being higher, particularly as many KEGG pathways contain genes found only in a subset of species.

Of the transcription factor families examined, approximately 60 % of the expected complement can be reliably identified within our dataset based on comparison of DNA binding domains, while additional more ambiguous assignments can be made for those transcription factors whose sequence was recovered away from these. The assignation by Blast2GO of our dataset into 34,846 BlastX hits, split into 17,967 GO categories at all GO levels, also suggests that we have uncovered a substantial proportion of the expressed transcriptome.

We therefore suspect that at least 60–70 % of the 24–72-hpf trochophore transcript complement is represented in our additive multiple-*k* assembly, providing a resource of considerable utility for a range of further investigations.

Conclusions

Using the Oases assembler in building an additive multiple-*k* assembly uncovered the sequence of a range of developmentally important genes. The additive multiple-*k* assembly technique also holds much promise. It increased the number of identifiable contigs, as well as their average length, relative to individual *k*-mer size builds, although the amount of data produced proved to be more difficult to analyse using Blast2GO, necessitating a higher-size cutoff value (500 bp).

Previous studies of the *P. lamarckii* gene complement have been limited to a small conventional EST screen (Takahashi et al. 2009). While this paper described 2,308 EST clusters, 1,103 with significant matches to the nr database, our dataset described 50,151 sequences, 34,846 of which had significant BlastX hits to the nr database, split into 17,967 GO categories at all levels of categorization. This represents a marked increase in the data available in this species in particular and the Annelida as a whole.

The large number of transcription factors discovered presents many opportunities for future studies and represents a significant increase on the extant annelid dataset. The identification of annelid homologues of numerous homeobox genes, including *Dmbx* and a HNF/*Hmbox* representative, presents opportunities for considering the conservation of these genes in the Lophotrochozoa and provides opportunities to test their ancestral developmental role. The BlastX top hit results found through Blast2GO analysis, which most closely match species noted as slow-evolving, support the hypothesis that polychaete annelids are slow-evolving compared to other taxa, as shown previously in this species (Takahashi et al. 2009).

The de novo construction of transcriptomes using Illumina data from non-model organisms is therefore a practical enterprise, in this case yielding a large annelid transcription factor dataset and adding much to our knowledge of these genes in this phylum.

Acknowledgments We thank the members of the Shimeld and Holland groups for their help and support in preparing this manuscript and two anonymous reviewers for their comments and suggestions. Sequencing was performed by the High-Throughput Genomics unit at the Wellcome Trust Centre for Human Genetics, Oxford. Supercomputing support was provided by the Oxford Supercomputing Center (http://www.oerc.ox.ac.uk/). We thank the Elizabeth Hannah Jenkinson Fund, which funded the sequencing, and the Clarendon Fund, which supported NJK in the course of this project.

References

Andrews S (2011) FastQC—a quality control tool for high throughput sequence data. Babraham Bioinformatics. http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/

Arenas-Mena C (2008) The transcription factors *HeBlimp* and *HeT-brain* of an indirectly developing polychaete suggest ancestral endodermal, gastrulation, and sensory cell-type specification roles. J Exp Zool B 310B(7):567–576

Arendt D, Technau U, Wittbrodt J (2001) Evolution of the bilaterian larval foregut. Nature 409:81–85

Arendt D, Denes AS, Jekely G, Tessmar-Raible K (2008) The evolution of nervous system centralization. Philos T Roy Soc B 363 (1496):1523–1528

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc Ser B (Methodol) 57(1):289–300

Bowles J, Schepers G, Koopman P (2000) Phylogeny of the Sox family of developmental transcription factors based on sequence and structural indicators. Dev Biol 227(2):239–255

Brusca R, Brusca G (2002) Invertebrates, 2nd edn. Sinauer, Sunderland

Burglin TR, Cassata G (2002) Loss and gain of domains during evolution of *cut* superclass homeobox genes. Int J Dev Biol 46(1):115–123

Carlsson P, Mahlapuu M (2002) Forkhead transcription factors: key players in development and metabolism. Dev Biol 250(1):1–23

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–552

Clamp M, Cuff J, Searle SM, Barton GJ (2004) The Jalview Java alignment editor. Bioinformatics 20(3):426–427

Conesa A, Gotz S, Garcia-Gomez J, Terol J, Talon M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18):3674–3676

Denes A, Jekely G, Steinmetz P, Raible F, Snyman H, Prud'homme B, Ferrier D, Balavoine G, Arendt D (2007) Molecular architecture of annelid nerve cord supports common origin of nervous system centralization in Bilateria. Cell 129:277–288

Emrich S, Barbazuk W, Li L, Schnable P (2007) Gene discovery and annotation using LCM-454 transcriptome sequencing. Genome Res 17(1):69–73

Feldmeyer B, Wheat C, Krezdorn N, Rotter B, Pfenninger M (2011) Short read Illumina data for the de novo assembly of a non-model snail species transcriptome (*Radix balthica*, Basommatophora, Pulmonata), and a comparison of assembler performance. BMC Genomics 12(1):317

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer ELL, Eddy SR, Bateman A (2010) The Pfam protein families database. Nucleic Acids Res 38(Suppl 1):D211–D222

Fischer A, Henrich T, Arendt D (2010) The normal development of *Platynereis dumerilii* (Nereididae, Annelida). Front Zool 7(1):31

Gehring WJ (1992) The homeobox in perspective. Trends Biochem Sci 17(8):277–280

Gotz S, Garcia-Gomez J, Terol J, Williams T, Nagaraj S, Nueda M, Robles M, Talon M, Dopazo J, Conesa A (2008) High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res 36:3420–3435

Gotz S, Arnold R, Sebastian-Leon P, Martin-Rodriguez S, Tischler P, Jehl M, Dopazo J, Rattei T, Conesa A (2011) B2G-FAR, a species-centered GO annotation repository. Bioinformatics 27 (7):919–924

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotech 29(7):644–652

Hansen KD, Brenner SE, Dudoit S (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res 38(12):e131

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9(9):868–877

Hui JHL, McDougall C, Monteiro AS, Holland PWH, Arendt D, Balavoine G, Ferrier DEK (2012) Extensive chordate and annelid macrosynteny reveals ancestral homeobox gene organization. Mol Biol Evol 29:157–165

Jager M, Queinnec E, Houliston E, Manuel M (2006) Expansion of the Sox gene family predated the emergence of the Bilateria. Mol Phylogenet Evol 39(2):468–477

JGI genome website http://genome.jgi-psf.org/

Kaestner KH, Knochel W, Martinez DE (2000) Unified nomenclature for the winged helix/forkhead transcription factors. Gene Dev 14 (2):142–146

Katoh K, Misawa K, Kuma KÄ, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30(14):3059–3066

Kerner P, Simionato E, Le Gouar M, Vervoort M (2009) Orthologs of key vertebrate neural genes are expressed during neurogenesis in the annelid *Platynereis dumerilii*. Evol Dev 11(5):513–524

Koopman P, Schepers G, Brenner S, Venkatesh B (2004) Origin and diversity of the Sox transcription factor gene family: genome-wide analysis in *Fugu rubripes*. Gene 328:177–186

Kumar S, Blaxter M (2010) Comparing de novo assemblers for 454 transcriptome data. BMC Genomics 11:571

Kurtz S (2011) The Vmatch large scale sequence analysis software. http://www.vmatch.de/

Langmead B, Trapnell C, Pop M, Salzberg S (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3):R25

Lartillot N, Lespinet O, Vervoort M, Adoutte A (2002) Expression pattern of *Brachyury* in the mollusc *Patella vulgata* suggests a conserved role in the establishments of the AP axis in Bilateria. Development 129(6):1411–1421

Lesch BJ, Bargmann CI (2010) The homeodomain protein *hmbx-1* maintains asymmetric gene expression in adult *C. elegans* olfactory neurons. Genes Dev 24(16):1802–1815

Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nat Rev Genet 12(10):671–682

Martin J, Bruno V, Fang Z, Meng X, Blow M, Zhang T, Sherlock G, Snyder M, Wang Z (2010) Rnnotator: an automated *de novo* transcriptome assembly pipeline from stranded RNA-Seq reads. BMC Genomics 11(1):663

McDougall C, Chen W-C, Shimeld S, Ferrier D (2006) The development of the larval nervous system, musculature and ciliary bands of *Pomatoceros lamarckii* (Annelida): heterochrony in polychaetes. Front Zool 3(1):16

McDougall C, Korchagina N, Tobin J, Ferrier D (2011) Annelid *Distal-less/Dlx* duplications reveal varied post-duplication fates. BMC Evol Biol 11(1):241

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 35(Web Server issue):W182–185. doi:10.1093/nar/gkm321

Papaioannou VE, Silver LM (1998) The T-box gene family. Bioessays 20(1):9–19

Paps J, Holland PWH, Shimeld SM (2012) A genome-wide view of transcription factor gene diversity in chordate evolution: less gene loss in amphioxus? Brief Funct Genom 11(2):177–186

Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B (2003) TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19:651–652

Putnam N, Butts T, Ferrier D, Furlong R, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu J (2008) The amphioxus genome and the evolution of the chordate karyotype. Nature 453:1064–1071

Raible F, Tessmar-Raible K, Osoegawa K, Wincker P, Jubin C, Balavoine G, Ferrier D, Benes V, De Jong P, Weissenbach J (2005) Vertebrate-type intron-rich genes in the marine annelid *Platynereis dumerilii*. Science 310:1325–1326

Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, Griffith M, Raymond A, Thiessen N, Cezard T, Butterfield YS, Newsome R, Chan SK, She R, Varhol R, Kamoh B, Prabhu A-L, Tam A, Zhao Y, Moore RA, Hirst M, Marra MA, Jones SJM, Hoodless PA, Birol I (2010) De novo assembly and analysis of RNA-seq data. Nat Meth 7 (11):909–912

Schmerer M, Savage RM, Shankland M (2009) *Paxβ:* a novel family of lophotrochozoan Pax genes. Evol Dev 11(6):689–696

Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. doi:10.1093/bioinformatics/bts094

Schuster S (2008) Next-generation sequencing transforms today's biology. Nat Meth 5:16–18

Segrove F (1941) The development of the Serpulid *Pomatoceros triqueter* L. Q J Microsc Sci 82:467–540

Shimeld SM, Boyle MJ, Brunet T, Luke GN, Seaver EC (2010a) Clustered *Fox* genes in lophotrochozoans and the evolution of the bilaterian *Fox* gene cluster. Dev Biol 340(2):234–248

Shimeld SM, Degnan B, Luke GN (2010b) Evolutionary genomics of the *Fox* genes: origin of gene families and the ancestry of gene clusters. Genomics 95(5):256–260

Small K, Brudno M, Hill M, Sidow A (2007) A haplome alignment and reference sequence of the highly polymorphic *Ciona savignyi* genome. Genome Biol 8(3):R41

Sodergren E, Weinstock GM, Davidson EH et al (2006) The genome of the sea urchin *Strongylocentrotus purpuratus*. Science 314 (5801):941–952

Struck TH, Paul C, Hill N, Hartmann S, Hosel C, Kube M, Lieb B, Meyer A, Tiedemann R, Purschke G, Bleidorn C (2011) Phylogenomic analyses unravel annelid evolution. Nature 471 (7336):95–98

Surget-Groba Y, Montoya-Burgos JI (2010) Optimization of *de novo* transcriptome assembly from next-generation sequencing data. Genome Res 20(10):1432–1440

Tagawa K, Humphreys T, Satoh N (2000) *T-brain* expression in the apical organ of hemichordate tornaria larvae suggests its evolutionary link to the vertebrate forebrain. J Exp Zool 288 (1):23–31

Takahashi T, Holland PWH (2004) Amphioxus and ascidian *Dmbx* homeobox genes give clues to the vertebrate origins of midbrain development. Development 131(14):3285–3294

Takahashi T, McDougall C, Troscianko J, Chen W-C, Jayaraman-Nagarajan A, Shimeld S, Ferrier D (2009) An EST screen from the annelid *Pomatoceros lamarckii* reveals patterns of gene loss and gain in animals. BMC Evol Biol 9(1):240

Takatori N, Butts T, Candiani S, Pestarino M, Ferrier D, Saiga H, Holland P (2008) Comprehensive survey and classification of homeobox genes in the genome of amphioxus, *Branchiostoma floridae*. Dev Genes Evol 218(11):579–590

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28(10):2731–2739

Tessmar-Raible K, Arendt D (2003) Emerging systems: between vertebrates and arthropods, the Lophotrochozoa. Curr Opin Genetics Dev 13:331–340

Tessmar-Raible K, Raible F, Christodoulou F, Guy K, Rembold M, Hausen H, Arendt D (2007) Conserved sensory-neurosecretory cell types in annelid and fish forebrain: insights into hypothalamus evolution. Cell 129:1389–1400

Vera J, Wheat C, Fescemyer H, Frilander M, Crawford D, Hanski I, Marden J (2008) Rapid transcriptome characterization for a non-model organism using 454 pyrosequencing. Mol Ecol 17 (7):1636–1647

Wang X-W, Luan J-B, Li J-M, Bao Y-Y, Zhang C-X, Liu S-S (2010) De novo characterization of a whitefly transcriptome and analysis of its gene expression during development. BMC Genomics 11(1):400

Waterhouse AM, Procter JB, Martin DMA, Ml C, Barton GJ (2009) Jalview Version 2, A multiple sequence alignment editor and analysis workbench. Bioinformatics 25(9):1189–1191

Winchell C, Valencia J, Jacobs D (2010) Expression of Distal-less, dachshund, and optomotor blind in *Neanthes arenaceodentata* (Annelida, Nereididae) does not support homology of appendage-forming mechanisms across the Bilateria. Dev Genes Evol 220(9):275–295

Zerbino DR (2010) Using the Velvet de novo assembler for short-read sequencing technologies. Curr Protoc Bioinformatics Chapter 11: Unit 11 15

Zerbino D, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829

Zhao Q-Y, Wang Y, Kong Y-M, Luo D, Li X, Hao P (2011) Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. BMC Bioinforma 12(Suppl 14):S2

Zhong YF, Holland PW (2011) HomeoDB2: functional expansion of a comparative homeobox gene database for evolutionary developmental biology. Evol Dev 13:567–568

Zhong YF, Butts T, Holland PW (2008) HomeoDB: a database of homeobox gene diversity. Evol Dev 10(5):516–518

General and Comparative Endocrinology 188 (2013) 16–22

Contents lists available at SciVerse ScienceDirect

General and Comparative Endocrinology

journal homepage: www.elsevier.com/locate/ygcen

Review

How are comparative genomics and the study of microRNAs changing our views on arthropod endocrinology and adaptations to the environment?

Nathan J. Kenny [a,1], Shan Quah [a,1], Peter W.H. Holland [a,*], Stephen S. Tobe [b,*], Jerome H.L. Hui [a,c,*]

[a] Department of Zoology, University of Oxford, South Parks Road, OX1 3PS, UK
[b] Department of Cell and Systems Biology, University of Toronto, Canada M5S 3G5
[c] School of Life Sciences, Chinese University of Hong Kong, Shatin, Hong Kong

ARTICLE INFO

Article history:
Available online 7 March 2013

Keywords:
Juvenile hormone
Comparative genomics
Next generation sequencing
MicroRNA
Arthropods
Comparative endocrinology

ABSTRACT

As the last few decades of work has shown, precise regulation of biosynthesis and release of arthropod hormones is essential to cope with environmental stresses and challenges. In crustaceans and insects, the sesquiterpenoids methyl farnesoate (MF), farnesoic acid (FA) and juvenile hormone (JH) regulate many developmental, physiological, and reproductive processes. In this review, we discuss how comparative genomics has and will impact our views on arthropod endocrinology. We will also highlight the current knowledge of regulation of genes involved in arthropod hormone biosynthesis by microRNAs, and describe the potential insights into arthropod endocrinology, evolution, and adaptation that are likely to come from the study of microRNAs.

1. Introduction

All multicellular animals produce hormones (from the Greek "hormon": excite) to control their developmental, physiological, and reproductive activities. The phylum Arthropoda (Greek: jointed leg) is highly speciose and comprises the majority of described extant animal species including arachnids, crustaceans, and insects. The diversity of the habitats that they have conquered perhaps is partly a result of the unique sesquiterpenoid hormonal system that has coevolved with them, which has been shown to be responsive to environmental stresses such as anoxia, temperature, salinity, and even environmental contaminants (e.g. Borst et al., 2001; LeBlanc, 2007; Lovett et al., 1997, 2001; Nagaraju and Borst, 2008).

Hormone production in bilaterians shares a conserved mevalonate biosynthetic pathway derived from simple acetate molecules. Different final products are generated in different species (i.e. cholesterol in vertebrates, juvenile hormone (JH) in insects, and methyl farnesoate (MF) and farnesoic acid (FA) in crustaceans, Fig. 1). Details of the pathway have been extensively reviewed, and will not be repeated here (Bellés et al., 2005; Hui et al., 2013; Tobe and Bendena, 1999). This article will focus on the comparison of JH in insects with the MF and FA in crustaceans. These hormones are thought to be produced in structurally similar organs (MF and FA in the mandibular organ and JH in the corpora allata), and are believed to serve similar functions in the regulation of gametogenesis/reproduction, metabolic activities, metamorphosis and ecdysteroid secretion for moulting (Laufer et al., 1987; Le Roux, 1968; Tiu et al., 2009, 2012; Tobe and Stay, 1985; Tobe et al., 1989). As the chemical structure of MF lacks the epoxide group found in JH-III, and as no JH has ever been identified in crustaceans, MF or FA are commonly thought to be the crustacean equivalents of "JH", and to date, JH is generally considered to be an evolutionary derivative of MF and a hormone unique to insects (see section 3.3).

One rate-limiting step in the biosynthesis of these hormones is thought to be the final step, the conversion to JH or MF through a *S*-adenosyl-methyltransferase (SAM)-dependent methylation (for details, see Hui et al., 2010, 2013; Tobe and Bendena, 1999). In insects, the methylation occurs by way of the juvenile hormone methyltransferase (JHAMT), which is involved in the conversion of JH acids or FA to JH (through MF) in a range of insects (Kinjoh et al., 2007; Marchal et al., 2011; Minakuchi et al., 2008; Niwa et al., 2008; Shinoda and Itoyama, 2003). Studies of knockdown and overexpression of JHAMT all suggest an essential role in main-

Abbreviations: JH, juvenile hormone; MF, methyl farnesoate; NGS, next generation sequencing; JHAMT, juvenile hormone methyltransferase; FAMeT, farnesoic acid *O*-methyltransferase; SAM, *S*-adenosyl-methyltransferase; RNAi, RNA interference.

* Corresponding authors. Address: School of Life Sciences, Chinese University of Hong Kong, Shatin, Hong Kong (J.H.L. Hui).
E-mail address: hui.jerome@gmail.com (J.H.L. Hui).

[1] These authors contributed equally.

0016-6480/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ygcen.2013.02.013

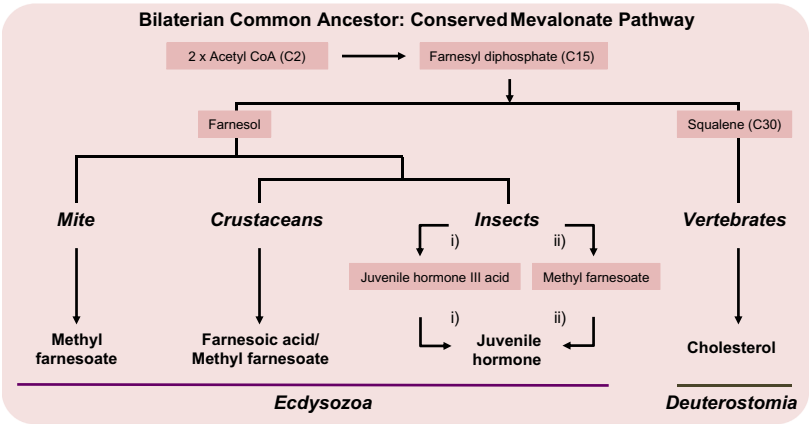


Fig. 1. Comparative endocrinology of the Mevalonate pathway within the Bilateria modified from Bellés et al. (2005), Hui et al. (2013), and Tobe and Bendena (1999). Superphyla and subphyla where the mevalonate pathway has been studied are shown. Intermediate molecules within the pathway are shown in boxed darker coloration. Note crustacean water flea *Daphnia pulex* (Hui et al., 2010) and mite *Tetranychus urticae* (Grbic et al., 2011) contain JHAMTs in their genomes, but may not produce JH.

taining normal development and metamorphosis (Kinjoh et al., 2007; Minakuchi et al., 2008; Niwa et al., 2008).

However, the corresponding situation in crustaceans is not as clear. For the past decade, the putative farnesoic acid *O*-methyltransferase (FAMeT) in crustaceans was assumed to play this role (e.g. Burtenshaw et al., 2008; Gunawardene et al., 2002; Holford et al., 2004; Hui et al., 2008; Wainwright et al., 1998). With the first systematic phylogenomic analysis of FAMeT and JHAMT across bilaterians (Hui et al., 2010) this received view of crustacean endocrinology has been completely revised (for full details, see review (Hui et al., 2013). In summary, the finding of a true JHAMT in the water flea *Daphnia pulex* reveals that, contrary to earlier opinion, crustaceans do possess the JHAMT enzyme and JHAMT is not unique to insects. This discovery of the first *bona fide* JHAMT in a crustacean also opens new avenues for research. Examples of questions raised by this discovery include those related to the expression, role and function of JHAMT in crustacean endocrinology. Further, *JHAMT* is known to be one of the most up-regulated genes in presence of kairomone signals in water fleas (a typical ecotoxicological model organism) (Miyakawa et al., 2010). Study of the interactions of JHAMT and biosynthesis of sesquiterpenoids in different arthropods under different environmental conditions may also provide useful information on ecotoxicology.

With the recent rapid advancement in sequencing technologies, and a simultaneous reduction in costs, the feasibility of applying comparative genomic approaches to the questions above is now a reality, even for smaller laboratories. Here we will discuss how these comparative genomic approaches have been made possible, how they may be performed, and how they might change our view of arthropod endocrinology as a whole. These studies are likely to focus on the clade Ecdysozoa within the animals, since it is already is clear that no JH biosynthetic pathway specific components have been observed in the lophotrochozoans or deuterostomes (Fig. 1).

In the present review, we will focus our discussion on the relatives of insects and crustaceans that can be used to test evolutionary hypotheses within these clades. Also, recent studies show that non-coding microRNAs have huge impact on many areas of biology by regulating genes at the post-transcriptional level, and one which is largely unexamined in this field (Hui et al., 2013). We will therefore also review the latest findings on the role of non-coding RNAs in arthropod endocrinology, and discuss how the study of microRNAs may contribute to our understanding of arthropod hormones and the adaptations of arthropods to their environment.

2. Comparative genomics

Recent years have seen an explosion in the amount of genomic data available from across the animal kingdom. This increase in data availability has revolutionised comparative biology, allowing inferences into the evolution and diversification of traits at a molecular level. Whereas initial sequencing efforts were concentrated on traditional model organisms such as flies and nematodes, increased sampling across the metazoan 'tree of life' has quickly shed much light on previously intractable problems in assessing the course of evolution, as we will illustrate with particular reference to the evolution of the hormonal system within the Arthropoda.

2.1. Next generation sequencing and assembly

Until recently, genome projects have depended on traditional Sanger-based shotgun or clone-by-clone sequencing. While effective, these methods have always been slow, labour intensive and expensive, and with recent technological advances, this approach has been almost entirely supplanted by next-generation sequencing (NGS) technologies Shendure and Ji, 2008. A range of NGS platforms are available for purchase or commercial use. Perhaps the most common providers are those utilizing Illumina or Roche 454 chemistry, although others (such as PacBio, Ion Torrent, SOLiD and Helioscope) are also available. Each source of NGS data has its own merits and drawbacks, and these have been covered in depth elsewhere (Glenn, 2011; Henson et al., 2012). Nanopore sequencing technology is likely to appear commercially in the near future, and the long read length provided by such an approach might further simplify genome assembly (Hayden, 2012).

A major difference between traditional Sanger approaches and NGS methods is the amount of DNA sequence generated. For example, a single DNA sequencing run on the most advanced Sanger platform (the ABI 3730XL machine) can yield up to 70 kb of DNA sequence derived from 96 reactions each generating ~700 bp. In contrast, a single lane run on an Illumina HiSeq generates over

35 Gb of sequence: a staggering half a million times as much data. The vast amount of data provided by a typical NGS project has necessitated the use of novel algorithms for assembling these data, although no firm consensus as to the best to use has been established. These assemblers generally rely on de Bruijn graph-based methods, which are memory-intensive but able to cope with large datasets (Wajid and Serpedin, 2012). Assemblers vary in performance, and their relative merits have been well reviewed elsewhere (Lin et al., 2011; Zhang et al., 2011). It is generally recommended that a number of assemblers are trialled, with selection guided by computational resources available, and once a preferred option is identified, optimisation is attempted.

The first genome to be entirely sequenced on next-generation platforms was that of the Giant Panda, *Ailuropoda melanoleuca* (Li et al., 2010), swiftly followed by that of the Turkey *Meleagris gallopavo* (Dalloul et al., 2010). Many other genomes have been sequenced since, although the release of these lags far behind the speed of sequencing. Many NGS-derived genomes are on the horizon, including a 5,000 insect genomes project (Arthropod Sequencing Consortium, 2012) and a vision for 10,000 vertebrate genomes (Genome 10K Community of Scientists, 2009). Furthermore, the amount of data generated by NGS sequencing, and its cost effectiveness, has brought genome sequencing into the realm of individual laboratory projects.

2.2. Metazoan interrelationships

One area on which NGS technology has been revolutionary has been the study of metazoan interrelationships. Early phylogenetic trees were often based on one or a small sample of genes, amplified via PCR, often using degenerate primers. Such early approaches were vulnerable to bias caused by variation in these loci in some of the species sampled, which often unduly affected the form of the final tree. More recently, the creation of expressed sequence tag (EST) libraries has allowed the comparison of a larger number of sites across an increased number of genes, ameliorating these stochastic effects (Telford and Copley, 2011). NGS, and in particular Roche 454 technology has proved particularly useful for the creation of EST datasets, with the long read length of 454 allowing the derivation of useful information even without assembly. Today, phylogenetic trees are therefore constructed on the basis of alignments of hundreds or thousands of proteins. The combination of large datasets and more complete sampling across the range of animal diversity has already solved some formerly intractable problems in phylogeny.

In Fig. 2, a consensus cladogram of metazoan relationships is shown, based upon three recent and extensive phylogenetic analyses (Dunn et al., 2008; Hejnol et al., 2009; Pick et al., 2010). Our general picture of animal evolution, drawn from such trees, is that there were several early branching lineages (collectively marked non-Bilateria in Fig. 2), plus a large clade of bilaterally symmetrical animals called Bilateria. Discounting a putative early diverging branch of this clade, the Acoelomorpha, the bulk of Bilateria is generally divided into three great superphyla: the Ecdysozoa, the Lophotrochozoa and the Deuterostomia. The present availability of genomic-scale information for each phylum, or its pending arrival, is indicated at right. Despite the increasing availability of next-generation sequencing techniques, it is clear that some clades within the Metazoa have been more extensively sampled than others. A number of factors contribute to this – in particular, some phyla are more medically, commercially or "scientifically" important than others.

Building a tree of evolutionary relationships is just the start for evolutionary study. When phylogenetic frameworks are well established and a number of genomic datasets are available, then patterns of character change can be inferred, such as the gain and loss of genes or biochemical pathways. This is aided considerably if phyla are sampled in some depth, since this avoids the problem that an absence in one species might not be typical of the entire group. Hence, long-held assertions can be revisited and their validity tested. A good example of a phylum with extensive genome sampling is the Arthropoda, where a number of key experiments can now be guided by a comparative genomic approach to assessing their evolution.

2.3. Arthropoda, Pancrustacea and comparative genomics

In terms of genomic resources, the Arthropoda is perhaps the best-sampled phylum within the Metazoa. Several species were the subject of intensive early genomic investigation as a consequence of their importance in agriculture and as developmental models, including *Drosophila melanogaster, Apis mellifera, Tribolium castaneum* and *Acyrthosiphon pisum*. More importantly for comparative studies, many additional arthropods have recently been the subject of genome sequencing, including crustaceans, such as the water flea *D. pulex* and the amphipod *Parhyale hawaiensis* (JGI genome website), and chelicerates, such as the tick *Ixodes scapularis* (Hill and Wikel, 2005) and the mite *Tetranychus urticae* (Grbic et al., 2011). Additional genome projects include the centipede *Strigamia maritimia* (BioProject PRJNA20501) and the tardigrade *Hypsibius dujardini* (PRJNA20353), plus numerous other species in individual laboratories. These taxa provide valuable outgroups for studies of insect trait evolution.

Fig. 3 shows the interrelationships within the Arthropoda, as determined by two recent analyses (Regier et al., 2010; Rota-Stabelli et al., 2011). Based on morphological evidence, myriapods (centipedes and millipedes) and hexapods (insects) were historically believed to be close relatives, on the basis of the similar arrangement of their head, mouthparts, limbs and trachea. Molecular phylogenetic analysis, however, showed that insects are more properly nested within the Crustacea (forming a clade renamed as the Pancrustacea), and quite distant from the Myriapoda. While some phylogenetic trees showed Myriapoda and Chelicerata (spiders) as forming a monophyletic clade, the most recent genomically-informed datasets using microRNA data conclude that Myriapoda are the sister group of the Pancrustacea, to the exclusion of the Chelicerata (Rota-Stabelli et al., 2011). This latter conclusion is still controversial, because inferring phylogenies from presence or absence of characters (such as particular microRNAs) is heavily dependent on sampling. Indeed, a similar study of the placement of Tardigrada (Water Bears) and Onychophora (Velvet Worms) relative to Arthropoda suggests that Velvet Worms may be the sister group of the Arthropoda (Campbell et al., 2011); however, since this is based on the shared presence of a single microRNA (miR-305), we suggest that wider Tardigrada sequencing is needed to strengthen this claim. The conclusion that seems very secure, and highly relevant to the study of endocrine evolution, is that insects and crustaceans are closely related, forming the clade Pancrustacea. Indeed, it is likely that the insects are evolutionarily nested within the crustaceans, and therefore derived from them (Regier et al., 2005). In short, insects are terrestrial crustaceans, suggesting JH is an evolutionary novelty evolving following the recruitment of the biosynthetic machinery to modify MF.

2.4. Arthropod hormonal systems and comparative genomics

With a firm phylogenetic framework in place, the reappraisal of many long-held beliefs about arthropod evolution has already begun. Arthropods, with their dependence on moulting for growth, and hormonally controlled reproductive cycles, have many unique characters. Understanding how and where they arose in the course of evolution is an excellent test of the power of a comparative

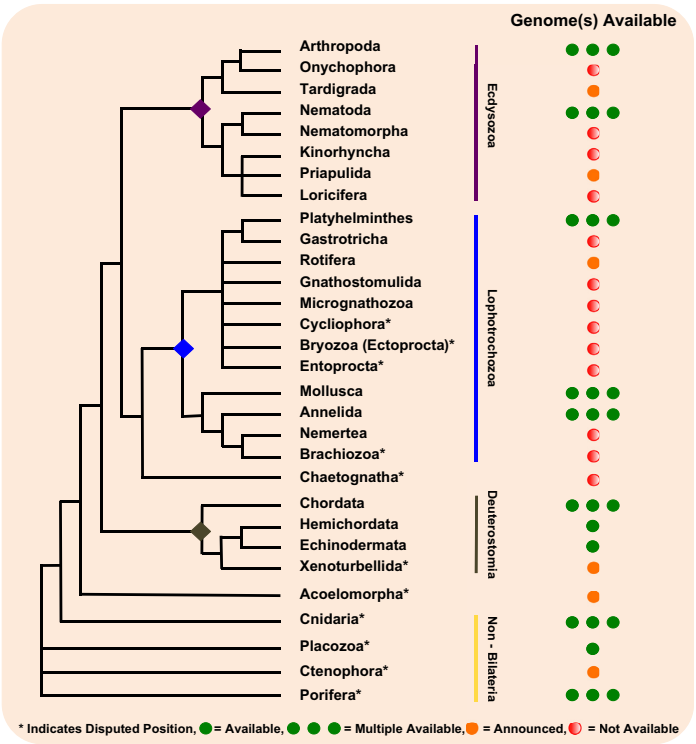


Fig. 2. Cladogram of metazoan evolutionary relationships, based on Dunn et al. (2008), Hejnol et al. (2009), and Pick et al. (2010). Problematica indicated with an asterisk. At right, current public availability of genomic datasets is indicated, according to the key found at the base of the figure.

genomic approach. The sesquiterpenoid biosynthetic pathway, as mentioned earlier, represents an interesting case in this regard.

Will studies on other ecdysozoan genomes provide any clue to the evolution of MF and JH biosynthesis, degradation, and regulation? The answer seems to be positive. With the discovery of JHAMT outside the Insecta using a comparative genomic approach (Hui et al., 2010), new questions have begun to be raised as to the role and evolution of JHAMT and other JH-biosynthetic, degradation, and regulatory components (see below and Hui et al., 2013). In the recent sequencing of the mite genome (Grbic et al., 2011), Smagghe and colleagues in the Mite Genome Consortium found unique JH biosynthetic genes, although hormonal assays and the absence of critical components (such as the juvenile hormone epoxidase gene) suggest only MF is present in this species. By using the genomic databases established in the wider Panarthropoda, it should be possible to determine when exactly these different genes arose. The increasing availability of genomes within the Ecdysozoa should act as a firm guide as to where functional work is necessary to dissect the evolution of the sesquiterpenoid pathway further.

3. MicroRNAs

3.1. MicroRNA biogenesis and function

MicroRNAs are short 21–23 nucleotide non-coding RNAs which act in post-transcriptional regulation of gene expression. They are initially transcribed from long primary transcripts which include one or more imperfect ~80 nucleotide hairpins (Lee et al., 2002, Fig. 4). Primary transcripts are cleaved in the nucleus by an RNase III enzyme, Drosha, to release each hairpin as precursor microRNA (Lee et al., 2003). The hairpin precursors are exported to the cytoplasm where they are further processed by a cytoplasmic RNase III enzyme, Dicer, to generate a 21–23 nucleotide RNA duplex with 3′ overhangs of 2 nucleotides at each end. Mature microRNAs of this duplex are then transferred to protein complex containing Argonaute, where it serves to target the protein complex to specific messenger RNA transcripts by partial complementary base pairing in animals (Eulalio et al., 2008).

3.2. MicroRNAs and the regulation of arthropod development and hormone productions

There is now a growing body of evidence that microRNAs are involved in the coordination of growth, developmental timing, and hormones production. Initial work in the nematode *Caenhorhabditis elegans* revealed that microRNAs contribute to the regulation of postembryonic development (Lau et al., 2001; Reinhart et al., 2000). For example, mutation in the let-7 microRNA gene causes a heterochronic defect at larval stage 4 (L4) to adult moult, resulting in a supernumerary moult to a fifth larval stage. Overexpression of let-7 causes an opposite heterochronic effect where there

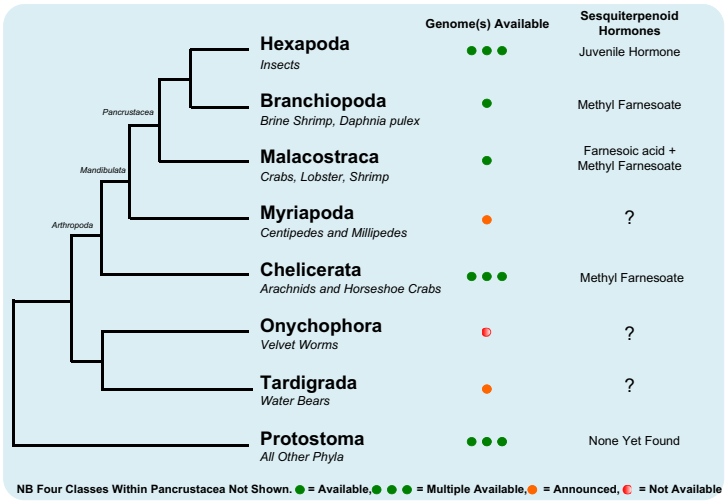


Fig. 3. Cladogram of arthropod evolutionary relationships, based on Regier et al. (2010) and Rota-Stabelli et al. (2011). Common names of typical members shown in italics below each branch label. Clade names shown in italics at some nodes. Four classes of Pancrustacea, lying within that clade, excluded for clarity. Also shown is genomic sequence availability and the presence of sesquiterpenoid hormones, when known.

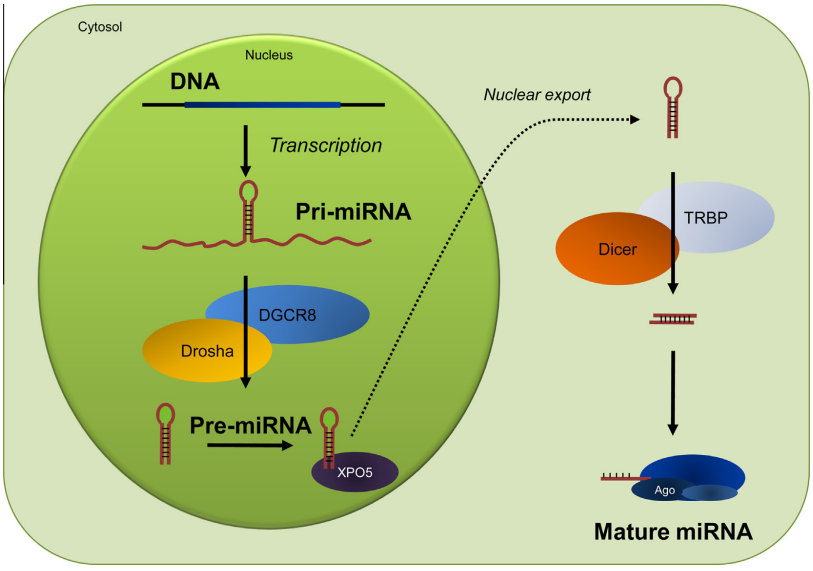


Fig. 4. Schematic diagram showing the biogenesis pathway of microRNAs.

is premature terminal differentiation of hypodermal cells after the L3 stage (Reinhart et al., 2000).

MicroRNAs have also been shown to be important in adult metamorphosis and hormonal level changes in insects. Knocking down Dicer-1 in the cockroach *Blatella germanica* using RNAi administered at the penultimate (5th) instar permitted normal moulting to the 6th instar, but generated a supernumerary moult to a 7th instar with functional prothoracic glands instead of the adult (Gomez-Orte and Belles, 2009). Since this nymphoid phenotype is similar to that obtained by treatment with JH, and knocking down Dicer-1 does not increase JH production, this suggests that Dicer-1 either acts downstream of, or in parallel to, the JH pathway.

Other examples of microRNAs playing roles in insect development and hormone production were elucidated in fly *D. melanogaster*, where a rise in ecdysteroid titre at the end of L3 triggers the initiation of metamorphosis (Sempere et al., 2002, 2003). During this period, there is an ecdysone-dependent increase in the expression of miR-100, miR-125 and let-7, with a similar ecdysone-mediated decrease in expression level of miR-34. On the other hand, in mutants lacking all Broad-Complex (BR-C), miR-100, miR-125 and let-7 are expressed at lower levels than in wild type flies, but the expression of miR-34 is enhanced. Considering the fly miR-125 is a homologue of the aforementioned *C. elegans* lin-4, and chromosomal clustering of miR-100, miR-125 and let-7 appears to be broadly conserved in animal phylogeny (Griffiths-Jones et al., 2011; Sempere et al., 2003), and JH blocks ecdysteroid-induced expression of BR-C RNA in some animals such as the tobacco hornworm *Manduca sexta* (Zhou et al., 1998); this evidence could suggest there is a conserved microRNA regulatory mechanism through Broad-Complex (BR-C) proteins on insect hormones.

The use of high-throughput sequencing has also facilitated the analysis of microRNA expression relative to the actions of JH and ecdysteroids. For example, deep sequencing of two microRNA libraries from *B. germanica* (one from ecdysone peak in the penultimate larval instar (N5) and one during the last larval instar (N6), where JH is present in N5 but not in N6) (Rubio et al., 2012; Treiblmayr et al., 2006) identified let-7–5p, miR-100–5p and miR-125–5p were highly expressed during the ecdysone peak in N6 but not in N5, comparable to the findings made in *D. melanogaster*. Deep sequencing of small RNAs isolated from different stages in the silkworm moth *Bombyx mori* also found several microRNAs with predicted roles in regulation of development, moulting and metamorphosis (Jagadeeswaran et al., 2010; Liu et al., 2010). For example, Bmo-miR-2998 is predicted to target JHAMT, while the regulator of JH titer (juvenile hormone esterase) is a predicted target of Bmo-miR-2766. Nevertheless, how these microRNAs regulate JH is currently unknown.

4. Conclusions

A host of key questions remain to be asked and answered regarding the evolution and regulation of arthropod hormone production. For example, how do the insects and other ecdysozoans produce different sesquiterpenoid products? How did the biosynthetic, degradation, and regulation on these hormones evolve from the other? Was there a fundamental microRNA-hormone axis established early in the evolution of arthropods? How has this become modified in different lineages? With the feasibility of sequencing technology, the increasing availability of genomic datasets, and rapid expanding knowledge of microRNAs, the framework with which to interpret the results of key experiments is now in place.

Acknowledgments

NJK is funded by a Clarendon Fund Scholarship (Oxford), and SQ is funded by the A*STAR National Science Scholarship (Singapore). "This work was funded by the Direct Grant (4053034) of the Chinese University of Hong Kong (JHLH).

References

Bellés, X., Martín, D., Piulachs, M.D., 2005. The mevalonate pathway and the synthesis of juvenile hormone in insects. Annu. Rev. Entomol. 50, 181–199.
Borst, D.W., Ogan, J.T., Tsukimura, B., Claerhout, T., Holford, K.C., 2001. Regulation of the crustacean mandibular organ. Am. Zool. 41, 430–441.
Burtenshaw, S.M., Su, P.P., Zhang, J.R., Tobe, S.S., Dayton, L., Bendena, W.G., 2008. A putative farnesoic acid O-methyltransferase (FAMeT) orthologue in *Drosophila melanogaster* (CG10527): relationship to juvenile hormone biosynthesis. Peptides 29, 242–251.
Campbell, L.I., Rota-Stabelli, O., Edgecombe, G.D., Marchioro, T., Longhorn, S.J., Telford, M.J., et al., 2011. MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda. Proc. Natl. Acad. Sci. USA 108, 15920–15924.
Dalloul, R.A., Long, J.A., Zimin, A.V., Aslam, L., Beal, K., Ann Blomberg, L., et al., 2010. Multi-platform next-generation sequencing of the domestic Turkey *Meleagris gallopavo*: genome assembly and analysis. PLoS Biol. 8, e1000475.
Dunn, C.W., Hejnol, A., Matus, D.Q., Pang, K., Browne, W.E., Smith, S.A., et al., 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452, 745–749.
Eulalio, A., Huntzinger, E., Izaurralde, E., 2008. GW182 interaction with Argonaute is essential for microRNA-mediated translational repression and mRNA decay. Nat. Struct. Mol. Biol. 15, 346–353.
Genome 10K Community of Scientists, 2009. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. J. Hered. 100, 659–674.
Glenn, T.C., 2011. Field guide to next-generation DNA sequencers. Mol. Ecol. Resour. 11, 759–769.
Gomez-Orte, E., Belles, X., 2009. MicroRNA-dependent metamorphosis in hemimetabolan insects. Proc. Natl. Acad. Sci. USA 106, 21678–21682.
Grbic, M., Van Leeuwen, T., Clark, R.M., Rombauts, S., Rouze, P., Grbic, V., et al., 2011. The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. Nature 479, 487–492.
Griffiths-Jones, S., Hui, J.H., Marco, A., Ronshaugen, M., 2011. MicroRNA evolution by arm switching. EMBO Rep. 12, 172–177.
Gunawardene, Y.I., Tobe, S.S., Bendena, W.G., Chow, B.K., Yagi, K.J., Chan, S.M., 2002. Function and cellular localization of farnesoic acid O-methyltransferase (FAMeT) in the shrimp, *Metapenaeus ensis*. Eur. J. Biochem. 269, 3587–3595.
Hayden, E.C., 2012. Nanopore genome sequencer makes its debut (News Article). Nature 482, 445.
Hejnol, A., Obst, M., Stamatakis, A., Ott, M., Rouse, G., Edgecombe, G., et al., 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc. Biol. Sci. 276, 4261–4270.
Henson, J., Tischler, G., Ning, Z., 2012. Next-generation sequencing and large genome assemblies. Pharmacogenomics 13, 901–915.
Hill, C.A., Wikel, S.K., 2005. The *Ixodes scapularis* genome project: an opportunity for advancing tick research. Trends Parasitol. 21, 151–153.
Holford, K.C., Edwards, K.A., Bendena, W.G., Tobe, S.S., Wang, Z., Borst, D.W., 2004. Purification and characterization of a mandibular organ protein from the American lobster, *Homarus americanus*: a putative farnesoic acid O-methyltransferase. Insect Biochem. Mol. Biol. 34, 785–798.
Hui, J.H.L., Tobe, S.S., Chan, S.M., 2008. Characterization of the putative farnesoic acid O-methyltransferase (LvFAMeT) cDNA from white shrimp *Litopenaeus vannamei:* evidence for its role in molting,. Peptides 29, 252–260.
Hui, J.H.L., Hayward, A., Bendena, W.G., Takahashi, T., Tobe, S.S., 2010. Evolution and functional divergence of enzymes involved in sesquiterpenoid hormone biosynthesis in crustaceans and insects. Peptides 31, 451–455.
Hui, J.H.L., Bendena, W.G., Tobe, S.S., 2013. Future perspectives on the research of juvenile hormones and sesquiterpenoids in arthropod endocrinology and ecotoxicology. In: Devillers, J. (Ed.), Juvenile Hormones and Juvenoids: Modelling Biological Effects and Environmental Fate. CRC Press, New York.
Jagadeeswaran, G., Zheng, Y., Sumathipala, N., Jiang, H., Arrese, E.L., Soulages, J.L., et al., 2010. Deep sequencing of small RNA libraries reveals dynamic regulation of conserved and novel microRNAs and microRNA-stars during silkworm development. BMC Genomics 11, 52.
JGI genome website. http://genome.jgi-psf.org/.
Kinjoh, T., Kaneko, Y., Itoyama, K., Mita, K., Hiruma, K., Shinoda, T., 2007. Control of juvenile hormone biosynthesis in *Bombyx mori:* cloning of the enzymes in the mevalonate pathway and assessment of their developmental expression in the corpora allata. Insect Biochem. Mol. Biol. 37, 808–818.
Lau, N.C., Lim, L.P., Weinstein, E.G., Bartel, D.P., 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. Science 294, 858–862.
Laufer, H., Borst, D., Baker, F.C., Reuter, C.C., Tsai, L.W., Schooley, D.A., Carrasco, C., Sinkus, M., 1987. Identification of a juvenile hormone-like compound in a crustacean. Science 235, 202–205.
LeBlanc, G.A., 2007. Crustacean endocrine toxicology: a review. Ecotoxicology 16, 61–81.
Le Roux, A., 1968. Description of d'organes mandibulaires nouveaux chez les crustaces decapods. C. R. Acad. Sci. Paris (D) 226, 1414–1417.
Lee, Y., Jeon, K., Lee, J., Kim, S., Kim, V.N., 2002. MicroRNA maturation : stepwise processing and subcellular localization. EMBO J. 21, 4663–4670.
Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., et al., 2003. The nuclear RNase III Drosha initiates microRNA processing. Nature 425, 1–5.
Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., et al., 2010. The sequence and de novo assembly of the giant panda genome. Nature 463, 311–317.
Lin, Y., Li, J., Shen, H., Zhang, L., Papasian, C.J., Deng, H.A., 2011. Comparative studies of de novo assembly tools for next-generation sequencing technologies. Bioinformatics 27, 2031–2037.
Liu, S., Li, D., Li, Q., Zhao, P., Xiang, Z., Xia, Q., 2010. MicroRNAs of *Bombyx mori* identified by Solexa sequencing. BMC Genomics 11, 148.
Lovett, D.L., Clifford, P.D., Borst, D.W., 1997. Physiological stress elevates hemolymph levels of methyl farnesoate in the green crab *Carcinus maenas*. Biol. Bull. 193, 266–267.

Lovett, D.L., Verzi, M.P., Clifford, P.D., Borst, D.W., 2001. Hemolymph levels of methyl farnesoate increase in response to osmotic stress in the green crab, *Carcinus maenas*. Comp. Biochem. Physiol. A Mol. Integr. Physiol. 128, 299–306.

Marchal, E., Zhang, J., Badisco, L., Verlinden, H., Hult, E.F., Van Wielendaele, P., Yagi, K.J., Tobe, S.S., Vanden Broeck, J., 2011. Final steps in juvenile hormone biosynthesis in the desert locust, *Schistocerca gregaria*. Insect Biochem. Mol. Biol. 41, 219–227.

Minakuchi, C., Namiki, T., Yoshiyama, M., Shinoda, T., 2008. RNAi-mediated knockdown of juvenile hormone acid O-methyltransferase gene causes precocious metamorphosis in the red flour beetle *Tribolium castaneum*. FEBS J. 275, 2919–2931.

Miyakawa, H., Imai, M., Sugimoto, N., Ishikawa, Y., Ishikawa, A., Ishigaki, H., et al., 2010. Gene up-regulation in response to predator kairomones in the water flea *Daphnia pulex*. BMC Dev. Biol. 10, 45.

Nagaraju, G.P., Borst, D.W., 2008. Methyl farnesoate couples environmental changes to testicular development in the crab *Carcinus maenas*. J. Exp. Biol. 211, 2773–2778.

Niwa, R., Niimi, T., Honda, N., Yoshiyama, M., Itoyama, K., Kataoka, H., Shinoda, T., 2008. Juvenile hormone acid O-methyltransferase in *Drosophila melanogaster*. Insect Biochem. Mol. Biol. 38, 714–720.

Pick, K.S., Philippe, H., Schreiber, F., Erpenbeck, D., Jackson, D.J., Wrede, P., et al., 2010. Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. Mol. Biol. Evol. 27, 1983–1987.

Regier, J.C., Shultz, J.W., Kambic, R.E., 2005. Pancrustacean phylogeny: hexapods are terrestrial crustaceans and maxillopods are not monophyletic. Proc. Biol. Sci. 272 (1561), 395–401.

Regier, J.C., Shultz, J.W., Zwick, A., Hussey, A., Ball, B., Wetzer, R., et al., 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature 463, 1079–1083.

Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., et al., 2000. The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. Nature 403, 901–906.

Rota-Stabelli, O., Campbell, L., Brinkmann, H., Edgecombe, G.D., Longhorn, S.J., Peterson, K.J., et al., 2011. A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. Proc. R. Soc. B Biol. Sci. 278, 298–306.

Rubio, M., de Horna, A., Belles, X., 2012. MicroRNAs in metamorphic and non-metamorphic transitions in hemimetabolan insect metamorphosis. BMC Genomics 13, 386.

Sempere, L.F., Dubrovsky, E.B., Dubrovskaya, V.A., Berger, E.M., Ambros, V., 2002. The expression of the let-7 small regulatory RNA is controlled by ecdysone during metamorphosis in *Drosophila melanogaster*. Dev. Biol. 244, 170–179.

Sempere, L.F., Sokol, N.S., Dubrovsky, E.B., Berger, E.M., Ambros, V., 2003. Temporal regulation of microRNA expression in *Drosophila melanogaster* mediated by hormonal signals and Broad-Complex gene activity. Dev. Biol. 259, 9–18.

Shendure, J., Ji, H., 2008. Next-generation DNA sequencing. Nat. Biotechnol. 26, 1135–1145.

Shinoda, T., Itoyama, K., 2003. Juvenile hormone acid methyltransferase: a key regulatory enzyme for insect metamorphosis. Proc. Natl. Acad. Sci. USA 100 (2003), 11986–11991.

Telford, M.J., Copley, R.R., 2011. Improving animal phylogenies with genomic data. Trends Genet. 27, 186–195.

The Arthropod Sequencing Consortium, The i5k Project, 2012. http://arthropodgenomes.org/wiki/Main_Page.

Tiu, S.H., Hui, H.L., Tsukimura, B., Tobe, S.S., He, J.G., Chan, S.M., 2009. Cloning and expression study of the lobster (*Homarus americanus*) vitellogenin: conservation in gene structure among decapods. Gen. Comp. Endocrinol. 160 (1), 36–46.

Tiu, S.H., Hult, E.F., Yagi, K.J., Tobe, S.S., 2012. Farnesoic acid and methyl farnesoate production during lobster reproduction: possible functional correlation with retinoid X receptor expression. Gen. Comp. Endocrinol. 175 (2), 259–269.

Tobe, S.S., Stay, B., 1985. Structure and regulation of the corpus allatum. Adv. Insect Physiol. 18, 305–432.

Tobe, S.S., Young, D.A., Khoo, H.W., 1989. Production of methyl farnesoate by the mandibular organs of the mud crab, *Scylla serrata*: validation of a radiochemical assay. Gen. Comp. Endocrinol. 73, 342–353.

Tobe, S.S., Bendena, W.G., 1999. The regulation of juvenile hormone production in arthropods. Functional and evolutionary perspectives. Ann. N. Y. Acad. Sci. 897, 300–310.

Treiblmayr, K., Pascual, N., Piulachs, M.D., Keller, T., Belles, X., 2006. Juvenile hormone titer versus juvenile hormone synthesis in female nymphs and adults of the German cockroach, *Blattella germanica*. J. Insect Sci. 6, 1–7.

Wainwright, G., Webster, S.G., Rees, H.H., 1998. Neuropeptide regulation of biosynthesis of the juvenoid, methyl farnesoate, in the edible crab, *Cancer pagurus*. Biochem. J. 334, 651–657.

Wajid, B., Serpedin, E., 2012. Review of general algorithmic features for genome assemblers for next generation sequencers. Genomics Proteomics Bioinformatics 10, 58–73.

Zhang, W., Chen, J., Yang, Y., Tang, Y., Shang, J., Shen, B., 2011. A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. PLoS One 6, e17915.

Zhou, B., Hiruma, K., Shinoda, T., Riddiford, L.M., 1998. Juvenile hormone prevents ecdysteroid-induced expression of broad complex RNAs in the epidermis of the tobacco hornworm, *Manduca sexta*. Dev. Biol. 203, 233–244.

Appendix E: Reciprocal Best Blast

This appendix contains the reciprocal blast hits for genes mentioned as present or absent in this thesis, but for which no phylogenetic evidence has been presented, particularly the *Lefty* and *Fox H* genes. Evidence is presented for *P. lamarckii, P. vulgata* and *B. plicatilis* in order by gene, first showing absence of evidence for the presence of *Lefty* in our samples (tblastn using *Strongylocentrotus purpuratus Lefty* ACA04466.1 and *Mus musculus* Lefty CAA03909.1 to search.

Secondly, Fox family hits from *P. vulgata* and *B. plicatilis* species are shown, using *Branchiostoma floridae Fox H*, ACE79158.1 as the known sequence query.

Thirdly, as these hit multiple other Fox family genes due to homology in the Forkhead Box region, the first three hits for each of these species by reciprocal blastx to the nr database is used to show that bona fide Fox H genes are (in the case of *P. vulgata*, see especially pages 293/294/295) or are not (*B. plicatilis*) present in these species. For *P. lamarckii* Fox information, please refer to the appropriate section of Chapter 3.

```
TBLASTN 2.2.27+


Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro A.
Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J.
Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs", Nucleic Acids Res. 25:3389-3402.



Database: Pomatcontigs500.fasta
           63,599 sequences; 75,827,742 total letters


Query= gi|167859062|gb|ACA04466.1| lefty [Strongylocentrotus purpuratus]

Length=399


***** No hits found *****


Lambda      K        H        a         alpha
   0.318    0.135    0.410    0.792     4.96

Gapped
Lambda      K        H        a         alpha    sigma
   0.267   0.0410    0.140     1.90     42.6     43.6

Effective search space used: 5430585060


Query= gi|2347040|emb|CAA03909.1| Lefty protein [Mus musculus]

Length=368


***** No hits found *****


Lambda      K        H        a         alpha
   0.320    0.134    0.418    0.792     4.96

Gapped
Lambda      K        H        a         alpha    sigma
   0.267   0.0410    0.140     1.90     42.6     43.6

Effective search space used: 4891278997


  Database: Pomatcontigs500.fasta
    Posted date:  Feb 7, 2014  2:01 AM
  Number of letters in database: 75,827,742
  Number of sequences in database:  63,599



Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Neighboring words threshold: 13
Window for multiple hits: 40

TBLASTN 2.2.27+


Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro A.
Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J.
Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs", Nucleic Acids Res. 25:3389-3402.



Database: abyss61-scaffoldstrim.fa
           310,316 sequences; 607,381,932 total letters
```

```
Query= gi|167859062|gb|ACA04466.1| lefty [Strongylocentrotus purpuratus]

Length=399


***** No hits found *****


Lambda      K        H        a         alpha
   0.318    0.135    0.410    0.792     4.96

Gapped
Lambda      K        H        a         alpha    sigma
   0.267   0.0410    0.140     1.90     42.6     43.6

Effective search space used: 45845649424


Query= gi|2347040|emb|CAA03909.1| Lefty protein [Mus musculus]

Length=368


***** No hits found *****


Lambda      K        H        a         alpha
   0.320    0.134    0.418    0.792     4.96

Gapped
Lambda      K        H        a         alpha    sigma
   0.267   0.0410    0.140     1.90     42.6     43.6

Effective search space used: 40733364776


  Database: abyss61-scaffoldstrim.fa
    Posted date:  Feb 7, 2014  2:49 AM
  Number of letters in database: 607,381,932
  Number of sequences in database:  310,316



Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Neighboring words threshold: 13
Window for multiple hits: 40



TBLASTN 2.2.27+


Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro A.
Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J.
Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs", Nucleic Acids Res. 25:3389-3402.



Database: 200plus200windownonred.fa
           232,330 sequences; 162,337,606 total letters


Query= gi|167859062|gb|ACA04466.1| lefty [Strongylocentrotus purpuratus]

Length=399


***** No hits found *****


Lambda      K        H        a         alpha
   0.318    0.135    0.410    0.792     4.96
```

```
Gapped
Lambda      K        H        a         alpha    sigma
   0.267   0.0410    0.140     1.90     42.6     43.6

Effective search space used: 8348683850


Query= gi|2347040|emb|CAA03909.1| Lefty protein [Mus musculus]

Length=368


***** No hits found *****



Lambda      K        H        a         alpha
   0.320    0.134    0.418    0.792     4.96

Gapped
Lambda      K        H        a         alpha    sigma
   0.267   0.0410    0.140     1.90     42.6     43.6

Effective search space used: 7456238335


  Database: 200plus200windownonred.fa
    Posted date:  Feb 7, 2014  4:55 AM
  Number of letters in database: 162,337,606
  Number of sequences in database:  232,330



Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Neighboring words threshold: 13
Window for multiple hits: 40
```

```
TBLASTN 2.2.27+


Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro A.
Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J.
Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs", Nucleic Acids Res. 25:3389-3402.



Database: abyss61-scaffoldstrim.fa
           310,316 sequences; 607,381,932 total letters



Query= gi|190576701|gb|ACE79158.1| winged helix/forkhead transcription
factor FoxH [Branchiostoma floridae]

Length=523
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

  7527519 3379 77835 251177+,...,3366549-                             106    9e-23
  8012618 8166 173447 7732053-,124N,7913320-                          98.2    5e-20
  7618526 2682 63681 2688144-,...,5039994-                            95.5    3e-19
  7630192 4890 83672 3106671+,...,4409103+                            94.7    7e-19
  4257531 3757 46166                                                  91.7    6e-18
  7731395 5227 66068 1062555+,7608726-,5772291-                       91.3    8e-18
  8006371 15120 196115 7703625-,110N,7871886+                         91.7    8e-18
  7567370 12658 159222 1205245+,...,1961569+                          90.9    1e-17
  8000780 3318 65995 7570868+,206N,7926430-                           90.5    1e-17
  7992002 5136 78656 5250086+,469N,7962848+                           90.5    2e-17
  7732903 4361 73828 1111724+,3586420+,6736938-                       84.0    2e-15
  8034524 4532 87611 7886927-,7931173-                                82.8    4e-15
  7968785 1533 19788 7672096-,1186732+,7544432+                       79.3    2e-14
  7661341 4519 70081 4490050-,...,6794340-                            76.3    4e-13
  8021194 7816 179275 7777470-,110N,7892722-                          71.2    2e-11
  7792175 1997 52624 3058566+,...,2719362-                            67.8    1e-10
  8032779 18784 296728 7867271+,...,7911134+                          68.2    1e-10
  4378817 5786 78337                                                  67.4    3e-10
  8017211 13869 395257 7755161-,...,7852083+                          67.4    3e-10
  8015547 5089 60318 7746431+,130N,7900220+                           56.6    5e-07


> 7527519 3379 77835 251177+,...,3366549-
Length=3379

 Score =  106 bits (265),  Expect = 9e-23, Method: Compositional matrix adjust.
 Identities = 46/82 (56%), Positives = 61/82 (74%), Gaps = 1/82 (1%)
 Frame = +1

Query  84    QRYPKPPYSYLALVVMAIQNAPEKKLPLKEIHEALKKMYPFFRGDYTGWKDSVRHNLSTY  143
             +R  KPPYSY+AL+VMAIQ  P ++  L EI++ L++ +PFFRG Y GWK+SVRHNLS
Sbjct  2260  RRSEKPPYSYIALIVMAIQANPTRRCTLSEIYQYLQQKFPFFRGTYQGWKNSVRHNLSLN  2439

Query  144   KCFYKVPKDPSRPFAKGNYWAV  165
             +CF K+PK   RP  KG+YW +
Sbjct  2440  ECFIKLPKGIGRP-GKGHYWTI  2502


> 8012618 8166 173447 7732053-,124N,7913320-
Length=8166

 Score = 98.2 bits (243),  Expect = 5e-20, Method: Compositional matrix adjust.
 Identities = 46/114 (40%), Positives = 69/114 (61%), Gaps = 10/114 (9%)
 Frame = -2

Query  55    LPAKRKKNAANVRWKNYRRTDEVEGKRKYQRYPKPPYSYLALVVMAIQNAPEKKLPLKEI  114
             +   R +     R KNYRR+           + KPPYSY++L+ MAIQ +P K   L EI
Sbjct  1106  MSMDRAQALNRARDKNYRRS---------YTHAKPPYSYISLITMAIQQSPNKMCTLSEI  954

Query  115   HEALKKMYPFFRGDYTGWKDSVRHNLSTYKCFYKVPKDPSRPFAKGNYWAVYED  168
             ++ +  ++PF+R +   W++S+RH+LS   CF KVP+ P RP  KG+YWA++ D
Sbjct  953   YQFIMDLFPFYRQNQQRWQNSIRHSLSFNDCFVKVPRTPDRP-GKGSYWALHPD  795


> 7618526 2682 63681 2688144-,...,5039994-
Length=2682
```

```
 Score = 95.5 bits (236),  Expect = 3e-19, Method: Compositional matrix adjust.
 Identities = 40/87 (46%), Positives = 63/87 (72%), Gaps = 1/87 (1%)
 Frame = +1

Query  79    GKRKYQRYPKPPYSYLALVVMAIQNAPEKKLPLKEIHEALKKMYPFFRGDYTGWKDSVRH  138
             G ++ ++  KPP+SY AL++MAI+ +PEK+L L +I+E + K +P+++ +  GW++S+RH
Sbjct  1273  GSKEEKKAEKPPFSYNALIMMAIRGSPEKRLTLSQIYEFIMKNFPYYKDNKQGWQNSIRH  1452

Query  139   NLSTYKCFYKVPKDPSRPFAKGNYWAV  165
             NLS  KCF KVP+    P  KGNYW +
Sbjct  1453  NLSLNKCFLKVPRHYDDP-GKGNYWML  1530


> 7630192 4890 83672 3106671+,...,4409103+
Length=4890

 Score = 94.7 bits (234),  Expect = 7e-19, Method: Compositional matrix adjust.
 Identities = 38/78 (49%), Positives = 58/78 (74%), Gaps = 1/78 (1%)
 Frame = -3

Query  88    KPPYSYLALVVMAIQNAPEKKLPLKEIHEALKKMYPFFRGDYTGWKDSVRHNLSTYKCFY  147
             KPPYSY+AL+ MAIQ++PEK++ L  I+  +   +PF+R +  GW++S+RHNLS  +CF
Sbjct  2371  KPPYSYIALIAMAIQSSPEKRVTLNGIYAFIMDRFPFYRENKQGWQNSIRHNLSLNECFM  2192

Query  148   KVPKDPSRPFAKGNYWAV  165
             K+P+D  +P  KG+YW +
Sbjct  2191  KIPRDDKKP-GKGSYWTL  2141


> 4257531 3757 46166
Length=3757

 Score = 91.7 bits (226),  Expect = 6e-18, Method: Compositional matrix adjust.
 Identities = 36/89 (40%), Positives = 63/89 (71%), Gaps = 1/89 (1%)
 Frame = -3

Query  77    VEGKRKYQRYPKPPYSYLALVVMAIQNAPEKKLPLKEIHEALKKMYPFFRGDYTGWKDSV  136
             +E  ++ ++  KPPYSY+AL+ MA++ AP++K+ L  I++ + + +P++  +  GW++S+
Sbjct  1802  IEDLQRREQPQKPPYSYIALIAMAVKAAPDRKVTLNGIYQFIMERFPYYHDNKQGWQNSI  1623

Query  137   RHNLSTYKCFYKVPKDPSRPFAKGNYWAV  165
             RHNLS   CF KVP++  +P  KGNYW +
Sbjct  1622  RHNLSLNDCFVKVPREKGKP-GKGNYWTL  1539


> 7731395 5227 66068 1062555+,7608726-,5772291-
Length=5227

 Score = 91.3 bits (225),  Expect = 8e-18, Method: Compositional matrix adjust.
 Identities = 40/79 (51%), Positives = 57/79 (72%), Gaps = 1/79 (1%)
 Frame = -2

Query  88    KPPYSYLALVVMAIQNAPEKKLPLKEIHEALKKMYPFFRGDYTGWKDSVRHNLSTYKCFY  147
             KPPYSY+AL  MAIQ + EK LPL +I++ +   +PF+R +   W++S+RHNLS   CF
Sbjct  3210  KPPYSYIALTAMAIQASGEKMLPLSDIYKFIMDNFPFYRKNTQRWQNSLRHNLSFNDCFI  3031

Query  148   KVPKDPSRPFAKGNYWAVY  166
             K+P+ P RP  KG+YWA++
Sbjct  3030  KIPRRPDRP-GKGSYWALH  2977


> 8006371 15120 196115 7703625-,110N,7871886+
Length=15120

 Score = 91.7 bits (226),  Expect = 8e-18, Method: Compositional matrix adjust.
 Identities = 36/79 (46%), Positives = 56/79 (71%), Gaps = 1/79 (1%)
 Frame = -3

Query  88     KPPYSYLALVVMAIQNAPEKKLPLKEIHEALKKMYPFFRGDYTGWKDSVRHNLSTYKCFY  147
              KPPYSY+AL+ MA++++P   + L EI+  +   +P+F+ +   W++S+RHNLS   CF
Sbjct  10870  KPPYSYIALITMAVESSPHGMMTLNEIYAFIMNRFPYFKQNQQRWQNSIRHNLSLNDCFI  10691

Query  148    KVPKDPSRPFAKGNYWAVY  166
              KVP+ P RP  KGNYW+++
Sbjct  10690  KVPRGPGRP-GKGNYWSLH  10637


> 7567370 12658 159222 1205245+,...,1961569+
Length=12658
```

```
 Score = 90.9 bits (224),  Expect = 1e-17, Method: Compositional matrix adjust.
 Identities = 38/88 (43%), Positives = 58/88 (66%), Gaps = 1/88 (1%)
 Frame = +1

Query  78    EGKRKYQRYPKPPYSYLALVVMAIQNAPEKKLPLKEIHEALKKMYPFFRGDYTGWKDSVR  137
             +  +K Q   KPPYSY+AL+ M+I  +P K+L L  I E +   +P++R  +  W++S+R
Sbjct  4912  DALKKKQNLVKPPYSYIALITMSILQSPRKRLTLSGICEFIMTRFPYYREKFPAWQNSIR  5091

Query  138   HNLSTYKCFYKVPKDPSRPFAKGNYWAV  165
             HNLS   CF K+P++P  P  KGNYW++
Sbjct  5092  HNLSLNDCFVKIPREPGNP-GKGNYWSL  5172


> 8000780 3318 65995 7570868+,206N,7926430-
Length=3318

 Score = 90.5 bits (223),  Expect = 1e-17, Method: Compositional matrix adjust.
 Identities = 37/82 (45%), Positives = 60/82 (73%), Gaps = 1/82 (1%)
 Frame = -1

Query  84    QRYPKPPYSYLALVVMAIQNAPEKKLPLKEIHEALKKMYPFFRGDYTGWKDSVRHNLSTY  143
             ++  KPP+SY AL++MAI+++ EK++ L +I+E + K +P+++ +  GW++S+RHNLS
Sbjct  2388  KKNDKPPFSYNALIMMAIRSSHEKRMTLSQIYEFIMKNFPYYKDNKQGWQNSIRHNLSLN  2209

Query  144   KCFYKVPKDPSRPFAKGNYWAV  165
             KCF KVP+    P  KGNYW +
Sbjct  2208  KCFLKVPRHYDDP-GKGNYWML  2146


> 7992002 5136 78656 5250086+,469N,7962848+
Length=5136

 Score = 90.5 bits (223),  Expect = 2e-17, Method: Compositional matrix adjust.
 Identities = 42/84 (50%), Positives = 58/84 (69%), Gaps = 4/84 (5%)
 Frame = -2

Query  72    RRTDEVEGKR---KYQRYPKPPYSYLALVVMAIQNAPEKKLPLKEIHEALKKMYPFFRGD  128
             RR  E+ G+R    Y+R+ KPPYSY+ ++ + I+ APE KL L  I  A K ++PFF+G+
Sbjct  4628  RRVSEL-GRRTVKNYKRHGKPPYSYVGMIALLIRMAPEHKLSLAGILAAFKDLFPFFQGE  4452

Query  129   YTGWKDSVRHNLSTYKCFYKVPKD  152
             Y GW+DSVRHNLS   CFY V ++
Sbjct  4451  YQGWRDSVRHNLSHNDCFYMVRQN  4380


> 7732903 4361 73828 1111724+,3586420+,6736938-
Length=4361

 Score = 84.0 bits (206),  Expect = 2e-15, Method: Compositional matrix adjust.
 Identities = 39/88 (44%), Positives = 57/88 (65%), Gaps = 4/88 (5%)
 Frame = +1

Query  88    KPPYSYLALVVMAIQNAPEKKLPLKEIHEALKKMYPFFRGDYTGWKDSVRHNLSTYKCFY  147
             KPPYSY+AL+ MAI+ + EK+L L  I+  +   +P++  +  GW++S+RHNLS  +CF
Sbjct  1507  KPPYSYVALIAMAIKESTEKRLTLSGIYNFIVAKFPYYEKNKKGWQNSIRHNLSLNECFV  1686

Query  148   KVPKDPSRPFAKGNYWAV---YEDKVPK  172
             KVP++      KGNYW V   +ED   K
Sbjct  1687  KVPREGGGE-RKGNYWTVDPAFEDMFEK  1767


> 8034524 4532 87611 7886927-,7931173-
Length=4532

 Score = 82.8 bits (203),  Expect = 4e-15, Method: Compositional matrix adjust.
 Identities = 42/98 (43%), Positives = 62/98 (63%), Gaps = 11/98 (11%)
 Frame = +1

Query  79    GKRKYQR---YPKPPYSYLALVVMAIQNAPEKKLPLKEIHEALKKMYPFFRG--DYTGWK  133
             G RK +R     KPPYSY+AL+ M I N+PE+KL L EI+  + + +P++R   +  GW+
Sbjct  1168  GSRKRKRPIPKGKPPYSYIALISMGIANSPERKLTLHEIYSFITERFPYYRDHPNSKGWR  1347

Query  134   DSVRHNLSTYKCFYKVPKDPSRPFAKGNYWAV---YED  168
              S+RHNL+   CF K+ +   +P  KG+ WA+   YED
Sbjct  1348  GSIRHNLALNDCFVKLDR---KPGMKGHQWAIDPDYED  1452


> 7968785 1533 19788 7672096-,1186732+,7544432+
```

```
Length=1533

 Score = 79.3 bits (194),  Expect = 2e-14, Method: Compositional matrix adjust.
 Identities = 33/65 (51%), Positives = 50/65 (77%), Gaps = 0/65 (0%)
 Frame = +2

Query  88   KPPYSYLALVVMAIQNAPEKKLPLKEIHEALKKMYPFFRGDYTGWKDSVRHNLSTYKCFY  147
            KPPYS+  L+ MA++++P+K+LP+K+I+  +   +P+F+   TGWK+SVRHNLS  KCF
Sbjct  722  KPPYSFSCLIFMAVEDSPQKRLPVKDIYNWILSHFPYFQNAPTGWKNSVRHNLSLNKCFK  901

Query  148  KVPKD  152
            KV K+
Sbjct  902  KVDKE  916


> 7661341 4519 70081 4490050-,...,6794340-
Length=4519

 Score = 76.3 bits (186),  Expect = 4e-13, Method: Compositional matrix adjust.
 Identities = 36/86 (42%), Positives = 52/86 (60%), Gaps = 3/86 (3%)
 Frame = -3

Query  81    RKYQRYPKPPYSYLALVVMAIQNAPEKKLPLKEIHEALKKMYPFFRGDYTGWKDSVRHNL  140
             R  Q  PKP  SY+ L+ MAI  A +KKL L +I++ +   Y +FR    GW++S+RHNL
Sbjct  3251  RYIQEEPKPSQSYIGLISMAILGAKDKKLILSDIYQWILDNYAYFRTRGPGWRNSIRHNL  3072

Query  141   STYKCFYKVPKDPSRPFAKGNYWAVY  166
             S   CF K  +  +    KG+YWA++
Sbjct  3071  SLNDCFIKSGRSAN---GKGHYWAIH  3003


> 8021194 7816 179275 7777470-,110N,7892722-
Length=7816

 Score = 71.2 bits (173),  Expect = 2e-11, Method: Compositional matrix adjust.
 Identities = 32/79 (41%), Positives = 48/79 (61%), Gaps = 3/79 (4%)
 Frame = +3

Query  88    KPPYSYLALVVMAIQNAPEKKLPLKEIHEALKKMYPFFRGDYTGWKDSVRHNLSTYKCFY  147
             KP  SY+AL+  AI  + EK+L L  I+  ++  +P++     GW++SVRHNLS  +CF
Sbjct  6831  KPSMSYIALIGKAILESSEKRLNLGSIYSWIESKFPYYVNRGQGWRNSVRHNLSLNECFI  7010

Query  148   KVPKDPSRPFAKGNYWAVY  166
             K  +       KGNYWA++
Sbjct  7011  KAGRCED---GKGNYWAIH  7058


> 7792175 1997 52624 3058566+,...,2719362-
Length=1997

 Score = 67.8 bits (164),  Expect = 1e-10, Method: Compositional matrix adjust.
 Identities = 27/76 (36%), Positives = 50/76 (66%), Gaps = 3/76 (4%)
 Frame = -3

Query  91   YSYLALVVMAIQNAPEKKLPLKEIHEALKKMYPFFRGDYTGWKDSVRHNLSTYKCFYKVP  150
            +SY+AL+ M I N+PEKK+ L +I++ +   +P++  +   W++S+RHNLS  +CF K
Sbjct  639  HSYIALISMXILNSPEKKVLLGDIYQYIMDNFPYYNNEVKAWRNSIRHNLSLNECFIKSR  460

Query  151  KDPSRPFAKGNYWAVY  166
            +  +    K N+W++
Sbjct  459  RSDN---GKHNFWSIH  421


> 8032779 18784 296728 7867271+,...,7911134+
Length=18784

 Score = 68.2 bits (165),  Expect = 1e-10, Method: Compositional matrix adjust.
 Identities = 27/79 (34%), Positives = 51/79 (65%), Gaps = 3/79 (4%)
 Frame = +3

Query  88    KPPYSYLALVVMAIQNAPEKKLPLKEIHEALKKMYPFFRGDYTGWKDSVRHNLSTYKCFY  147
             KP  SY+ L+ MAIQ++ ++K+ L +I++ +   +P+++   W++SVRHNLS  +CF
Sbjct  2694  KPSLSYIGLISMAIQSSSQRKMLLSDIYQWIVNNFPYYKMEDRSWRNSVRHNLSLNECFI  2873

Query  148   KVPKDPSRPFAKGNYWAVY  166
             K  +  +    K +YW ++
Sbjct  2874  KCGRSEN---GKSHYWTIH  2921
```

```
> 4378817 5786 78337
Length=5786

 Score = 67.4 bits (163),  Expect = 3e-10, Method: Compositional matrix adjust.
 Identities = 30/78 (38%), Positives = 48/78 (62%), Gaps = 1/78 (1%)
 Frame = -1

Query  88    KPPYSYLALVVMAIQNAPEKKLPLKEIHEALKKMYPFFRGDYTGWKDSVRHNLSTYKCFY  147
             KPP+SY  L+ MAI    EKK+ +  I++ +   Y +++   + WK+S+RHNLS  K F
Sbjct  2228  KPPHSYATLISMAINETKEKKINISSIYKWITDKYSYYKMADSHWKNSIRHNLSLNKRFE  2049

Query  148   KVPKDPSRPFAKGNYWAV  165
             K+P++ +    KG YW +
Sbjct  2048  KIPREQNES-NKGGYWRI  1998


> 8017211 13869 395257 7755161-,...,7852083+
Length=13869

 Score = 67.4 bits (163),  Expect = 3e-10, Method: Compositional matrix adjust.
 Identities = 34/90 (38%), Positives = 52/90 (58%), Gaps = 2/90 (2%)
 Frame = -1

Query  77    VEGKR-KYQRYPKPPYSYLALVVMAIQNAPEKKLPLKEIHEALKKMYPFFRGDYTGWKDS  135
             +E K  K     KPPYSY  L+ MAI    EKK+ +  I++ +   + +++   + WK+S
Sbjct  5046  IENKNYKNNSCIKPPYSYATLISMAINETKEKKINISSIYKWITDNFSYYKMADSHWKNS  4867

Query  136   VRHNLSTYKCFYKVPKDPSRPFAKGNYWAV  165
             +RHNLS  K F KV + P+   +KG YW +
Sbjct  4866  IRHNLSLNKRFEKVSRLPNES-SKGGYWKI  4780



Lambda      K        H        a         alpha
   0.314    0.131    0.402    0.792     4.96

Gapped
Lambda      K        H        a         alpha    sigma
   0.267   0.0410    0.140     1.90     42.6     43.6

Effective search space used: 65428602540


  Database: abyss61-scaffoldstrim.fa
    Posted date:  Feb 7, 2014  2:49 AM
  Number of letters in database: 607,381,932
  Number of sequences in database:  310,316



Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Neighboring words threshold: 13
Window for multiple hits: 40

TBLASTN 2.2.27+


Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro A.
Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J.
Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs", Nucleic Acids Res. 25:3389-3402.



Database: 200plus200windownonred.fa
           232,330 sequences; 162,337,606 total letters



Query= gi|190576701|gb|ACE79158.1| winged helix/forkhead transcription
factor FoxH [Branchiostoma floridae]

Length=523
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

  NODE_265519_length_13224_cov_52.508545                              92.4    1e-18
  NODE_440046_length_40245_cov_35.859535                              84.3    4e-16
```

```
  NODE_2802840_length_447_cov_47.935123                                61.2    3e-10


> NODE_265519_length_13224_cov_52.508545
Length=13284

 Score = 92.4 bits (228),  Expect = 1e-18, Method: Compositional matrix adjust.
 Identities = 49/166 (30%), Positives = 89/166 (54%), Gaps = 1/166 (1%)
 Frame = -2

Query  1     MVSESLPTDADLQGSTPPCLTRMDSPTSQPPSKRPRVatdtdssetassGKRASLPAKRK  60
             +++ S P  A      PP      S T QP  K+ +   +   S+ + +  + ++
Sbjct  3041  LMAHSQPAQAYQPAFNPPTFNGTKSHTQQPNKKKIKKDKNRLKSQISEAEAKQAIEKGLV  2862

Query  61    KNAANVRWKNYRRTDEVEGKRKYQRYPKPPYSYLALVVMAIQNAPEKKLPLKEIHEALKK  120
                 +V   ++  T  +  +++     KPPYSY+AL+ MAI+++P   + L EI++ ++
Sbjct  2861  TEVDHVDLSHFSGTSTISQQKRRFAEVKPPYSYIALITMAIESSPTGMMTLNEIYQFIEN  2682

Query  121   MYPFFRGDYTGWKDSVRHNLSTYKCFYKVPKDPSRPFAKGNYWAVY  166
              +P+F+ +   W++S+RHNLS   CF KV K+  +P  KGNYWA++
Sbjct  2681  RFPYFKENTQRWQNSIRHNLSLNDCFLKVSKNAGKP-GKGNYWALH  2547


> NODE_440046_length_40245_cov_35.859535
Length=40305

 Score = 84.3 bits (207),  Expect = 4e-16, Method: Compositional matrix adjust.
 Identities = 42/100 (42%), Positives = 64/100 (64%), Gaps = 11/100 (11%)
 Frame = -1

Query  77    VEGKRKYQRYP----------KPPYSYLALVVMAIQNAPEKKLPLKEIHEALKKMYPFFR  126
             V+ +RK++  P          KPPYSY+AL  MAIQ + EK LPL +I++ +   +P++R
Sbjct  7275  VDFRRKFKIMPRPGKATYGDEKPPYSYIALTAMAIQASNEKMLPLSDIYKFIMDKFPYYR  7096

Query  127   GDYTGWKDSVRHNLSTYKCFYKVPKDPSRPFAKGNYWAVY  166
              +   W++S+RHNLS   CF K+P+   RP  KG+YWA++
Sbjct  7095  KNTQKWQNSLRHNLSFNDCFIKIPRRQDRP-GKGSYWALH  6979


> NODE_2802840_length_447_cov_47.935123
Length=467

 Score = 61.2 bits (147),  Expect = 3e-10, Method: Compositional matrix adjust.
 Identities = 30/82 (37%), Positives = 54/82 (66%), Gaps = 4/82 (5%)
 Frame = +3

Query  88   KPPYSYLALVVMAIQNAPEKKLPLKEIHEALKKMYPFFRGDYTG-WKDSVRHNLSTYKCF  146
            KPPY++  L+ +AI+++  K+L +KEI++ +     +++   +G WK+S+R NLS+ +CF
Sbjct  42   KPPYTFSCLIFLAIESSCNKRLCVKEINQWIVDNIAYYKNVPSGSWKNSIRFNLSSNQCF  221

Query  147  YKVPKD--PSRPFA-KGNYWAV  165
             KV K+    R F+ KG+ W +
Sbjct  222  SKVDKNLLTMRDFSGKGSLWCI  287



Lambda      K        H        a         alpha
   0.314    0.131    0.402    0.792     4.96

Gapped
Lambda      K        H        a         alpha    sigma
   0.267   0.0410    0.140     1.90     42.6     43.6

Effective search space used: 11545637325


  Database: 200plus200windownonred.fa
    Posted date:  Feb 7, 2014  4:55 AM
  Number of letters in database: 162,337,606
  Number of sequences in database:  232,330



Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Neighboring words threshold: 13
Window for multiple hits: 40
```

```
BLASTX 2.2.29+
Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro
A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and
David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new
generation of protein database search programs", Nucleic
Acids Res. 25:3389-3402.


RID: F8YAX2SF014


Database: All non-redundant GenBank CDS
translations+PDB+SwissProt+PIR+PRF excluding environmental samples
from WGS projects
           36,556,296 sequences; 12,906,967,781 total letters
Query= 7527519 3379 77835 251177+,...,3366549-

Length=1767


                                                                    Score     E
Sequences producing significant alignments:                        (Bits)  Value

gb|ESO83035.1|  hypothetical protein LOTGIDRAFT_134143 [Lottia...    110   2e-25
ref|XP_005912155.1|  PREDICTED: forkhead box protein H1-like i...    107   2e-21
ref|XP_005722570.1|  PREDICTED: forkhead box protein H1-like i...    107   3e-21
ref|XP_004546619.1|  PREDICTED: forkhead box protein H1-like i...    106   3e-21
emb|CBN81873.1|  Forkhead box protein H1 [Dicentrarchus labrax]      105   1e-20
ref|XP_003443542.1|  PREDICTED: forkhead box protein H1-like [...    105   1e-20
ref|NP_001017084.1|  forkhead box protein H1 [Xenopus (Siluran...    104   2e-20
ref|XP_005534230.1|  PREDICTED: forkhead box protein H1-like, ...  99.4   5e-20
gb|ACE79158.1|  winged helix/forkhead transcription factor Fox...    100   4e-19
gb|ACE79141.1|  winged helix/forkhead transcription factor Fox...    100   4e-19

ALIGNMENTS
>gb|ESO83035.1| hypothetical protein LOTGIDRAFT_134143 [Lottia gigantea]
Length=72

 Score =   110 bits (274),  Expect = 2e-25, Method: Compositional matrix adjust.
 Identities = 44/70 (63%), Positives = 60/70 (86%), Gaps = 0/70 (0%)
 Frame = +1

Query  64   VKNYKRHGKPPYSYVGMIALLIRMAPEHKLSLAGILAAFKDLFPFFQGEYQGWRDSVRHN  243
            +K+YKRH KPPYSY+GM+AL+I+ +P  + SLAGI+    D+FPFFQGEY+GW+DSVRHN
Sbjct  1    MKSYKRHDKPPYSYLGMVALIIQCSPGRQQSLAGIIDTLTDMFPFFQGEYKGWKDSVRHN  60

Query  244  LSHNDCFYMV  273
            ++++DCFY V
Sbjct  61   MTNSDCFYKV  70


>ref|XP_005912155.1| PREDICTED: forkhead box protein H1-like isoform X1 [Haplochromis
burtoni]
 ref|XP_005912156.1| PREDICTED: forkhead box protein H1-like isoform X2 [Haplochromis
burtoni]
Length=514

 Score =   107 bits (267),  Expect = 2e-21, Method: Compositional matrix adjust.
 Identities = 82/237 (35%), Positives = 112/237 (47%), Gaps = 51/237 (22%)
 Frame = +1

Query  31   RRRVSELGRRTVKNYKRHGKPPYSYVGMIALLIRMAPEHKLSLAGILAAFKDLFPFFQGE  210
             R     G    KNY+R+ KPPYSY+ MIA++I+ +PE KL+LA IL     LFPFF+G
Sbjct  114  EREKRNAGNGKKKNYQRYPKPPYSYLAMIAMVIQRSPEKKLTLAEILKEISTLFPFFKGN  173

Query  211  YQGWRDSVRHNLSHNDCFYMVEVHKVLRIYIKVAPRPRLAFMNIYLLQNLSDPFPLQVPQ  390
            Y+GWRDSVRHNLS  DCF  V                                L DP     PQ
Sbjct  174  YKGWRDSVRHNLSSYDCFVKV-----------------------------LKDP---GKPQ  202

Query  391  VINNVVKTKKCDWAVHLNKLPSDALRKQEKSSKDTDYQYQYASSLTEHFGLPQIEYLLPC  570
               N         WAV L+++P + L++Q  +    D +  +A  L          Y++
Sbjct  203  GKGNF-------WAVELSRIPVELLKRQNTAVSRQD-ETIFAQDLA--------PYIMQG  246

Query  571  MKLQPRTLPLHGITPLKASTSQVFKTPQSTRALKK-KRSFTIDSLLNSKDLKPVSST  738
             K +    P   +T L  ++S     PQ      K   SF IDSLL+S  L+P+S++
Sbjct  247  HKPE-SEAPPEPVTSLPPTSSGNSFPPQDDLYRPKLDSSFAIDSLLHS--LRPISAS  300


>ref|XP_005722570.1| PREDICTED: forkhead box protein H1-like isoform X1 [Pundamilia
```

```
nyererei]
 ref|XP_005722571.1| PREDICTED: forkhead box protein H1-like isoform X2 [Pundamilia
nyererei]
Length=514

 Score =   107 bits (266),  Expect = 3e-21, Method: Compositional matrix adjust.
 Identities = 82/237 (35%), Positives = 112/237 (47%), Gaps = 51/237 (22%)
 Frame = +1

Query  31   RRRVSELGRRTVKNYKRHGKPPYSYVGMIALLIRMAPEHKLSLAGILAAFKDLFPFFQGE  210
            R     G    KNY+R+ KPPYSY+ MIA++I+ +PE KL+LA IL     LFPFF+G
Sbjct  114  EREKRNAGNGKKKNYQRYPKPPYSYLAMIAMVIQRSPEKKLTLAEILKEISTLFPFFKGN  173

Query  211  YQGWRDSVRHNLSHNDCFYMVEVHKVLRIYIKVAPRPRLAFMNIYLLQNLSDPFPLQVPQ  390
            Y+GWRDSVRHNLS  DCF  V                            L DP     PQ
Sbjct  174  YKGWRDSVRHNLSSYDCFVKV----------------------------LKDP---GKPQ  202

Query  391  VINNVVKTKKCDWAVHLNKLPSDALRKQEKSSKDTDYQYQYASSLTEHFGLPQIEYLLPC  570
               N        WAV L+++P + L++Q  +    D +  +A  L          Y++
Sbjct  203  GKGNF-------WAVELSRIPLELLKRQNTAVSRQD-ETIFAQDLA--------PYIMQG  246

Query  571  MKLQPRTLPLHGITPLKASTSQVFKTPQSTRALKK-KRSFTIDSLLNSKDLKPVSST  738
             K +    P   +T L  ++S     PQ      K   SF IDSLL+S  L+P+S++
Sbjct  247  HKPE-SEAPPEPVTSLPPTSSGNSFPPQDDLYRPKLDSSFAIDSLLHS--LRPISAS  300


>ref|XP_004546619.1| PREDICTED: forkhead box protein H1-like isoform X1 [Maylandia
zebra]
 ref|XP_004546620.1| PREDICTED: forkhead box protein H1-like isoform X2 [Maylandia
zebra]
Length=514

 Score =   106 bits (265),  Expect = 3e-21, Method: Compositional matrix adjust.
 Identities = 82/237 (35%), Positives = 112/237 (47%), Gaps = 51/237 (22%)
 Frame = +1

Query  31   RRRVSELGRRTVKNYKRHGKPPYSYVGMIALLIRMAPEHKLSLAGILAAFKDLFPFFQGE  210
            R     G    KNY+R+ KPPYSY+ MIA++I+ +PE KL+LA IL     LFPFF+G
Sbjct  114  EREKRNAGNGKKKNYQRYPKPPYSYLAMIAMVIQRSPEKKLTLAEILKEISTLFPFFKGN  173

Query  211  YQGWRDSVRHNLSHNDCFYMVEVHKVLRIYIKVAPRPRLAFMNIYLLQNLSDPFPLQVPQ  390
            Y+GWRDSVRHNLS  DCF  V                            L DP     PQ
Sbjct  174  YKGWRDSVRHNLSSYDCFVKV----------------------------LKDP---GKPQ  202

Query  391  VINNVVKTKKCDWAVHLNKLPSDALRKQEKSSKDTDYQYQYASSLTEHFGLPQIEYLLPC  570
               N        WAV L+++P + L++Q  +    D +  +A  L          Y++
Sbjct  203  GKGNF-------WAVELSRIPLELLKRQNTAVSRQD-ETIFAQDLA--------PYIMQG  246

Query  571  MKLQPRTLPLHGITPLKASTSQVFKTPQSTRALKK-KRSFTIDSLLNSKDLKPVSST  738
             K +    P   +T L  ++S     PQ      K   SF IDSLL+S  L+P+S++
Sbjct  247  HKPE-SEAPPEPVTSLPPTSSGNSFPPQDDLYRPKLDSSFAIDSLLHS--LRPISAS  300


>emb|CBN81873.1| Forkhead box protein H1 [Dicentrarchus labrax]
Length=509

 Score =   105 bits (261),  Expect = 1e-20, Method: Compositional matrix adjust.
 Identities = 86/250 (34%), Positives = 118/250 (47%), Gaps = 57/250 (23%)
 Frame = +1

Query  25   AKRRRVSELGRRTVKNYKRHGKPPYSYVGMIALLIRMAPEHKLSLAGILAAFKDLFPFFQ  204
            ++R + S  G +  KNY+R+ KPPYSY+ MIA++I+ +PE KL+L+ IL     LFPFF+
Sbjct  104  SEREKSSNCGGKK-KNYQRYPKPPYSYLAMIAMVIQRSPEKKLTLSEILKEISTLFPFFK  162

Query  205  GEYQGWRDSVRHNLSHNDCFYMVEVHKVLRIYIKVAPRPRLAFMNIYLLQNLSDPFPLQV  384
            G Y+GWRDSVRHNLS  DCF  V                              L DP
Sbjct  163  GNYKGWRDSVRHNLSSYDCFVKV----------------------------LKDP---GK  191

Query  385  PQVINNVVKTKKCDWAVHLNKLPSDALRKQEKSSKDTDYQYQYASSLTEHFGLPQIEYLL  564
            PQ   N        WAV L+++P + L++Q  +    D +  +A  L          Y+L
Sbjct  192  PQGKGNF-------WAVELSRVPLELLKRQNTAVSRQD-ETIFAQDLA--------PYIL  235

Query  565  PCMKLQPRTLPLHGITPLKASTSQVFKTPQSTRALKK-KRSFTIDSLLNSKDLKPVSSTI  741
               K +    P   + PL    S+     PQ      K   SF IDSLL+S  L+P
Sbjct  236  QGHKPESEPPPA-SVNPLPPMRSRNPSPPQEDLFRPKLDSSFAIDSLLHS--LRP-----  287

Query  742  PTTASDALVS  771
            P+ + D  VS
Sbjct  288  PSASGDVDVS  297
```

```
>ref|XP_003443542.1| PREDICTED: forkhead box protein H1-like [Oreochromis niloticus]
Length=514

 Score =   105 bits (261),  Expect = 1e-20, Method: Compositional matrix adjust.
 Identities = 80/236 (34%), Positives = 111/236 (47%), Gaps = 49/236 (21%)
 Frame = +1

Query  31   RRRVSELGRRTVKNYKRHGKPPYSYVGMIALLIRMAPEHKLSLAGILAAFKDLFPFFQGE  210
            R     G    KNY+R+ KPPYSY+ MIA++I+ +PE KL+LA IL     LFPFF+G
Sbjct  114  EREKRNAGSGKKKNYQRYPKPPYSYLAMIAMVIQRSPEKKLTLAEILKEISTLFPFFKGN  173

Query  211  YQGWRDSVRHNLSHNDCFYMVEVHKVLRIYIKVAPRPRLAFMNIYLLQNLSDPFPLQVPQ  390
            Y+GWRDSVRHNLS  DCF  V                             L DP     PQ
Sbjct  174  YKGWRDSVRHNLSSYDCFLKV----------------------------LKDP---GKPQ  202

Query  391  VINNVVKTKKCDWAVHLNKLPSDALRKQEKSSKDTDYQYQYASSLTEHFGLPQIEYLLPC  570
               N        WAV L+++P + L++Q  +    D +  +A  L  +     ++   P
Sbjct  203  GKGNF-------WAVELSRIPLELLKRQNTAVSRQD-ETIFAQDLAPYI----MQGHNPE  250

Query  571  MKLQPRTLPLHGITPLKASTSQVFKTPQSTRALKKKRSFTIDSLLNSKDLKPVSST  738
             +  P   P+  + P   S+   F         K   SF IDSLL+S  L+P+S++
Sbjct  251  SEAPPE--PVTSLPP--KSSGNSFPMQDDLYRPKLDSSFAIDSLLHS--LRPISAS  300


>ref|NP_001017084.1| forkhead box protein H1 [Xenopus (Silurana) tropicalis]
 sp|Q28GC4.1|FOXH1_XENTR RecName: Full=Forkhead box protein H1; AltName: Full=Forkhead
activin signal transducer 1; Short=Fast-1
 emb|CAJ81979.1| forkhead box H1 [Xenopus (Silurana) tropicalis]
Length=515

 Score =   104 bits (260),  Expect = 2e-20, Method: Compositional matrix adjust.
 Identities = 78/224 (35%), Positives = 106/224 (47%), Gaps = 53/224 (24%)
 Frame = +1

Query  67   KNYKRHGKPPYSYVGMIALLIRMAPEHKLSLAGILAAFKDLFPFFQGEYQGWRDSVRHNL  246
            KNY R+ KPPYSY+ MIAL+I+ +PE +L L+ IL     LFPFF+G+Y GW+DS+RHNL
Sbjct  103  KNYHRYAKPPYSYLAMIALVIQNSPEKRLKLSQILKEVSTLFPFFKGDYMGWKDSIRHNL  162

Query  247  SHNDCFYMVEVHKVLRIYIKVAPRPRLAFMNIYLLQNLSDPFPLQVPQVINNVVKTKKCD  426
            S NDCF  V                                L DP     PQ   N
Sbjct  163  SSNDCFKKV----------------------------LKDP---GKPQAKGNF-------  184

Query  427  WAVHLNKLPSDALRKQEK--SSKDTDYQYQYASSLTEHFGLPQIEYLLPCMKLQPRTLPL  600
            W V ++++P DA++ Q    +   +DY  Q  +    H      +Y    +   P  +P
Sbjct  185  WTVDVSRIPLDAMKLQNTALTRGGSDYFVQDLAPYILH----NYKYEHNVVVYAPHHMPS  240

Query  601  HGITPLKASTSQVFKTP-QSTRALKKKRSFTIDSLLNSKDLKPV  729
            H      AS+  + + P Q+    K   SF IDSLLN  DL+ V
Sbjct  241  H------ASSLPLAEDPHQTNTGGKLNTSFMIDSLLN--DLQDV  276


>ref|XP_005534230.1| PREDICTED: forkhead box protein H1-like, partial [Pseudopodoces
humilis]
Length=232

 Score = 99.4 bits (246),  Expect = 5e-20, Method: Compositional matrix adjust.
 Identities = 46/89 (52%), Positives = 57/89 (64%), Gaps = 0/89 (0%)
 Frame = +1

Query  7    TRMEPEAKRRRVSELGRRTVKNYKRHGKPPYSYVGMIALLIRMAPEHKLSLAGILAAFKD  186
            +R +P    R  S       + Y+RH KPPYSY+ +IAL+IR AP  +L LA I+   +
Sbjct  80   SRSDPRCGSRCGSRSSPGRPQRYRRHPKPPYSYLALIALVIRAAPGRRLKLAQIIQQLRS  139

Query  187  LFPFFQGEYQGWRDSVRHNLSHNDCFYMV  273
            LFPFF G YQGW+DSVRHNLS N CF  V
Sbjct  140  LFPFFGGGYQGWKDSVRHNLSSNPCFAKV  168


>gb|ACE79158.1| winged helix/forkhead transcription factor FoxH [Branchiostoma
floridae]
Length=523

 Score =   100 bits (248),  Expect = 4e-19, Method: Compositional matrix adjust.
 Identities = 60/161 (37%), Positives = 84/161 (52%), Gaps = 45/161 (28%)
 Frame = +1

Query  34   RRVSEL-GRRTVKNYKRHGKPPYSYVGMIALLIRMAPEHKLSLAGILAAFKDLFPFFQGE  210
```

```
            RR  E+ G+R    Y+R+ KPPYSY+ ++ + I+ APE KL L  I  A K ++PFF+G+
Sbjct  72   RRTDEVEGKR---KYQRYPKPPYSYLALVVMAIQNAPEKKLPLKEIHEALKKMYPFFRGD  128

Query  211  YQGWRDSVRHNLSHNDCFYMVEVHKVLRIYIKVAPRPRLAFMNIYLLQNLSDPFPLQVPQ  390
            Y GW+DSVRHNLS   CFY V                          ++ S PF
Sbjct  129  YTGWKDSVRHNLSTYKCFYKVP-------------------------KDPSRPF------  157

Query  391  VINNVVKTKKCDWAVHLNKLPSDALRKQEKSSK---DTDYQ  504
                    K   WAV+ +K+P DAL+KQ+ SS+   + DY+
Sbjct  158  -------AKGNYWAVYEDKVPKDALKKQDSSSEREGEADYE  191


>gb|ACE79141.1| winged helix/forkhead transcription factor FoxH [Branchiostoma
floridae]
Length=501

 Score =   100 bits (248),  Expect = 4e-19, Method: Compositional matrix adjust.
 Identities = 60/161 (37%), Positives = 84/161 (52%), Gaps = 45/161 (28%)
 Frame = +1

Query  34   RRVSEL-GRRTVKNYKRHGKPPYSYVGMIALLIRMAPEHKLSLAGILAAFKDLFPFFQGE  210
            RR  E+ G+R    Y+R+ KPPYSY+ ++ + I+ APE KL L  I  A K ++PFF+G+
Sbjct  50   RRTDEVEGKR---KYQRYPKPPYSYLALVVMAIQNAPEKKLPLKEIHEALKKMYPFFRGD  106

Query  211  YQGWRDSVRHNLSHNDCFYMVEVHKVLRIYIKVAPRPRLAFMNIYLLQNLSDPFPLQVPQ  390
            Y GW+DSVRHNLS   CFY V                          ++ S PF
Sbjct  107  YTGWKDSVRHNLSTYKCFYKVP-------------------------KDPSRPF------  135

Query  391  VINNVVKTKKCDWAVHLNKLPSDALRKQEKSSK---DTDYQ  504
                    K   WAV+ +K+P DAL+KQ+ SS+   + DY+
Sbjct  136  -------AKGNYWAVYEDKVPKDALKKQDSSSEREGEADYE  169


Query= 8012618 8166 173447 7732053-,124N,7913320-

Length=8166


                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value


ALIGNMENTS
Query= 7618526 2682 63681 2688144-,...,5039994-

Length=2682


                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value


ALIGNMENTS
  Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF
excluding environmental samples from WGS projects
    Posted date:  Feb 5, 2014  1:43 PM
  Number of letters in database: 12,906,967,781
  Number of sequences in database:  36,556,296

Lambda      K        H
   0.317    0.132    0.397
Gapped
Lambda      K        H
   0.267   0.0410    0.140
Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Sequences: 36556296
Number of Hits to DB: 12464349822
Number of extensions: 247037333
Number of successful extensions: 482575
Number of sequences better than 10: 78
Number of HSP's better than 10 without gapping: 0
Number of HSP's gapped: 480621
Number of HSP's successfully gapped: 156
Length of database: 12906967781
T: 12
A: 40
X1: 16 (7.3 bits)
X2: 38 (14.6 bits)
```

```
X3: 64 (24.7 bits)
S1: 41 (20.4 bits)
ka-blk-alpha gapped: 1.9
ka-blk-alpha ungapped: 0.7916
ka-blk-alpha_v gapped: 42.6028
ka-blk-alpha_v ungapped: 4.96466
ka-blk-sigma gapped: 43.6362

BLASTX 2.2.29+
Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro
A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and
David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new
generation of protein database search programs", Nucleic
Acids Res. 25:3389-3402.


RID: F8YAX2SF014


Database: All non-redundant GenBank CDS
translations+PDB+SwissProt+PIR+PRF excluding environmental samples
from WGS projects
           36,556,296 sequences; 12,906,967,781 total letters
Query= 7527519 3379 77835 251177+,...,3366549-

Length=1767


                                                                    Score     E
Sequences producing significant alignments:                        (Bits)  Value


ALIGNMENTS
Query= 8012618 8166 173447 7732053-,124N,7913320-

Length=8166


                                                                    Score     E
Sequences producing significant alignments:                        (Bits)  Value

emb|CAD45552.1|  fork head protein [Patella vulgata]                 665   0.0
gb|ESO85449.1|  hypothetical protein LOTGIDRAFT_183845 [Lottia...    565   2e-178
gb|EKC29500.1|  Hepatocyte nuclear factor 3-beta [Crassostrea ...    321   3e-93
gb|ADC35038.1|  forkhead box A [Themiste lageniformis]               322   2e-92
gb|ACE79136.1|  winged helix/forkhead transcription factor Fox...    285   2e-80
emb|CAA65368.1|  AmHNF-3-1 protein [Branchiostoma floridae]          280   4e-79
ref|NP_001158426.1|  forkhead box A [Saccoglossus kowalevskii]...    277   5e-78
gb|AFK11361.1|  hepatocyte nuclear factor 3-beta [Callorhinchu...    276   3e-77
ref|NP_001098162.1|  hepatocyte nuclear factor 3-beta [Oryzias...    275   4e-77
gb|ENN79246.1|  hypothetical protein YQE_04282, partial [Dendr...    275   1e-76

ALIGNMENTS
>emb|CAD45552.1| fork head protein [Patella vulgata]
Length=435

 Score =   665 bits (1715),  Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 315/317 (99%), Positives = 315/317 (99%), Gaps = 0/317 (0%)
 Frame = -2

Query  1127  RQMDPNMMSMDRAQALNRARDKNYRRSYTHAKPPYSYISLITMAIQQSPNKMCTLSEIYQ  948
             RQMDPNMMSMDRAQALNRARDKNYRRSYTHAKPPYSYISLITMAIQQSPNKMCTLSEIYQ
Sbjct  119   RQMDPNMMSMDRAQALNRARDKNYRRSYTHAKPPYSYISLITMAIQQSPNKMCTLSEIYQ  178

Query  947   FIMDLFPFYRQNQQRWQNSIRHSLSFNDCFVKVPRTPDRPGKGSYWALHPDSGNMFENGC  768
             FIMDLFPFYRQNQQRWQNSIRHSLSFNDCFVKVPRTPDRPGKGSYWALHPDSGNMFENGC
Sbjct  179   FIMDLFPFYRQNQQRWQNSIRHSLSFNDCFVKVPRTPDRPGKGSYWALHPDSGNMFENGC  238

Query  767   YLRRQKRFKCLKKESMRSSHDDDPSCGMSNGQNSADSTPTSTGDGTPLSAPNTPPHPVEQ  588
             YLRRQKRFKCLKKESMRSSHDDDPSCGMSNGQNSADSTPTSTGDGTPLSAPNTPPHPVEQ
Sbjct  239   YLRRQKRFKCLKKESMRSSHDDDPSCGMSNGQNSADSTPTSTGDGTPLSAPNTPPHPVEQ  298

Query  587   SQMAQPKTEQPTHQNHSQAHVQQQDMSHLTHSSQNQMGNCGTELTSMRQLHHDMSMNHGL  408
             SQM QPKTEQPTHQNHSQAHVQQQDMSHLTHSSQNQMGNCGTELTSMRQLHHDMSMNHGL
Sbjct  299   SQMTQPKTEQPTHQNHSQAHVQQQDMSHLTHSSQNQMGNCGTELTSMRQLHHDMSMNHGL  358

Query  407   NLAPGQLNHPHSFNHPFSITNLMSENKMDLKMYEAISGYGAYTQMSPMSMPKEASPPMNA  228
             NLAPGQLNHPHSFNHPFSITNLMSENKMD KMYEAISGYGAYTQMSPMSMPKEASPPMNA
Sbjct  359   NLAPGQLNHPHSFNHPFSITNLMSENKMDFKMYEAISGYGAYTQMSPMSMPKEASPPMNA  418
```

```
Query  227   QDGSYYKTYAPHSTASL  177
             QDGSYYKTYAPHSTASL
Sbjct  419   QDGSYYKTYAPHSTASL  435


 Score = 82.4 bits (202),  Expect = 1e-12, Method: Compositional matrix adjust.
 Identities = 40/40 (100%), Positives = 40/40 (100%), Gaps = 0/40 (0%)
 Frame = -2

Query  1481  MLSAKPGSYDPTSSGGYSMASMTSINTMGGVGPMSNMNYP  1362
             MLSAKPGSYDPTSSGGYSMASMTSINTMGGVGPMSNMNYP
Sbjct  1     MLSAKPGSYDPTSSGGYSMASMTSINTMGGVGPMSNMNYP  40


>gb|ESO85449.1| hypothetical protein LOTGIDRAFT_183845 [Lottia gigantea]
Length=437

 Score =   565 bits (1457),  Expect = 2e-178, Method: Compositional matrix adjust.
 Identities = 283/322 (88%), Positives = 295/322 (92%), Gaps = 7/322 (2%)
 Frame = -2

Query  1127  RQMDPNMMSMDRAQALNRARDKNYRRSYTHAKPPYSYISLITMAIQQSPNKMCTLSEIYQ  948
             R MDPNMMSMDRAQALNRARDKNYRRSYTHAKPPYSYISLITMAIQQSPNKMCTLSEIYQ
Sbjct  118   RPMDPNMMSMDRAQALNRARDKNYRRSYTHAKPPYSYISLITMAIQQSPNKMCTLSEIYQ  177

Query  947   FIMDLFPFYRQNQQRWQNSIRHSLSFNDCFVKVPRTPDRPGKGSYWALHPDSGNMFENGC  768
             FIMDLFPFYRQNQQRWQNSIRHSLSFNDCFVKVPRTPDRPGKGSYWALHPDSGNMFENGC
Sbjct  178   FIMDLFPFYRQNQQRWQNSIRHSLSFNDCFVKVPRTPDRPGKGSYWALHPDSGNMFENGC  237

Query  767   YLRRQKRFKCLKKESMRSSHDDDPSCGMSNGQNSADSTPTSTGDGTPLSAPNTPPH--PV  594
             YLRRQKRFKCLKKESMRS   DDP+C M +G  SADSTPTSTGDGT LSAPN+PP+
Sbjct  238   YLRRQKRFKCLKKESMRSQDPDDPNCDMQDGPGSADSTPTSTGDGTSLSAPNSPPNHQVE  297

Query  593   EQSQMAQPKTEQPTHQNHSQAHVQQQDMSHLTHS--SQNQMGNCGTELTSMRQLHHDMSM  420
             +  QM QPKTE   +Q H QAH+QQQDMSHLTHS  SQ QMGNCGTELTSMRQLHHD+SM
Sbjct  298   QVQQMVQPKTEH--NQAHPQAHLQQQDMSHLTHSQASQQQMGNCGTELTSMRQLHHDISM  355

Query  419   N-HGLNLAPGQLNHPHSFNHPFSITNLMSENKMDLKMYEAISGYGAYTQMSPMSMPKEAS  243
             N HGLNLAPGQLNHPHSFNHPFSITNLMSENKMDLKMYEA+SGY +YTQMSP+S+PKEAS
Sbjct  356   NHHGLNLAPGQLNHPHSFNHPFSITNLMSENKMDLKMYEALSGYQSYTQMSPISLPKEAS  415

Query  242   PPMNAQDGSYYKTYAPHSTASL  177
             PPMN QDGSYYKTYAPHSTASL
Sbjct  416   PPMNPQDGSYYKTYAPHSTASL  437


 Score = 73.2 bits (178),  Expect = 1e-09, Method: Compositional matrix adjust.
 Identities = 35/39 (90%), Positives = 36/39 (92%), Gaps = 0/39 (0%)
 Frame = -2

Query  1481  MLSAKPGSYDPTSSGGYSMASMTSINTMGGVGPMSNMNY  1365
             MLS KPGSYDPTS+GGYSMASM  INTMGGVGPMSNMNY
Sbjct  1     MLSTKPGSYDPTSTGGYSMASMPGINTMGGVGPMSNMNY  39


>gb|EKC29500.1| Hepatocyte nuclear factor 3-beta [Crassostrea gigas]
Length=401

 Score =   321 bits (823),  Expect = 3e-93, Method: Compositional matrix adjust.
 Identities = 189/317 (60%), Positives = 214/317 (68%), Gaps = 33/317 (10%)
 Frame = -2

Query  1109  MMSMDRAQALNRARDKNYRRSYTHAKPPYSYISLITMAIQQSPNKMCTLSEIYQFIMDLF  930
             M S +RAQA+NRARDK+YRRSYTHAKPPYSYISLITMAIQQSPNKMCTLSEIYQFIMDLF
Sbjct  112   MESYNRAQAINRARDKSYRRSYTHAKPPYSYISLITMAIQQSPNKMCTLSEIYQFIMDLF  171

Query  929   PFYRQNQQRWQNSIRHSLSFNDCFVKVPRTPDRPGKGSYWALHPDSGNMFENGCYLRRQK  750
             PFYRQNQQRWQNSIRHSLSFNDCFVKVPRTPDRPGKGSYW LHPDSGNMFENGCYLRRQK
Sbjct  172   PFYRQNQQRWQNSIRHSLSFNDCFVKVPRTPDRPGKGSYWTLHPDSGNMFENGCYLRRQK  231

Query  749   RFKCLKKESMRSSHDDDPSCGMSNGQN-SADSTPTSTGDGTPLSAPNTPPHPVEQSQMAQ  573
             RFKCLKKE +R S        +S  ++ S    P S      LS P   PH  E    AQ
Sbjct  232   RFKCLKKEMIRQS--------LSKSEDGSVIENPHSPRSDQSLSPP---PHSPEDLPNAQ  280

Query  572   PKTEQPTHQNHSQAHVQQQDMSHLTHSSQNQMGNCGTELTSMRQLHHDMSMNHGLNLAPG  393
                EQP          ++ ++   TH++            SMR       ++   N   G
Sbjct  281   ---EQPD---------RKPEIPITTHNTTTPPSQHQEMPPSMRHHPQHDPLSQAYNN--G  326
```

```
Query  392   QLNHPHSFNHPFSITNLMSEN-KMDLK-MYEAISGYGAYTQ---MSPMSMPKEASPPMNA  228
             Q N PHSF+HPFSIT+L+++N KMD+K MYE  + Y  Y     + PM+   E+S PM
Sbjct  327   QYN-PHSFHHPFSITSLITDNAKMDMKSMYEMPNMYAGYNNSMALPPMAK-TESSTPMPL  384

Query  227   QDGSYYKTYAPHSTASL  177
              D  YYKTYAPHSTASL
Sbjct  385   SDNGYYKTYAPHSTASL  401


>gb|ADC35038.1| forkhead box A [Themiste lageniformis]
Length=471

 Score =   322 bits (824),  Expect = 2e-92, Method: Compositional matrix adjust.
 Identities = 189/353 (54%), Positives = 227/353 (64%), Gaps = 46/353 (13%)
 Frame = -2

Query  1106  MSMDRAQALNRARDKNYRRSYTHAKPPYSYISLITMAIQQSPNKMCTLSEIYQFIMDLFP  927
             M  ++ QALNRAR+K YRRSYTHAKPPYSYISLITMAIQQSP+KMCTLSEIYQFIMDLFP
Sbjct  122   MDFNQQQALNRAREKTYRRSYTHAKPPYSYISLITMAIQQSPSKMCTLSEIYQFIMDLFP  181

Query  926   FYRQNQQRWQNSIRHSLSFNDCFVKVPRTPDRPGKGSYWALHPDSGNMFENGCYLRRQKR  747
             FYRQNQQRWQNSIRHSLSFNDCFVKVPRTPDRPGKGSYW LHPD+GNMFENGCYLRRQKR
Sbjct  182   FYRQNQQRWQNSIRHSLSFNDCFVKVPRTPDRPGKGSYWTLHPDAGNMFENGCYLRRQKR  241

Query  746   FKCLKKESMRSSHD-----DDPSCGMSNGQNSAD------STPTSTGDGTPLSAPNTPPH  600
             FKCLKKE +R   D      +P    S+G+   +      +TP    D T +++   P H
Sbjct  242   FKCLKKEELRQGLDMEDMNGEPMSPGSDGECRPNENLQPATTPPGVQDRTAVASQGPPQH  301

Query  599   PVEQSQMAQPKTEQPTHQ-------NHSQAHVQQQDMSHLTHSSQNQMGNCGTELTSMRQ  441
               +  +  +PKTE  T Q        H+     QQ   H  H  Q Q  +    +
Sbjct  302   --QTLEPIEPKTEPVTQQQPPSVPVTHAHQLEPQQHPHHPNHHQQPQQQHNHHPVEDQGP  359

Query  440   LHHDMSMNH---------GLNLAPGQLN------HPHSFNHPFSITNLMS-----ENKMD  321
             L + M ++H         G+N  P  +N      HP  F+HPFS TNLMS     ++K+D
Sbjct  360   LRNPMELSHVSSQASVMSGMNPPPPLMNAMSMNSHP-GFHHPFSFTNLMSSQQVHDSKVD  418

Query  320   LKMYEAISGYGAYTQ---MSPMSMPKEASPPMN--AQDGSYYKTYAPHSTASL  177
             +KMYE +   G Y+    MSPM++PKE+S  ++    DG YYKTY PHSTA+L
Sbjct  419   VKMYENMQAAGMYSHPYNMSPMNVPKESSSAISTPTPDGGYYKTYTPHSTAAL  471


>gb|ACE79136.1| winged helix/forkhead transcription factor FoxAa [Branchiostoma
floridae]
Length=407

 Score =   285 bits (728),  Expect = 2e-80, Method: Compositional matrix adjust.
 Identities = 170/317 (54%), Positives = 199/317 (63%), Gaps = 27/317 (9%)
 Frame = -2

Query  1085  ALNR----ARDKNYRRSYTHAKPPYSYISLITMAIQQSPNKMCTLSEIYQFIMDLFPFYR  918
             ALNR     R+K YRRSYTHAKPPYSYISLITM+IQ SPNKM TL+EIYQFIMDLFP+YR
Sbjct  90    ALNRNAIAEREKAYRRSYTHAKPPYSYISLITMSIQSSPNKMVTLAEIYQFIMDLFPYYR  149

Query  917   QNQQRWQNSIRHSLSFNDCFVKVPRTPDRPGKGSYWALHPDSGNMFENGCYLRRQKRFKC  738
             QNQQRWQNSIRHSLSFNDCFVKVPRTPDRPGKGSYW LHP++GNMFENGCYLRRQKRFKC
Sbjct  150   QNQQRWQNSIRHSLSFNDCFVKVPRTPDRPGKGSYWTLHPEAGNMFENGCYLRRQKRFKC  209

Query  737   LKKESMRSSHD----DDPSCGMSNGQNSAD------STPTSTGDGT----PLSAPNTP-P  603
              KK +M+ +      D P+ G  N   S        STPT+T +G     PL   NTP P
Sbjct  210   EKKLAMKMAQQQAARDTPNPGTENSAVSPTTTAEPASTPTTTSNGASTLQPLQPINTPSP  269

Query  602   HPVEQSQMAQPKTEQPTHQNHSQAHVQQQDMSHLTHSSQNQMGNCGTELTSMRQLHHDMS  423
             +P EQ Q    + +              Q       H+ Q+Q    G     + Q    +
Sbjct  270   NPQEQQQHQHHQHQQQHQQQPQVQTTQPQDMQQHAQQHQ----GLPARPIPQ-QSSLP  324

Query  422   MNHGLNLAPGQLNHPHSFNHPFSITNLMS-ENKMDLKMYEAISGYGAYTQMSPMSMPKEA  246
             M+ G   +P  L    H F HPFSI+NLMS E+K DLK Y A+ GY  Y  MSP  +PK
Sbjct  325   MSMGGYFSPEHLRAAHGFTHPFSISNLMSQEHKPDLKEYAAM-GYSGYNSMSPTGVPK-T  382

Query  245   SPPMNAQDGSYYKTYAP  195
             +  M++    YY+ Y P
Sbjct  383   TMSMDSMGTDYYQGYVP  399


>emb|CAA65368.1| AmHNF-3-1 protein [Branchiostoma floridae]
Length=403

 Score =   280 bits (717),  Expect = 4e-79, Method: Compositional matrix adjust.
 Identities = 166/313 (53%), Positives = 195/313 (62%), Gaps = 23/313 (7%)
```

```
 Frame = -2

Query  1085  ALNR----ARDKNYRRSYTHAKPPYSYISLITMAIQQSPNKMCTLSEIYQFIMDLFPFYR  918
             ALNR     R+K YRRSYTHAKPPYSYISLITM+IQ SPNKM TL+EIYQFIMDLFP+YR
Sbjct  90    ALNRNAIAEREKAYRRSYTHAKPPYSYISLITMSIQSSPNKMVTLAEIYQFIMDLFPYYR  149

Query  917   QNQQRWQNSIRHSLSFNDCFVKVPRTPDRPGKGSYWALHPDSGNMFENGCYLRRQKRFKC  738
             QNQQRWQNSIRHSLSFNDCFVKVPRTPDRPGKGSYW LHP++GNMFENGCYLRRQKRFKC
Sbjct  150   QNQQRWQNSIRHSLSFNDCFVKVPRTPDRPGKGSYWTLHPEAGNMFENGCYLRRQKRFKC  209

Query  737   LKKESMRSSHDD------DPSCGMSNGQNSADSTPTSTGDGT----PLSAPNTP-PHPVE  591
              KK +M+ +          P   +    +   STPT+T +G     PL   NTP P+P E
Sbjct  210   EKKLAMKMAQQQAAPTPPTPGRELRRLADYHGSTPTTTSNGASTLQPLQPINTPSPNPQE  269

Query  590   QSQMAQPKTEQPTHQNHSQAHVQQQDMSHLTHSSQNQMGNCGTELTSMRQLHHDMSMNHG  411
             Q Q    + +              Q      H+ Q+Q    G     + Q    + M+ G
Sbjct  270   QQQHHQHQHQQQHQQQPQVQTTQPQDMQQHAQQHQ----GLPARPIPQ-QSSLPMSMG  324

Query  410   LNLAPGQLNHPHSFNHPFSITNLMS-ENKMDLKMYEAISGYGAYTQMSPMSMPKEASPPM  234
                +P  L   H F HPFSI+NLMS E+K DLK Y A+ GY  Y  MSP  +PK     M
Sbjct  325   GYFSPEHLRAAHGFTHPFSISNLMSQEHKPDLKEYAAM-GYSGYNSMSPTGVPKTTM-SM  382

Query  233   NAQDGSYYKTYAP  195
             ++    YY+ Y P
Sbjct  383   DSMGTDYYQGYVP  395


>ref|NP_001158426.1| forkhead box A [Saccoglossus kowalevskii]
 gb|ACG76356.1| forkhead box A protein [Saccoglossus kowalevskii]
Length=404

 Score =   277 bits (709),  Expect = 5e-78, Method: Compositional matrix adjust.
 Identities = 162/291 (56%), Positives = 189/291 (65%), Gaps = 29/291 (10%)
 Frame = -2

Query  1097  DRAQALNRAR--DKNYRRSYTHAKPPYSYISLITMAIQQSPNKMCTLSEIYQFIMDLFPF  924
             ++  A+NR R  +K YRRSYTHAKPPYSYISLITMAIQ SPNKM TLS+IYQFIMDLFPF
Sbjct  104   NQMNAVNRVRTDNKTYRRSYTHAKPPYSYISLITMAIQSSPNKMVTLSDIYQFIMDLFPF  163

Query  923   YRQNQQRWQNSIRHSLSFNDCFVKVPRTPDRPGKGSYWALHPDSGNMFENGCYLRRQKRF  744
             YRQNQQRWQNSIRHSLSFNDCF+KVPRTPDRPGKGS+W LHPDSGNMFENGCYLRRQKRF
Sbjct  164   YRQNQQRWQNSIRHSLSFNDCFLKVPRTPDRPGKGSFWTLHPDSGNMFENGCYLRRQKRF  223

Query  743   KCLKKESMRSSHDDDPSCGMSNGQNSADSTPTSTGDGTP-LSAPNTPPHPVEQSQMAQPK  567
             KC K+E+ R +  D        G+       T+   G+P ++    T      Q QM QPK
Sbjct  224   KCPKREAQRVTQRD--------GRTDEHGEETTGSTGSPHMTDTTTAQQQQTQQQMPQPK  275

Query  566   TEQPTHQNHSQAHVQQQDMSHLTHSSQNQMGNCGTELTSMRQLHHDMSMNHGLN---LAP  396
              E      H+ +      +  +T  S     N  T L   R +    +   G+    L P
Sbjct  276   VEP-----HTMSPPPPNSIQSITSMS----NNVPTSLAMARPIPQQAAYVSGMGAHMLGP  326

Query  395   GQLNHPHSFNHPFSITNLMSENKMDLKMYEAISGYGAYTQMSPM-SMPKEA  246
             GQ  HP SFNHPFSITN+MS  +   K YEA+ GY  Y  MSP+ SMPK +
Sbjct  327   GQ--HP-SFNHPFSITNIMSPAEHG-KAYEAM-GYPGYNTMSPISSMPKSS  372


>gb|AFK11361.1| hepatocyte nuclear factor 3-beta [Callorhinchus milii]
Length=430

 Score =   276 bits (706),  Expect = 3e-77, Method: Compositional matrix adjust.
 Identities = 160/303 (53%), Positives = 194/303 (64%), Gaps = 31/303 (10%)
 Frame = -2

Query  1082  LNRARD-KNYRRSYTHAKPPYSYISLITMAIQQSPNKMCTLSEIYQFIMDLFPFYRQNQQ  906
             LNR+RD K YRRSYTHAKPPYSYISLITMAIQQSP+KM TLSEIYQ+IMDLFPFYRQNQQ
Sbjct  137   LNRSRDPKTYRRSYTHAKPPYSYISLITMAIQQSPSKMLTLSEIYQWIMDLFPFYRQNQQ  196

Query  905   RWQNSIRHSLSFNDCFVKVPRTPDRPGKGSYWALHPDSGNMFENGCYLRRQKRFKCLKKE  726
             RWQNSIRHSLSFNDCF+KVPR+PD+PGKGS+W LHPDSGNMFENGCYLRRQKRFKC KK+
Sbjct  197   RWQNSIRHSLSFNDCFLKVPRSPDKPGKGSFWTLHPDSGNMFENGCYLRRQKRFKCEKKQ  256

Query  725   SMRSSHDDDPSCGMSNGQNSADSTPTSTGDGTPLSAPNTPPHPVEQS--QMAQPKTEQPT  552
             +++S+ +       +    S  S+ ++ G+ +P S+  +P    ++S   M       P
Sbjct  257   ALKSAQETTRKTSETGSTGSNSSSESTNGNESPHSSA-SPGREQKRSLVDMKSASALSPE  315

Query  551   HQNHSQAHVQQQDMSHLTHSSQNQMGNCGTELTSMRQLHHDMSMNHGLNLAPGQLNHPHS  372
             H + S +  Q Q   HL H   + + +  +E     Q H          L P   +H +S
Sbjct  316   HSHTSSSVAQAQQ--HLMHQHHSVLSHIASEA----QAH----------LKP---DHHYS  356
```

```
Query  371   FNHPFSITNLMSE----NKMDLKMYEAISGYGAY----TQMSPMSMPKEASPPMNAQDGS  216
             FNHPFSI NLMS     +KMDLK YE +  Y  Y    T   PMS          A D S
Sbjct  357   FNHPFSINNLMSSEQQHHKMDLKAYEQVMHYSNYGSPMTANLPMSSKTVLDTSAIASDAS  416

Query  215   YYK  207
             YY+
Sbjct  417   YYQ  419


>ref|NP_001098162.1| hepatocyte nuclear factor 3-beta [Oryzias latipes]
 sp|O42097.1|FOXA2_ORYLA RecName: Full=Hepatocyte nuclear factor 3-beta; Short=HNF-3-
beta;
Short=HNF-3B; AltName: Full=Forkhead box protein A2; AltName:
Full=Me-HNF3B
 dbj|BAA23579.1| Me-HNF3B [Oryzias latipes]
Length=415

 Score =   275 bits (703),  Expect = 4e-77, Method: Compositional matrix adjust.
 Identities = 151/274 (55%), Positives = 179/274 (65%), Gaps = 39/274 (14%)
 Frame = -2

Query  1082  LNRARD-KNYRRSYTHAKPPYSYISLITMAIQQSPNKMCTLSEIYQFIMDLFPFYRQNQQ  906
             +NR+RD K YRRSYTHAKPPYSYISLITMAIQQSP+KM TL+EIYQ+IMDLFPFYRQNQQ
Sbjct  133   INRSRDPKTYRRSYTHAKPPYSYISLITMAIQQSPSKMLTLAEIYQWIMDLFPFYRQNQQ  192

Query  905   RWQNSIRHSLSFNDCFVKVPRTPDRPGKGSYWALHPDSGNMFENGCYLRRQKRFKCLKKE  726
             RWQNSIRHSLSFNDCF+KVPR+PD+PGKGS+W LHPDSGNMFENGCYLRRQKRFKC KK
Sbjct  193   RWQNSIRHSLSFNDCFLKVPRSPDKPGKGSFWTLHPDSGNMFENGCYLRRQKRFKCEKKM  252

Query  725   SMRSSHDDDPSCGMSNGQNSADSTPTSTGDGTPLSAPNTPPHPVEQSQMAQPKTEQPTHQ  546
             SM+     +P     +G ++  S+ +  G+ +P S  ++  H    S M   +   P H
Sbjct  253   SMK-----EPGRKGGDGGSANSSSDSCNGNESPHSNSSSGEHKRSLSDMKGSQALSPEHT  307

Query  545   NHSQAHVQQQDMSHLTHSSQNQMGNCGTELTSMRQLHHDMSMNHGLNLAPGQLNHPHSFN  366
               S     Q  MS                        HH + + H  +L P    H +SFN
Sbjct  308   APSPVSQGQHLMSQ----------------------HHSV-LAHEAHLKP---EHHYSFN  341

Query  365   HPFSITNLMSE----NKMDLKMYEAI---SGYGA  285
             HPFSI NLMS     +KMDLK YE +   SGYG+
Sbjct  342   HPFSINNLMSSEQQHHKMDLKTYEQVMHYSGYGS  375


>gb|ENN79246.1| hypothetical protein YQE_04282, partial [Dendroctonus ponderosae]
 gb|ERL93739.1| hypothetical protein D910_11025 [Dendroctonus ponderosae]
Length=436

 Score =   275 bits (702),  Expect = 1e-76, Method: Compositional matrix adjust.
 Identities = 161/289 (56%), Positives = 182/289 (63%), Gaps = 25/289 (9%)
 Frame = -2

Query  1085  ALNRA--RDKNYRRSYTHAKPPYSYISLITMAIQQSPNKMCTLSEIYQFIMDLFPFYRQN  912
             AL RA   +K YRRSYTHAKPPYSYISLITMAIQ SP KM TLSEIYQFIMDLFPFYRQN
Sbjct  128   ALQRAVRTEKPYRRSYTHAKPPYSYISLITMAIQNSPQKMLTLSEIYQFIMDLFPFYRQN  187

Query  911   QQRWQNSIRHSLSFNDCFVKVPRTPDRPGKGSYWALHPDSGNMFENGCYLRRQKRFKCLK  732
             QQRWQNSIRHSLSFNDCFVKVPRTPD+PGKGS+W+LHPDSGNMFENGCYLRRQKRFK  K
Sbjct  188   QQRWQNSIRHSLSFNDCFVKVPRTPDKPGKGSFWSLHPDSGNMFENGCYLRRQKRFKDEK  247

Query  731   KESMRSSHDDDPSCGMSNGQNSADSTPTSTGDGTP--LSAP---------NTPPHPVEQS  585
             KE +R  H        M N  +S   TP  +G+     L  P         N  P  ++
Sbjct  248   KEVIRQQHKSPSHHSMDNSNSSGKKTPLHSGEDKSHQLEKPSENNLSSMMNIHPSKLDVD  307

Query  584   QMAQPKT-EQPTHQNHSQAHVQQQDMSHLTHSSQNQMGNCGTELTSMRQLHHDMSMNHGL  408
             QM+     E   HQ H       Q+MSH   S      N    LT  + + H  S+NH L
Sbjct  308   QMSMMNANELNMHQQH-----HHQNMSHEELSVMINRSNHLASLTEHQAMLHHNSINHHL  362

Query  407   NLAPGQLNHPHSFNHPFSITNLM-SENKMDLKMYEAISGYGAYTQMSPM  264
                P         NHPFSIT L+ +E+K D+KMY A  GYG Y  +SP+
Sbjct  363   KQEPTGYT---PSNHPFSITRLLPTESKSDIKMY-ADMGYG-YNTLSPL  406


Query= 7618526 2682 63681 2688144-,...,5039994-

Length=2682


                                                                  Score     E
Sequences producing significant alignments:                      (Bits)  Value
```

ALIGNMENTS
  Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
    Posted date: Feb 5, 2014 1:43 PM
  Number of letters in database: 12,906,967,781
  Number of sequences in database: 36,556,296

Lambda   K   H
   0.317   0.132   0.397
Gapped
Lambda   K   H
   0.267   0.0410   0.140
Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Sequences: 36556296
Number of Hits to DB: 12464349822
Number of extensions: 247037333
Number of successful extensions: 482575
Number of sequences better than 10: 78
Number of HSP's better than 10 without gapping: 0
Number of HSP's gapped: 480621
Number of HSP's successfully gapped: 156
Length of database: 12906967781
T: 12
A: 40
X1: 16 (7.3 bits)
X2: 38 (14.6 bits)
X3: 64 (24.7 bits)
S1: 41 (20.4 bits)
ka-blk-alpha gapped: 1.9
ka-blk-alpha ungapped: 0.7916
ka-blk-alpha_v gapped: 42.6028
ka-blk-alpha_v ungapped: 4.96466
ka-blk-sigma gapped: 43.6362

BLASTX 2.2.29+
Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro
A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and
David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new
generation of protein database search programs", Nucleic
Acids Res. 25:3389-3402.

RID: F8YAX2SF014

Database: All non-redundant GenBank CDS
translations+PDB+SwissProt+PIR+PRF excluding environmental samples
from WGS projects
           36,556,296 sequences; 12,906,967,781 total letters
Query= 7527519 3379 77835 251177+,...,3366549-

Length=1767

                                                                      Score   E
Sequences producing significant alignments: (Bits) Value

ALIGNMENTS
Query= 8012618 8166 173447 7732053-,124N,7913320-

Length=8166

                                                                      Score   E
Sequences producing significant alignments: (Bits) Value

ALIGNMENTS
Query= 7618526 2682 63681 2688144-,...,5039994-

Length=2682

                                                                      Score   E
Sequences producing significant alignments: (Bits) Value

```
gb|EKC40931.1|  Fork head domain transcription factor slp2 [Cr...    227   2e-62
gb|ESP02936.1|  hypothetical protein LOTGIDRAFT_59807, partial...    215   3e-62
gb|AAP79301.1|  brain factor 1 [Saccoglossus kowalevskii]            208   2e-56
ref|XP_002735197.1|  PREDICTED: brain factor 1 [Saccoglossus k...    208   3e-56
gb|AEZ03828.1|  FoxG, partial [Terebratalia transversa]              208   6e-56
ref|XP_002610230.1|  hypothetical protein BRAFLDRAFT_286825 [B...    206   3e-55
gb|AAC18392.1|  transcription factor BF-1 [Branchiostoma flori...    206   3e-55
gb|ELT94106.1|  hypothetical protein CAPTEDRAFT_139421, partia...    196   9e-55
ref|XP_003706042.1|  PREDICTED: uncharacterized protein LOC100...    202   8e-54
gb|ADG26725.1|  forkhead box protein G1 [Platynereis dumerilii]      199   1e-53

ALIGNMENTS
>gb|EKC40931.1| Fork head domain transcription factor slp2 [Crassostrea gigas]
Length=403

 Score =   227 bits (578),  Expect = 2e-62, Method: Compositional matrix adjust.
 Identities = 171/380 (45%), Positives = 225/380 (59%), Gaps = 41/380 (11%)
 Frame = +1

Query  1012  FSISRVLGEDIG-----CSPAQADDNCEKAELDGSSSFREDFVDVVSEEADDGPSFDGNI  1176
             FSISRVLGED+       +  + DD+ E  ++    S   D V+ +  E  +   ++
Sbjct  16    FSISRVLGEDLNDESDDMAMNEHDDSSENLDIQADDSCNVD-VESLESEMQNSEYYEKVF  74

Query  1177  SVNDD---SGDDMRDDTLEDLEDKQSENDGDLKDTGSKEEKKAEKPPFSYNALIMMAIRG  1347
               N D   S  + +D+ + + +   S+ + +     S + KK+EKPPFSYNALIMMAIR
Sbjct  75    PKNCDLPPSSPETKDNDIGNDDGDDSDRNNNKDSQQSNDSKKSEKPPFSYNALIMMAIRS  134

Query  1348  SPEKRLTLSQIYEFIMKNFPYYKDNKQGWQNSIRHNLSLNKCFLKVPRHYDDPGKGNYWM  1527
             S EKRLTL+ IYEFIMKNFPYYKDNKQGWQNSIRHNLSLNKCF+KVPRHYDDPGKGNYWM
Sbjct  135   SAEKRLTLNGIYEFIMKNFPYYKDNKQGWQNSIRHNLSLNKCFVKVPRHYDDPGKGNYWM  194

Query  1528  LDPCCEDVFIGGTTGklrrrstlatrsrlaalkraGFPHISVPGYPYPADRLHPYAACTP  1707
             LDP C+DVFIGGTTGKLRRRST A+RSRLAALKRAGF     P YPY     H   + +
Sbjct  195   LDPSCDDVFIGGTTGKLRRRSTTASRSRLAALKRAGFAGYQSPVYPYFG---HTGKSGSY  251

Query  1708  LYALPGF-------PGTRYPSLPYSAYP-----GALSFAN--PSDRS-APKPLSFSVDRL  1842
             ++  P         P +   SL YS +P     G LS ++  P++ S +P+  +FSVDRL
Sbjct  252   IWPFPSLFHLGSASPASTSGSLRYSGFPMYPYRGLLSPSSSLPANSSPSPRTTNFSVDRL  311

Query  1843  LSDSVRSSENASESLPTAQALPRQHSNSAIFRNPLILPTFSAID---PRNMEMYLNQLRS  2013
             L   + S +  ++S  +    P Q + S           FS +D   P N  + L+  +
Sbjct  312   LGLDINSQQQVNKSNSSFLQWPPQAAMS-----------FSQMDQSSPNNRGLDLSVFKG  360

Query  2014  FPLSPFAGMGALASPNRLLK  2073
             F   P + + A  SP+ + +
Sbjct  361   FHGLPVSPVSAFTSPDVMTQ  380


>gb|ESP02936.1| hypothetical protein LOTGIDRAFT_59807, partial [Lottia gigantea]
Length=105

 Score =   215 bits (547),  Expect = 3e-62, Method: Compositional matrix adjust.
 Identities = 98/98 (100%), Positives = 98/98 (100%), Gaps = 0/98 (0%)
 Frame = +1

Query  1279  KEEKKAEKPPFSYNALIMMAIRGSPEKRLTLSQIYEFIMKNFPYYKDNKQGWQNSIRHNL  1458
             KEEKKAEKPPFSYNALIMMAIRGSPEKRLTLSQIYEFIMKNFPYYKDNKQGWQNSIRHNL
Sbjct  1     KEEKKAEKPPFSYNALIMMAIRGSPEKRLTLSQIYEFIMKNFPYYKDNKQGWQNSIRHNL  60

Query  1459  SLNKCFLKVPRHYDDPGKGNYWMLDPCCEDVFIGGTTG  1572
             SLNKCFLKVPRHYDDPGKGNYWMLDPCCEDVFIGGTTG
Sbjct  61    SLNKCFLKVPRHYDDPGKGNYWMLDPCCEDVFIGGTTG  98


>gb|AAP79301.1| brain factor 1 [Saccoglossus kowalevskii]
Length=356

 Score =   208 bits (530),  Expect = 2e-56, Method: Compositional matrix adjust.
 Identities = 143/278 (51%), Positives = 180/278 (65%), Gaps = 35/278 (13%)
 Frame = +1

Query  1165  DGNISVNDDSGDDMRD--DTLE----------DLEDKQSENDGDLKDTGSK------EEK  1290
             +GN+  +D+  + + D  +T+E          + E +  ++DG+ K TGS        K
Sbjct  6     NGNLHDSDEMAEKVHDNGETMESNDENANHSGNRESESKKDDGN-KQTGSSNGSSPPRNK  64

Query  1291  KAEKPPFSYNALIMMAIRGSPEKRLTLSQIYEFIMKNFPYYKDNKQGWQNSIRHNLSLNK  1470
               EKPPFSYNALIMMAIR SPEKRLTL+ IYEFIMK+FPYY++NKQGWQNSIRHNLSLNK
Sbjct  65    YGEKPPFSYNALIMMAIRQSPEKRLTLNGIYEFIMKHFPYYRENKQGWQNSIRHNLSLNK  124
```

```
Query  1471  CFLKVPRHYDDPGKGNYWMLDPCCEDVFIGGTTGklrrrstlatrsrlaalkraGFPHIS  1650
             CF+KVPRHYDDPGKGNYWMLDP  +DVFIGGTTGKLRRRST A+R+RLA LKR   P +
Sbjct  125   CFVKVPRHYDDPGKGNYWMLDPSSDDVFIGGTTGKLRRRSTSASRNRLAQLKR--HPRLH  182

Query  1651  VPGYPYPADRLHPYAACTPL-YALPGFPGTRYPS--LPYSAYPGALSFANPSDRSAPKPL  1821
               GYP  +D + PY    P  + LP  P     S  L Y+A  G L  ++ +    P P
Sbjct  183   GGGYPLQSD-IKPYPMYWPASHMLPSLPQHAAASNALRYTATAGGLHTSHYNSLFTPSPT  241

Query  1822  --------SFSVDRLL-SDSVRSSENASESL-PTAQAL  1905
                     +FSVDRL+ +D+  +S +   +L P+AQ L
Sbjct  242   MSRPLGSHNFSVDRLIGTDASYASPHHQNTLSPSAQTL  279


>ref|XP_002735197.1| PREDICTED: brain factor 1 [Saccoglossus kowalevskii]
Length=380

 Score =   208 bits (529),  Expect = 3e-56, Method: Compositional matrix adjust.
 Identities = 143/278 (51%), Positives = 180/278 (65%), Gaps = 35/278 (13%)
 Frame = +1

Query  1165  DGNISVNDDSGDDMRD--DTLE----------DLEDKQSENDGDLKDTGSK------EEK  1290
             +GN+  +D+  + + D  +T+E          + E +  ++DG+ K TGS          K
Sbjct  6     NGNLHDSDEMAEKVHDNGETMESNDENANHSGNRESESKKDDGN-KQTGSSNGSSPPRNK  64

Query  1291  KAEKPPFSYNALIMMAIRGSPEKRLTLSQIYEFIMKNFPYYKDNKQGWQNSIRHNLSLNK  1470
               EKPPFSYNALIMMAIR SPEKRLTL+ IYEFIMK+FPYY++NKQGWQNSIRHNLSLNK
Sbjct  65    YGEKPPFSYNALIMMAIRQSPEKRLTLNGIYEFIMKHFPYYRENKQGWQNSIRHNLSLNK  124

Query  1471  CFLKVPRHYDDPGKGNYWMLDPCCEDVFIGGTTGklrrrstlatrsrlaalkraGFPHIS  1650
             CF+KVPRHYDDPGKGNYWMLDP  +DVFIGGTTGKLRRRST A+R+RLA LKR   P +
Sbjct  125   CFVKVPRHYDDPGKGNYWMLDPSSDDVFIGGTTGKLRRRSTSASRNRLAQLKR--HPRLH  182

Query  1651  VPGYPYPADRLHPYAACTPL-YALPGFPGTRYPS--LPYSAYPGALSFANPSDRSAPKPL  1821
               GYP  +D + PY    P  + LP  P     S  L Y+A  G L  ++ +    P P
Sbjct  183   GGGYPLQSD-IKPYPMYWPASHMLPSLPQHAAASNALRYTATAGGLHTSHYNSLFTPSPT  241

Query  1822  --------SFSVDRLL-SDSVRSSENASESL-PTAQAL  1905
                     +FSVDRL+ +D+  +S +   +L P+AQ L
Sbjct  242   MSRPLGSHNFSVDRLIGTDASYASPHHQNTLSPSAQTL  279


>gb|AEZ03828.1| FoxG, partial [Terebratalia transversa]
Length=418

 Score =   208 bits (530),  Expect = 6e-56, Method: Compositional matrix adjust.
 Identities = 146/295 (49%), Positives = 181/295 (61%), Gaps = 34/295 (12%)
 Frame = +1

Query  1012  FSISRV-LGEDIGCSPAQADDNCEKAELDGSSSFREDFVDVVSEEADDGPSFDGNISVND  1188
             FSI+ + LG+ +     Q      KA           D ++   EEA +
Sbjct  42    FSINSIMLGDTVMKDQGQTSTEILKARQADELRVHNDNLNSNLEEASE----------S  90

Query  1189  DSGDDMRDDTLEDLEDKQSENDGDLKDTGSKEEKKAEKPPFSYNALIMMAIRGSPEKRLT  1368
             D  ++  DD+ ++ E+ +  +DG  K+   + + K +KPPFSYNALIMMAIR SPEKRLT
Sbjct  91    DVENNFADDSTDNAENTEPLDDG--KNEKVETKDKPDKPPFSYNALIMMAIRSSPEKRLT  148

Query  1369  LSQIYEFIMKNFPYYKDNKQGWQNSIRHNLSLNKCFLKVPRHYDDPGKGNYWMLDPCCED  1548
             L+ IYEFIM NFPYY+DNKQGWQNSIRHNLSLNKCF+KVPRHYDDPGKGNYWMLDP  +D
Sbjct  149   LNGIYEFIMTNFPYYRDNKQGWQNSIRHNLSLNKCFVKVPRHYDDPGKGNYWMLDPSSDD  208

Query  1549  VFIGGTTGklrrrstlatrsrlaalkraGFPHISVPGYPYPADRLHPY----AACTPLYA  1716
             VFIGGTTGKLRRRST A+RSRLAA KRAG P +  PG+ +      P+     +  P+ +
Sbjct  209   VFIGGTTGKLRRRSTAASRSRLAAFKRAGIPRL--PGFGFETFGKTPFMWPGVSVNPMVS  266

Query  1717  LP---------GFPGTRYPSLPYSAYPGALSFANPSDRSAPK--PLSFSVDRLLS  1848
             L          G  G  Y   PY+A   + + A+   RS P   P SFSVDRLLS
Sbjct  267   LQQQAVAMQRFGATGGSY---PYNALFPSNTVASLPPRSPPSTVPSSFSVDRLLS  318


>ref|XP_002610230.1| hypothetical protein BRAFLDRAFT_286825 [Branchiostoma floridae]
 gb|EEN66240.1| hypothetical protein BRAFLDRAFT_286825 [Branchiostoma floridae]
Length=402

 Score =   206 bits (524),  Expect = 3e-55, Method: Compositional matrix adjust.
 Identities = 155/304 (51%), Positives = 184/304 (61%), Gaps = 28/304 (9%)
 Frame = +1

Query  1009  PFSISRVLGEDIGCSPAQADDNCEKAELDGSSSFREDFVDVVSEEADDGPSFDGNISVND  1188
```

```
            PFSI R+L + +             AEL   S         V+    +G     N +VN+
Sbjct  19   PFSIRRMLSQPL-----------HTAELPTVSLTAGHAPAPVTCRETNGDHSHHNKAVNE  67

Query  1189 DS----GDDMRDDTLEDLEDKQSENDGDLKDTGSKEE-KKAEKPPFSYNALIMMAIRGSP  1353
             G D  D  ++    K SE +   K+   KEE KK EKPPFSYNALIMMAIR SP
Sbjct  68   HELGKDGKDAPDSKVDTDVPKDSEQEDFKKEKDDKEEGKKHEKPPFSYNALIMMAIRQSP  127

Query  1354 EKRLTLSQIYEFIMKNFPYYKDNKQGWQNSIRHNLSLNKCFLKVPRHYDDPGKGNYWMLD  1533
            EKRLTL+ IYEFIMKNFPYY++NKQGWQNSIRHNLSLNKCF+KVPRHYDDPGKGNYWMLD
Sbjct  128  EKRLTLNGIYEFIMKNFPYYRENKQGWQNSIRHNLSLNKCFVKVPRHYDDPGKGNYWMLD  187

Query  1534 PCCEDVFIGGTTGklrrrstlatrsrlaalkraGFPH-ISVPGYPYPADRLHPYAACTP-  1707
            P  +DVFIGGTTGKLRRRST A RSRLA  +  G  +   V  +P  AD+ + Y    P
Sbjct  188  PSSDDVFIGGTTGKLRRRSTAAARSRLAFRRGFGVRYPAGVMEWPA-ADKTNCYWTTHPP  246

Query  1708 ---LYALPGF-PGTRYPSLPYSAYPGALSFANPSDRSAPKPLSFSVDRLLS-DSVRSSEN  1872
                Y+LP   PG  Y S P S+ PG   F   S  S+    +FSV+RLLS D+ R++
Sbjct  247  ANGGYSLPQHSPGFHY-SPPPSSTPG---FGFTSPHSSTPQHNFSVERLLSTDTSRAAPV  302

Query  1873 ASES  1884
             S S
Sbjct  303  CSLS  306


>gb|AAC18392.1| transcription factor BF-1 [Branchiostoma floridae]
Length=402

 Score =   206 bits (524),  Expect = 3e-55, Method: Compositional matrix adjust.
 Identities = 155/304 (51%), Positives = 184/304 (61%), Gaps = 28/304 (9%)
 Frame = +1

Query  1009 PFSISRVLGEDIGCSPAQADDNCEKAELDGSSSFREDFVDVVSEEADDGPSFDGNISVND  1188
            PFSI R+L + +             AEL   S         V+    +G     N +VN+
Sbjct  19   PFSIRRMLSQPL-----------HTAELPTVSLTAGHAPAPVTCRETNGDHSHHNKAVNE  67

Query  1189 DS----GDDMRDDTLEDLEDKQSENDGDLKDTGSKEE-KKAEKPPFSYNALIMMAIRGSP  1353
             G D  D  ++    K SE +   K+   KEE KK EKPPFSYNALIMMAIR SP
Sbjct  68   HELGKDGKDAPDSKVDTDVPKDSEQEDFKKEKDDKEEGKKHEKPPFSYNALIMMAIRQSP  127

Query  1354 EKRLTLSQIYEFIMKNFPYYKDNKQGWQNSIRHNLSLNKCFLKVPRHYDDPGKGNYWMLD  1533
            EKRLTL+ IYEFIMKNFPYY++NKQGWQNSIRHNLSLNKCF+KVPRHYDDPGKGNYWMLD
Sbjct  128  EKRLTLNGIYEFIMKNFPYYRENKQGWQNSIRHNLSLNKCFVKVPRHYDDPGKGNYWMLD  187

Query  1534 PCCEDVFIGGTTGklrrrstlatrsrlaalkraGFPH-ISVPGYPYPADRLHPYAACTP-  1707
            P  +DVFIGGTTGKLRRRST A RSRLA  +  G  +   V  +P  AD+ + Y    P
Sbjct  188  PSSDDVFIGGTTGKLRRRSTAAARSRLAFRRGFGVRYPAGVMDWPA-ADKTNCYWTTHPP  246

Query  1708 ---LYALPGF-PGTRYPSLPYSAYPGALSFANPSDRSAPKPLSFSVDRLLS-DSVRSSEN  1872
                Y+LP   PG  Y S P S+ PG   F   S  S+    +FSV+RLLS D+ R++
Sbjct  247  ANGGYSLPQHSPGFHY-SPPPSSTPG---FGFTSPHSSTPQHNFSVERLLSTDTSRAAPV  302

Query  1873 ASES  1884
             S S
Sbjct  303  CSLS  306


>gb|ELT94106.1| hypothetical protein CAPTEDRAFT_139421, partial [Capitella teleta]
Length=157

 Score =   196 bits (498),  Expect = 9e-55, Method: Compositional matrix adjust.
 Identities = 91/125 (73%), Positives = 105/125 (84%), Gaps = 2/125 (2%)
 Frame = +1

Query  1198 DDMRDDTLEDLEDKQSENDGDLKDTGSKEEKKAEKPPFSYNALIMMAIRGSPEKRLTLSQ  1377
            DD     +E  E  +SE++G+ K+   K+  K EKPP+SYNALIMMAIR +PEKRLTL+
Sbjct  26   DDKEAPPVERKEQVKSEHEGEKKE--QKDSGKGEKPPYSYNALIMMAIRSAPEKRLTLNG  83

Query  1378 IYEFIMKNFPYYKDNKQGWQNSIRHNLSLNKCFLKVPRHYDDPGKGNYWMLDPCCEDVFI  1557
            IYEFIMKNFPYY++NKQGWQNSIRHNLSLNKCF+KVPRHYDDPGKGNYWMLDP  +DVFI
Sbjct  84   IYEFIMKNFPYYRENKQGWQNSIRHNLSLNKCFVKVPRHYDDPGKGNYWMLDPSADDVFI  143

Query  1558 GGTTG  1572
            GGTTG
Sbjct  144  GGTTG  148


>ref|XP_003706042.1| PREDICTED: uncharacterized protein LOC100878233 [Megachile
rotundata]
Length=426
```

```
 Score =   202 bits (515),  Expect = 8e-54, Method: Compositional matrix adjust.
 Identities = 106/189 (56%), Positives = 125/189 (66%), Gaps = 18/189 (10%)
 Frame = +1

Query  1012  FSISRVLGEDIGCSPAQADDNCEKAELDGSSSFREDFVDVVSEEADDGPSFDGNISVNDD  1191
             FSI  +L E    +PA +       E+    S  ED     SE++ D       + V  D
Sbjct  27    FSIRSILPEACAGTPAPSVSRSTSPEI----SHVED-----SEDSSD-------LDVTGD  70

Query  1192  SGDDM--RDDTLEDLEDKQSENDGDLKDTGSKEEKKAEKPPFSYNALIMMAIRGSPEKRL  1365
              G++    D +        S    D KD  S E+KK EKPP+SYNALIMMAIR SPEKRL
Sbjct  71    GGNETPPLDCSRNATNSVSSSEPKDTKDRQSDEKKKCEKPPYSYNALIMMAIRQSPEKRL  130

Query  1366  TLSQIYEFIMKNFPYYKDNKQGWQNSIRHNLSLNKCFLKVPRHYDDPGKGNYWMLDPCCE  1545
             TL+ IYE+IM++FPYY++NKQGWQNSIRHNLSLNKCF+KVPRHYDDPGKGNYWMLDP  E
Sbjct  131   TLNGIYEYIMRHFPYYENNKQGWQNSIRHNLSLNKCFVKVPRHYDDPGKGNYWMLDPSSE  190

Query  1546  DVFIGGTTG  1572
             DVFIGGTTG
Sbjct  191   DVFIGGTTG  199


>gb|ADG26725.1| forkhead box protein G1 [Platynereis dumerilii]
Length=328

 Score =   199 bits (507),  Expect = 1e-53, Method: Compositional matrix adjust.
 Identities = 114/197 (58%), Positives = 132/197 (67%), Gaps = 10/197 (5%)
 Frame = +1

Query  1006  NPFSISRVLGEDIGCSP-----AQADDNCEKAELDGSSSFREDFVDVVSEEADDGPSFDG  1170
             +PFSIS +LGE++G +P     A    +        S           V    DD
Sbjct  14    HPFSISYMLGENLGSNPNNPNQAHHITSAAVPSPSSSIDDASSDAVDVETVDDDVTDDVH  73

Query  1171  NISVNDDSGDDMRDDT-LEDLE-DKQSENDGDLKDTGSK-EEKKAEKPPFSYNALIMMAI  1341
             N+  +DD  DD+     L   E +K+ E D   KD    + E+KKAEKPPFSYNALIMMAI
Sbjct  74    NVHYSDD--DDLESGRPLSSPEGEKKEEKDLAEKDKPKEGEQKKAEKPPFSYNALIMMAI  131

Query  1342  RGSPEKRLTLSQIYEFIMKNFPYYKDNKQGWQNSIRHNLSLNKCFLKVPRHYDDPGKGNY  1521
             R SPEKRLTL+ IYEFIMKNFPYY++NKQGWQNSIRHNLSLNKCF+KVPRHYDDPGKGNY
Sbjct  132   RSSPEKRLTLNGIYEFIMKNFPYYRENKQGWQNSIRHNLSLNKCFVKVPRHYDDPGKGNY  191

Query  1522  WMLDPCCEDVFIGGTTG  1572
             WMLDP  +DVFIGGTTG
Sbjct  192   WMLDPSSDDVFIGGTTG  208


  Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF
excluding environmental samples from WGS projects
    Posted date:  Feb 5, 2014  1:43 PM
  Number of letters in database: 12,906,967,781
  Number of sequences in database:  36,556,296

Lambda      K        H
   0.317    0.132    0.397
Gapped
Lambda      K        H
   0.267   0.0410    0.140
Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Sequences: 36556296
Number of Hits to DB: 12464349822
Number of extensions: 247037333
Number of successful extensions: 482575
Number of sequences better than 10: 78
Number of HSP's better than 10 without gapping: 0
Number of HSP's gapped: 480621
Number of HSP's successfully gapped: 156
Length of database: 12906967781
T: 12
A: 40
X1: 16 (7.3 bits)
X2: 38 (14.6 bits)
X3: 64 (24.7 bits)
S1: 41 (20.4 bits)
ka-blk-alpha gapped: 1.9
ka-blk-alpha ungapped: 0.7916
ka-blk-alpha_v gapped: 42.6028
ka-blk-alpha_v ungapped: 4.96466
ka-blk-sigma gapped: 43.6362
```

```
BLASTX 2.2.29+
Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro
A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and
David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new
generation of protein database search programs", Nucleic
Acids Res. 25:3389-3402.


RID: F8YNGPUD014


Database: All non-redundant GenBank CDS
translations+PDB+SwissProt+PIR+PRF excluding environmental samples
from WGS projects
           36,556,296 sequences; 12,906,967,781 total letters
Query= NODE_265519_length_13224_cov_52.508545

Length=3600


                                                                      Score     E
Sequences producing significant alignments:                         (Bits)  Value

gb|EOA95422.1|  Forkhead box protein B1, partial [Anas platyrh...    160   6e-42
ref|XP_006020106.1|  PREDICTED: forkhead box protein B1 [Allig...    159   4e-41
emb|CAA50745.1|  fkh-5 [Mus musculus]                                162   2e-42

ALIGNMENTS
>gb|EOA95422.1| Forkhead box protein B1, partial [Anas platyrhynchos]
Length=98

 Score =   160 bits (404),  Expect = 6e-42, Method: Compositional matrix adjust.
 Identities = 68/98 (69%), Positives = 83/98 (85%), Gaps = 0/98 (0%)
 Frame = -2

Query  2780  KPPYSYIALITMAIESSPTGMMTLNEIYQFIENRFPYFKENTQRWQNSIRHNLSLNDCFL  2601
              KPPYSYI+L  MAI+SSP  M+ L+EIY+FI +RFPY++ENTQRWQNS+RHNLS NDCF+
Sbjct  1     KPPYSYISLTAMAIQSSPEKMLPLSEIYKFIMDRFPYYRENTQRWQNSLRHNLSFNDCFI  60

Query  2600  KVSKNAGKPGKGNYWALHPKAGDMFGNGSFLRRSKRFK  2487
              K+ +   +PGKG++WALHP  GDMF NGSFLRR KRFK
Sbjct  61    KIPRRPDQPGKGSFWALHPSCGDMFENGSFLRRRKRFK  98


>ref|XP_006020106.1| PREDICTED: forkhead box protein B1 [Alligator sinensis]
Length=136

 Score =   159 bits (402),  Expect = 4e-41, Method: Compositional matrix adjust.
 Identities = 67/103 (65%), Positives = 85/103 (83%), Gaps = 0/103 (0%)
 Frame = -2

Query  2792  FAEVKPPYSYIALITMAIESSPTGMMTLNEIYQFIENRFPYFKENTQRWQNSIRHNLSLN  2613
              +++ KPPYSYI+L  MAI+SS   M+ L+EIY+FI +RFPY++ENTQRWQNS+RHNLS N
Sbjct  9     YSDQKPPYSYISLTAMAIQSSAEKMLPLSEIYKFIMDRFPYYRENTQRWQNSLRHNLSFN  68

Query  2612  DCFLKVSKNAGKPGKGNYWALHPKAGDMFGNGSFLRRSKRFKT  2484
              DCF+K+ +   +PGKG++WALHP  GDMF NGSFLRR KRFK
Sbjct  69    DCFIKIPRRPDQPGKGSFWALHPDCGDMFENGSFLRRRKRFKV  111


>emb|CAA50745.1| fkh-5 [Mus musculus]
Length=111

 Score =   162 bits (409),  Expect = 2e-42, Method: Compositional matrix adjust.
 Identities = 68/103 (66%), Positives = 86/103 (83%), Gaps = 0/103 (0%)
 Frame = -2

Query  2792  FAEVKPPYSYIALITMAIESSPTGMMTLNEIYQFIENRFPYFKENTQRWQNSIRHNLSLN  2613
              +++ KPPYSYI+L  MAI+SSP  M+ L+EIY+FI +RFPY++ENTQRWQNS+RHNLS N
Sbjct  6     YSDQKPPYSYISLTAMAIQSSPEKMLPLSEIYKFIMDRFPYYRENTQRWQNSLRHNLSFN  65

Query  2612  DCFLKVSKNAGKPGKGNYWALHPKAGDMFGNGSFLRRSKRFKT  2484
              DCF+K+ +   +PGKG++WALHP  GDMF NGSFLRR KRFK
Sbjct  66    DCFIKIPRRPDQPGKGSFWALHPSCGDMFENGSFLRRRKRFKV  108


Query= NODE_440046_length_40245_cov_35.859535
```

Length=7200

Sequences producing significant alignments:	Score (Bits)	E Value
gb|ESP00592.1| hypothetical protein LOTGIDRAFT_186344, partia...	150	6e-38
gb|EPY38106.1| flagellar associated protein [Angomonas deanei]	86.3	1e-37
gb|ABX89143.1| forkhead B [Patiria miniata]	152	2e-37
ref|XP_005089018.1| PREDICTED: forkhead box protein B1-like [...	155	1e-36
ref|NP_001158435.1| forkhead box B1 [Saccoglossus kowalevskii...	154	1e-36
ref|NP_999797.1| winged helix transcription factor Forkhead-l...	152	7e-36
gb|EKC18729.1| Forkhead box protein B1 [Crassostrea gigas]	151	8e-36
gb|EOA95422.1| Forkhead box protein B1, partial [Anas platyrh...	144	8e-36
ref|XP_006020106.1| PREDICTED: forkhead box protein B1 [Allig...	145	9e-36
emb|CAA50745.1| fkh-5 [Mus musculus]	144	9e-36

ALIGNMENTS
>gb|ESP00592.1| hypothetical protein LOTGIDRAFT_186344, partial [Lottia gigantea]
Length=110

 Score = 150 bits (380), Expect = 6e-38, Method: Compositional matrix adjust.
 Identities = 70/75 (93%), Positives = 72/75 (96%), Gaps = 0/75 (0%)
 Frame = -1

```
Query  7200  SYIALTAMAIQASNEKMLPLSDIYKFIMDKFPYYRKNTQKWQNSLRHNLSFNDCFIKIPR  7021
             SYIALTAMAIQAS EKMLPLSDIYKFIMD FP+YRKNTQ+WQNSLRHNLSFNDCFIKIPR
Sbjct  17    SYIALTAMAIQASGEKMLPLSDIYKFIMDNFPFYRKNTQRWQNSLRHNLSFNDCFIKIPR  76

Query  7020  RQDRPGKGSYWALHP  6976
             R DRPGKGSYWALHP
Sbjct  77    RPDRPGKGSYWALHP  91
```

>gb|EPY38106.1| flagellar associated protein [Angomonas deanei]
Length=698

 Score = 86.3 bits (212), Expect(3) = 1e-37, Method: Compositional matrix adjust.
 Identities = 40/92 (43%), Positives = 65/92 (71%), Gaps = 0/92 (0%)
 Frame = -1

```
Query  1191  QLKII*KSRAGLSLLAYCHYSLQDFVNASDCYEQLSQMYPEQDQYKLYFAQSLYKCGLNT  1012
             QL+   KSRA LSL+AYC+Y   D+  +S+ Y++L ++ P  ++Y++Y+AQ+LYK GL
Sbjct  82    QLEDFPKSRAALSLIAYCYYMQGDYAESSNFYDELVKLCPGVEEYRVYYAQALYKAGLYQ  141

Query  1011  EAMKICSQIETPSLQFKVIQLQAAIKYAEDDI  916
             E+ K+C+ IE+   Q ++ +LQA + Y +DDI
Sbjct  142   ESSKVCTMIESEQFQTRLTKLQAGVAYEQDDI  173
```

 Score = 64.7 bits (156), Expect(3) = 1e-37, Method: Compositional matrix adjust.
 Identities = 38/88 (43%), Positives = 64/88 (73%), Gaps = 0/88 (0%)
 Frame = -2

```
Query  857  KDCIEKSPSDDASTEMNKACILFKEKNYekalekfkkakKIIGNRANISYNIAVCYYQLK  678
            K  +EKS  +D  T + + C+L++E +YE+ALEKFK+A+  +G + ++ YNIA+ YY+++
Sbjct  177  KAFLEKSAVEDPDTIVLQGCVLYQEGSYEEALEKFKEAQTALGPKMDLVYNIALTYYRMQ  236

Query  677  QYDYSLRHIADIIEKGIKENPGIETRSN  594
            QY  SL  IA+II+KG++++P +   SN
Sbjct  237  QYGSSLSQIAEIIDKGVQDHPELGIGSN  264
```

 Score = 55.8 bits (133), Expect(3) = 1e-37, Method: Compositional matrix adjust.
 Identities = 30/70 (43%), Positives = 40/70 (57%), Gaps = 2/70 (3%)
 Frame = -3

```
Query  559  ELSVGMQTDGIEVASVGNTLVLHESCLVEAFNLKAAIQFNLKNSKILSENNKI*GFHIKC  380
            EL +G  T+G E  SVGNT +L ES L+EAFNLKAAI F +K      E+  +     +
Sbjct  258  ELGIGSNTEGNEARSVGNTQLLKESALIEAFNLKAAIDFQMKKPAEAKED--LADMPPRT  315

Query  379  EHSITKVVIH  350
            E  +  V +H
Sbjct  316  EEELDPVTLH  325
```

>gb|ABX89143.1| forkhead B [Patiria miniata]
Length=206

```
 Score =   152 bits (384),  Expect = 2e-37, Method: Compositional matrix adjust.
 Identities = 69/75 (92%), Positives = 72/75 (96%), Gaps = 0/75 (0%)
 Frame = -1

Query  7200  SYIALTAMAIQASNEKMLPLSDIYKFIMDKFPYYRKNTQKWQNSLRHNLSFNDCFIKIPR  7021
             SYI+LTAMAIQ S EKMLPLSDIYKFIMD+FPYYRKNTQ+WQNSLRHNLSFNDCFIKIPR
Sbjct  17    SYISLTAMAIQNSGEKMLPLSDIYKFIMDRFPYYRKNTQRWQNSLRHNLSFNDCFIKIPR  76

Query  7020  RQDRPGKGSYWALHP  6976
             R DRPGKGSYWALHP
Sbjct  77    RPDRPGKGSYWALHP  91


>ref|XP_005089018.1| PREDICTED: forkhead box protein B1-like [Aplysia californica]
Length=353

 Score =   155 bits (391),  Expect = 1e-36, Method: Compositional matrix adjust.
 Identities = 69/75 (92%), Positives = 72/75 (96%), Gaps = 0/75 (0%)
 Frame = -1

Query  7200  SYIALTAMAIQASNEKMLPLSDIYKFIMDKFPYYRKNTQKWQNSLRHNLSFNDCFIKIPR  7021
             SYIALTAMAIQ+S EKMLPLSDIYKFIMD FP+YRKNTQ+WQNSLRHNLSFNDCFIKIPR
Sbjct  17    SYIALTAMAIQSSGEKMLPLSDIYKFIMDNFPFYRKNTQRWQNSLRHNLSFNDCFIKIPR  76

Query  7020  RQDRPGKGSYWALHP  6976
             R DRPGKGSYWALHP
Sbjct  77    RPDRPGKGSYWALHP  91


>ref|NP_001158435.1| forkhead box B1 [Saccoglossus kowalevskii]
 gb|ACH68432.1| forkhead box B protein [Saccoglossus kowalevskii]
Length=324

 Score =   154 bits (389),  Expect = 1e-36, Method: Compositional matrix adjust.
 Identities = 69/75 (92%), Positives = 73/75 (97%), Gaps = 0/75 (0%)
 Frame = -1

Query  7200  SYIALTAMAIQASNEKMLPLSDIYKFIMDKFPYYRKNTQKWQNSLRHNLSFNDCFIKIPR  7021
             SYIALTAMAIQ+S EKMLPLSDIYKFIMD+FP+YRKNTQ+WQNSLRHNLSFNDCFIKIPR
Sbjct  17    SYIALTAMAIQSSTEKMLPLSDIYKFIMDRFPFYRKNTQRWQNSLRHNLSFNDCFIKIPR  76

Query  7020  RQDRPGKGSYWALHP  6976
             R DRPGKGSYWALHP
Sbjct  77    RPDRPGKGSYWALHP  91


>ref|NP_999797.1| winged helix transcription factor Forkhead-1 [Strongylocentrotus
purpuratus]
 gb|AAD34014.1|AF149706_1 winged helix transcription factor Forkhead-1
[Strongylocentrotus
purpuratus]
Length=360

 Score =   152 bits (385),  Expect = 7e-36, Method: Compositional matrix adjust.
 Identities = 68/75 (91%), Positives = 73/75 (97%), Gaps = 0/75 (0%)
 Frame = -1

Query  7200  SYIALTAMAIQASNEKMLPLSDIYKFIMDKFPYYRKNTQKWQNSLRHNLSFNDCFIKIPR  7021
             SYI+LTAMAIQ+S EKMLPLSDIYKFIMD+FPYYRKNTQ+WQNSLRHNLSFNDCF+KIPR
Sbjct  17    SYISLTAMAIQSSQEKMLPLSDIYKFIMDRFPYYRKNTQRWQNSLRHNLSFNDCFLKIPR  76

Query  7020  RQDRPGKGSYWALHP  6976
             R DRPGKGSYWALHP
Sbjct  77    RPDRPGKGSYWALHP  91


>gb|EKC18729.1| Forkhead box protein B1 [Crassostrea gigas]
Length=302

 Score =   151 bits (381),  Expect = 8e-36, Method: Compositional matrix adjust.
 Identities = 68/75 (91%), Positives = 72/75 (96%), Gaps = 0/75 (0%)
 Frame = -1

Query  7200  SYIALTAMAIQASNEKMLPLSDIYKFIMDKFPYYRKNTQKWQNSLRHNLSFNDCFIKIPR  7021
             SYIALTAMAIQ S EKMLPLSDIYKFIMD+FP+YR+NTQ+WQNSLRHNLSFNDCFIKIPR
Sbjct  17    SYIALTAMAIQNSAEKMLPLSDIYKFIMDRFPFYRQNTQRWQNSLRHNLSFNDCFIKIPR  76

Query  7020  RQDRPGKGSYWALHP  6976
             R DRPGKGSYWALHP
```

```
Sbjct  77    RPDRPGKGSYWALHP  91


>gb|EOA95422.1| Forkhead box protein B1, partial [Anas platyrhynchos]
Length=98

 Score =   144 bits (362),  Expect = 8e-36, Method: Compositional matrix adjust.
 Identities = 65/75 (87%), Positives = 73/75 (97%), Gaps = 0/75 (0%)
 Frame = -1

Query  7200  SYIALTAMAIQASNEKMLPLSDIYKFIMDKFPYYRKNTQKWQNSLRHNLSFNDCFIKIPR  7021
              SYI+LTAMAIQ+S EKMLPLS+IYKFIMD+FPYYR+NTQ+WQNSLRHNLSFNDCFIKIPR
Sbjct  5     SYISLTAMAIQSSPEKMLPLSEIYKFIMDRFPYYRENTQRWQNSLRHNLSFNDCFIKIPR  64

Query  7020  RQDRPGKGSYWALHP  6976
              R D+PGKGS+WALHP
Sbjct  65    RPDQPGKGSFWALHP  79


 Score = 38.5 bits (88),  Expect = 8.5, Method: Compositional matrix adjust.
 Identities = 15/18 (83%), Positives = 17/18 (94%), Gaps = 0/18 (0%)
 Frame = -3

Query  4639  CGNMFDNGSLLRRRKRFK  4586
              CG+MF+NGS LRRRKRFK
Sbjct  81    CGDMFENGSFLRRRKRFK  98


>ref|XP_006020106.1| PREDICTED: forkhead box protein B1 [Alligator sinensis]
Length=136

 Score =   145 bits (366),  Expect = 9e-36, Method: Compositional matrix adjust.
 Identities = 65/75 (87%), Positives = 73/75 (97%), Gaps = 0/75 (0%)
 Frame = -1

Query  7200  SYIALTAMAIQASNEKMLPLSDIYKFIMDKFPYYRKNTQKWQNSLRHNLSFNDCFIKIPR  7021
              SYI+LTAMAIQ+S EKMLPLS+IYKFIMD+FPYYR+NTQ+WQNSLRHNLSFNDCFIKIPR
Sbjct  17    SYISLTAMAIQSSAEKMLPLSEIYKFIMDRFPYYRENTQRWQNSLRHNLSFNDCFIKIPR  76

Query  7020  RQDRPGKGSYWALHP  6976
              R D+PGKGS+WALHP
Sbjct  77    RPDQPGKGSFWALHP  91


>emb|CAA50745.1| fkh-5 [Mus musculus]
Length=111

 Score =   144 bits (363),  Expect = 9e-36, Method: Compositional matrix adjust.
 Identities = 65/75 (87%), Positives = 73/75 (97%), Gaps = 0/75 (0%)
 Frame = -1

Query  7200  SYIALTAMAIQASNEKMLPLSDIYKFIMDKFPYYRKNTQKWQNSLRHNLSFNDCFIKIPR  7021
              SYI+LTAMAIQ+S EKMLPLS+IYKFIMD+FPYYR+NTQ+WQNSLRHNLSFNDCFIKIPR
Sbjct  14    SYISLTAMAIQSSPEKMLPLSEIYKFIMDRFPYYRENTQRWQNSLRHNLSFNDCFIKIPR  73

Query  7020  RQDRPGKGSYWALHP  6976
              R D+PGKGS+WALHP
Sbjct  74    RPDQPGKGSFWALHP  88


 Score = 40.4 bits (93),  Expect = 2.0, Method: Compositional matrix adjust.
 Identities = 16/20 (80%), Positives = 18/20 (90%), Gaps = 0/20 (0%)
 Frame = -3

Query  4639  CGNMFDNGSLLRRRKRFKTL  4580
              CG+MF+NGS LRRRKRFK L
Sbjct  90    CGDMFENGSFLRRRKRFKVL  109


Query= NODE_2802840_length_447_cov_47.935123

Length=467

                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value


ALIGNMENTS
```

```
  Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF
excluding environmental samples from WGS projects
    Posted date:  Feb 5, 2014  1:43 PM
  Number of letters in database: 12,906,967,781
  Number of sequences in database:  36,556,296

Lambda      K        H
   0.318    0.134    0.401
Gapped
Lambda      K        H
   0.267   0.0410    0.140
Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Sequences: 36556296
Number of Hits to DB: 8139247339
Number of extensions: 129717475
Number of successful extensions: 242958
Number of sequences better than 10: 47
Number of HSP's better than 10 without gapping: 0
Number of HSP's gapped: 242808
Number of HSP's successfully gapped: 127
Length of database: 12906967781
T: 12
A: 40
X1: 16 (7.3 bits)
X2: 38 (14.6 bits)
X3: 64 (24.7 bits)
S1: 41 (20.4 bits)
ka-blk-alpha gapped: 1.9
ka-blk-alpha ungapped: 0.7916
ka-blk-alpha_v gapped: 42.6028
ka-blk-alpha_v ungapped: 4.96466
ka-blk-sigma gapped: 43.6362




BLASTX 2.2.29+
Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro
A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and
David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new
generation of protein database search programs", Nucleic
Acids Res. 25:3389-3402.


RID: F8YNGPUD014


Database: All non-redundant GenBank CDS
translations+PDB+SwissProt+PIR+PRF excluding environmental samples
from WGS projects
           36,556,296 sequences; 12,906,967,781 total letters
Query= NODE_265519_length_13224_cov_52.508545

Length=3600


                                                                      Score     E
Sequences producing significant alignments:                         (Bits)  Value


ALIGNMENTS
Query= NODE_440046_length_40245_cov_35.859535

Length=7200


                                                                      Score     E
Sequences producing significant alignments:                         (Bits)  Value


ALIGNMENTS
Query= NODE_2802840_length_447_cov_47.935123

Length=467
```

```
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

gb|ADB22674.1|  fork-head box N2/3 transcription factor [Sacco...    122   9e-31
ref|XP_002739254.1|  PREDICTED: fork-head box N2/3 transcripti...    123   3e-30
ref|XP_002597237.1|  hypothetical protein BRAFLDRAFT_148243 [B...    117   8e-30
gb|ADA79649.1|  Foxn2/3b forkhead box transcription factor, pa...    117   2e-29
gb|AFS65550.1|  FoxN2/3, partial [Parastichopus parvimensis]         117   9e-29
gb|ESP02315.1|  hypothetical protein LOTGIDRAFT_138633 [Lottia...    115   1e-28
gb|ACE79140.1|  winged helix/forkhead transcription factor Fox...    119   1e-28
gb|EKC18744.1|  Forkhead box protein N3 [Crassostrea gigas]          118   2e-28
ref|XP_004698812.1|  PREDICTED: forkhead box protein N3 isofor...    116   6e-28
ref|XP_004698813.1|  PREDICTED: forkhead box protein N3 isofor...    116   6e-28

ALIGNMENTS
>gb|ADB22674.1| fork-head box N2/3 transcription factor [Saccoglossus kowalevskii]
Length=278

 Score =   122 bits (305),  Expect = 9e-31, Method: Compositional matrix adjust.
 Identities = 57/101 (56%), Positives = 75/101 (74%), Gaps = 4/101 (4%)
 Frame = +3

Query  27   QDATTKPPYTFSCLIFLAIESSCNKRLCVKEINQWIVDNIAYYKNVPSGSWKNSIRFNLS  206
            Q   +KPP++FSCLIF+A+E S NKRL VK+I QWI+D+  Y++N P+G WKNS+R NLS
Sbjct  117  QQLNSKPPFSFSCLIFMAVEDSPNKRLPVKDIYQWILDHFPYFQNAPTG-WKNSVRHNLS  175

Query  207  SNQCFSKVDKNLLTMRDFSGKGSLWCINPIYRPMLLESLAR  329
             N+CF KV+K         GKGSLWCI+P YRP LL++L +
Sbjct  176  LNKCFKKVEKE---KGQTIGKGSLWCIDPEYRPNLLQALKK  213


>ref|XP_002739254.1| PREDICTED: fork-head box N2/3 transcription factor [Saccoglossus
kowalevskii]
Length=510

 Score =   123 bits (309),  Expect = 3e-30, Method: Compositional matrix adjust.
 Identities = 57/101 (56%), Positives = 75/101 (74%), Gaps = 4/101 (4%)
 Frame = +3

Query  27   QDATTKPPYTFSCLIFLAIESSCNKRLCVKEINQWIVDNIAYYKNVPSGSWKNSIRFNLS  206
            Q   +KPP++FSCLIF+A+E S NKRL VK+I QWI+D+  Y++N P+G WKNS+R NLS
Sbjct  117  QQLNSKPPFSFSCLIFMAVEDSPNKRLPVKDIYQWILDHFPYFQNAPTG-WKNSVRHNLS  175

Query  207  SNQCFSKVDKNLLTMRDFSGKGSLWCINPIYRPMLLESLAR  329
             N+CF KV+K         GKGSLWCI+P YRP LL++L +
Sbjct  176  LNKCFKKVEKE---KGQTIGKGSLWCIDPEYRPNLLQALKK  213


>ref|XP_002597237.1| hypothetical protein BRAFLDRAFT_148243 [Branchiostoma floridae]
 gb|EEN53249.1| hypothetical protein BRAFLDRAFT_148243 [Branchiostoma floridae]
Length=191

 Score =   117 bits (294),  Expect = 8e-30, Method: Compositional matrix adjust.
 Identities = 57/102 (56%), Positives = 73/102 (72%), Gaps = 4/102 (4%)
 Frame = +3

Query  24   HQDATTKPPYTFSCLIFLAIESSCNKRLCVKEINQWIVDNIAYYKNVPSGSWKNSIRFNL  203
            HQ   +KPPY+FSCLIF+AIE S +KRL VKEI  WI+++  Y+ N P+G WKNS+R NL
Sbjct  73   HQHRNSKPPYSFSCLIFMAIEDSPSKRLPVKEIYNWILEHFPYFVNAPTG-WKNSVRHNL  131

Query  204  SSNQCFSKVDKNLLTMRDFSGKGSLWCINPIYRPMLLESLAR  329
            S N+CF KV+K         GKGSLW I+P YRP LL++L +
Sbjct  132  SLNKCFKKVEKE---KGQSIGKGSLWMIDPAYRPNLLQALKK  170


>gb|ADA79649.1| Foxn2/3b forkhead box transcription factor, partial [Patiria
miniata]
Length=219

 Score =   117 bits (293),  Expect = 2e-29, Method: Compositional matrix adjust.
 Identities = 59/110 (54%), Positives = 75/110 (68%), Gaps = 9/110 (8%)
 Frame = +3

Query  24   HQDA--------TTKPPYTFSCLIFLAIESSCNKRLCVKEINQWIVDNIAYYKNVPSGSW  179
            HQD          +KPP++FSCLIF+AIE S  KRL VK+I QWI D+  Y+ N P+G W
Sbjct  43   HQDGPVDPKRHINSKPPFSFSCLIFMAIEDSPLKRLPVKDIYQWIQDHFPYFHNAPTG-W  101

Query  180  KNSIRFNLSSNQCFSKVDKNLLTMRDFSGKGSLWCINPIYRPMLLESLAR  329
            KNS+R NLS N+CF KVDK    +    GKGSLWC++P YRP LL++L +
```

```
Sbjct  102  KNSVRHNLSLNKCFRKVDKIKGQISLSIGKGSLWCVDPDYRPNLLQALRK  151


>gb|AFS65550.1| FoxN2/3, partial [Parastichopus parvimensis]
Length=281

 Score =   117 bits (292),  Expect = 9e-29, Method: Compositional matrix adjust.
 Identities = 54/97 (56%), Positives = 73/97 (75%), Gaps = 3/97 (3%)
 Frame = +3

Query  39   TKPPYTFSCLIFLAIESSCNKRLCVKEINQWIVDNIAYYKNVPSGSWKNSIRFNLSSNQC  218
            +KPP++FSCLIF+AIE S ++RL VK+I  WI ++  YY+  P+G WKNS+R NLS N+C
Sbjct  104  SKPPFSFSCLIFMAIEDSLHQRLPVKDIYHWIEEHFPYYRTAPAG-WKNSVRHNLSLNKC  162

Query  219  FSKVDKNLLTMRDFSGKGSLWCINPIYRPMLLESLAR  329
            F KVDK  L  +   GKGSLWC++P YRP LL++L +
Sbjct  163  FQKVDK--LRGQQTLGKGSLWCVDPDYRPNLLQALRK  197


>gb|ESP02315.1| hypothetical protein LOTGIDRAFT_138633 [Lottia gigantea]
Length=208

 Score =   115 bits (287),  Expect = 1e-28, Method: Compositional matrix adjust.
 Identities = 55/97 (57%), Positives = 71/97 (73%), Gaps = 4/97 (4%)
 Frame = +3

Query  39   TKPPYTFSCLIFLAIESSCNKRLCVKEINQWIVDNIAYYKNVPSGSWKNSIRFNLSSNQC  218
            +KPPY+FSCLIF+A+E S  KRL VK+I  WI+ +  Y++N P+G WKNS+R NLS N+C
Sbjct  89   SKPPYSFSCLIFMAVEDSPMKRLPVKDIYNWILSHFPYFQNAPTG-WKNSVRHNLSLNKC  147

Query  219  FSKVDKNLLTMRDFSGKGSLWCINPIYRPMLLESLAR  329
            F KVDK         GKGSLWCI+P YRP LL++L +
Sbjct  148  FKKVDKE---KGQTIGKGSLWCIDPDYRPNLLQALRK  181


>gb|ACE79140.1| winged helix/forkhead transcription factor FoxN2/3 [Branchiostoma
floridae]
Length=535

 Score =   119 bits (297),  Expect = 1e-28, Method: Compositional matrix adjust.
 Identities = 57/103 (55%), Positives = 73/103 (71%), Gaps = 4/103 (4%)
 Frame = +3

Query  24   HQDATTKPPYTFSCLIFLAIESSCNKRLCVKEINQWIVDNIAYYKNVPSGSWKNSIRFNL  203
            HQ   +KPPY+FSCLIF+AIE S +KRL VKEI  WI+++  Y+ N P+G WKNS+R NL
Sbjct  120  HQHRNSKPPYSFSCLIFMAIEDSPSKRLPVKEIYNWILEHFPYFVNAPTG-WKNSVRHNL  178

Query  204  SSNQCFSKVDKNLLTMRDFSGKGSLWCINPIYRPMLLESLARI  332
            S N+CF KV+K         GKGSLW I+P YRP LL++L +
Sbjct  179  SLNKCFKKVEKE---KGQSIGKGSLWMIDPAYRPNLLQALKKT  218


>gb|EKC18744.1| Forkhead box protein N3 [Crassostrea gigas]
Length=487

 Score =   118 bits (295),  Expect = 2e-28, Method: Compositional matrix adjust.
 Identities = 57/99 (58%), Positives = 75/99 (76%), Gaps = 6/99 (6%)
 Frame = +3

Query  39   TKPPYTFSCLIFLAIESSCNKRLCVKEINQWIVDNIAYYKNVPSGSWKNSIRFNLSSNQC  218
            +KPPY+FSCLIF+AIE S +KRL VK+I  WI+++  Y+++ P+G WKNS+R NLS N+C
Sbjct  131  SKPPYSFSCLIFMAIEDSPHKRLPVKDIYSWILNHFPYFQHAPTG-WKNSVRHNLSLNKC  189

Query  219  FSKVDKNLLTMRDFS-GKGSLWCINPIYRPMLLESLARI  332
            F KVDK     R  S GKGSLWCI+P YRP LL++L +
Sbjct  190  FKKVDKE----RGQSIGKGSLWCIDPDYRPNLLQALRKT  224


>ref|XP_004698812.1| PREDICTED: forkhead box protein N3 isoform X1 [Echinops telfairi]
Length=488

 Score =   116 bits (291),  Expect = 6e-28, Method: Compositional matrix adjust.
 Identities = 58/103 (56%), Positives = 73/103 (71%), Gaps = 6/103 (6%)
 Frame = +3

Query  27   QDATTKPPYTFSCLIFLAIESSCNKRLCVKEINQWIVDNIAYYKNVPSGSWKNSIRFNLS  206
            Q+   KPPY+FSCLIF+AIE S  KRL VKEI  WI+++  Y+ N P+G WKNS+R NLS
Sbjct  109  QNPNCKPPYSFSCLIFMAIEDSPTKRLPVKEIYNWILEHFPYFANAPTG-WKNSVRHNLS  167
```

```
Query  207  SNQCFSKVDKNLLTMRDFS-GKGSLWCINPIYRPMLLESLARI  332
            N+CF KVDK     R  S GKGSLWCI+P YR  L+++L +
Sbjct  168  LNKCFKKVDKE----RSQSIGKGSLWCIDPEYRQNLIQALKKT  206


>ref|XP_004698813.1| PREDICTED: forkhead box protein N3 isoform X2 [Echinops telfairi]
Length=466

 Score =   116 bits (291),  Expect = 6e-28, Method: Compositional matrix adjust.
 Identities = 58/103 (56%), Positives = 73/103 (71%), Gaps = 6/103 (6%)
 Frame = +3

Query  27   QDATTKPPYTFSCLIFLAIESSCNKRLCVKEINQWIVDNIAYYKNVPSGSWKNSIRFNLS  206
            Q+   KPPY+FSCLIF+AIE S  KRL VKEI  WI+++  Y+ N P+G WKNS+R NLS
Sbjct  109  QNPNCKPPYSFSCLIFMAIEDSPTKRLPVKEIYNWILEHFPYFANAPTG-WKNSVRHNLS  167

Query  207  SNQCFSKVDKNLLTMRDFS-GKGSLWCINPIYRPMLLESLARI  332
            N+CF KVDK     R  S GKGSLWCI+P YR  L+++L +
Sbjct  168  LNKCFKKVDKE----RSQSIGKGSLWCIDPEYRQNLIQALKKT  206


  Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF
excluding environmental samples from WGS projects
    Posted date:  Feb 5, 2014  1:43 PM
  Number of letters in database: 12,906,967,781
  Number of sequences in database:  36,556,296

Lambda      K        H
   0.318    0.134    0.401
Gapped
Lambda      K        H
   0.267   0.0410    0.140
Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Sequences: 36556296
Number of Hits to DB: 8139247339
Number of extensions: 129717475
Number of successful extensions: 242958
Number of sequences better than 10: 47
Number of HSP's better than 10 without gapping: 0
Number of HSP's gapped: 242808
Number of HSP's successfully gapped: 127
Length of database: 12906967781
T: 12
A: 40
X1: 16 (7.3 bits)
X2: 38 (14.6 bits)
X3: 64 (24.7 bits)
S1: 41 (20.4 bits)
ka-blk-alpha gapped: 1.9
ka-blk-alpha ungapped: 0.7916
ka-blk-alpha_v gapped: 42.6028
ka-blk-alpha_v ungapped: 4.96466
ka-blk-sigma gapped: 43.6362


BLASTX 2.2.29+
Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro
A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and
David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new
generation of protein database search programs", Nucleic
Acids Res. 25:3389-3402.


RID: F8YNGPUD014


Database: All non-redundant GenBank CDS
translations+PDB+SwissProt+PIR+PRF excluding environmental samples
from WGS projects
           36,556,296 sequences; 12,906,967,781 total letters
Query= NODE_265519_length_13224_cov_52.508545

Length=3600


                                                                  Score     E
Sequences producing significant alignments:                      (Bits)  Value
```

```
gb|EKC36648.1|  Forkhead box protein B1 [Crassostrea gigas]          220   3e-60
ref|XP_005102249.1|  PREDICTED: forkhead box protein B2-like [...    222   2e-59
gb|ELU18648.1|  hypothetical protein CAPTEDRAFT_131123, partia...    208   1e-58
gb|ESP01313.1|  hypothetical protein LOTGIDRAFT_99760, partial...    191   9e-53
ref|XP_003727565.1|  PREDICTED: forkhead box protein A4 [Stron...    196   3e-51
ref|NP_001164676.1|  fork-head box A/B transcription factor [S...    184   1e-47
ref|XP_001631592.1|  predicted protein [Nematostella vectensis...    171   2e-45
gb|AFP87438.1|  forkhead domain protein A-B-like protein [Nema...    177   2e-45
emb|CBY19345.1|  unnamed protein product [Oikopleura dioica]        172   3e-43
gb|ESO06836.1|  hypothetical protein HELRODRAFT_156792 [Helobd...    164   5e-43

ALIGNMENTS
>gb|EKC36648.1| Forkhead box protein B1 [Crassostrea gigas]
Length=318

 Score =   220 bits (560),  Expect = 3e-60, Method: Compositional matrix adjust.
 Identities = 98/137 (72%), Positives = 117/137 (85%), Gaps = 1/137 (1%)
 Frame = -2

Query  2861  TEVDHVDLSHFSGTSTISQQKRRFAEVKPPYSYIALITMAIESSPTGMMTLNEIYQFIEN  2682
              +++D  D+  +SGTST+SQQKRRFA+VKPPYSYIALITM+IESS +GMMTLNEIY FI N
Sbjct  52    SDMDEYDMK-YSGTSTLSQQKRRFADVKPPYSYIALITMSIESSTSGMMTLNEIYAFIMN  110

Query  2681  RFPYFKENTQRWQNSIRHNLSLNDCFLKVSKNAGKPGKGNYWALHPKAGDMFGNGSFLRR  2502
              RFPYFK+N QRWQNSIRHNLSLNDCF+K+ +  G+PGKGNYWALHP  GDMFGNGSFLRR
Sbjct  111   RFPYFKDNQQRWQNSIRHNLSLNDCFVKIPRAPGRPGKGNYWALHPGCGDMFGNGSFLRR  170

Query  2501  SKRFKTSSPHKKDELNI  2451
              +KRFK     ++D  ++
Sbjct  171   AKRFKIQRQKREDPAHV  187


>ref|XP_005102249.1| PREDICTED: forkhead box protein B2-like [Aplysia californica]
Length=472

 Score =   222 bits (566),  Expect = 2e-59, Method: Compositional matrix adjust.
 Identities = 98/127 (77%), Positives = 112/127 (88%), Gaps = 0/127 (0%)
 Frame = -2

Query  2831  FSGTSTISQQKRRFAEVKPPYSYIALITMAIESSPTGMMTLNEIYQFIENRFPYFKENTQ  2652
              +SGTST+SQQ RRFA+VKPPYSYIALITMAIESS +GMMTLNEIY FI NRFPYFKEN Q
Sbjct  131   YSGTSTLSQQARRFADVKPPYSYIALITMAIESSHSGMMTLNEIYSFIMNRFPYFKENQQ  190

Query  2651  RWQNSIRHNLSLNDCFLKVSKNAGKPGKGNYWALHPKAGDMFGNGSFLRRSKRFKTSSPH  2472
              RWQNSIRHNLSLNDCF+K+++  G+PGKGNYWALHP  GDMFGNGSFLRR+KRFK + P
Sbjct  191   RWQNSIRHNLSLNDCFVKIARAPGRPGKGNYWALHPACGDMFGNGSFLRRAKRFKLTRPK  250

Query  2471  KKDELNI  2451
               ++  +I
Sbjct  251   SENSSHI  257


>gb|ELU18648.1| hypothetical protein CAPTEDRAFT_131123, partial [Capitella teleta]
Length=118

 Score =   208 bits (529),  Expect = 1e-58, Method: Compositional matrix adjust.
 Identities = 93/114 (82%), Positives = 103/114 (90%), Gaps = 0/114 (0%)
 Frame = -2

Query  2828  SGTSTISQQKRRFAEVKPPYSYIALITMAIESSPTGMMTLNEIYQFIENRFPYFKENTQR  2649
              SGTST+SQQKRRFA+VKPPYSYIALITM++ESS +GMMTLNEIY FI  RFPYFK+N QR
Sbjct  2     SGTSTLSQQKRRFADVKPPYSYIALITMSLESSTSGMMTLNEIYAFIMKRFPYFKDNQQR  61

Query  2648  WQNSIRHNLSLNDCFLKVSKNAGKPGKGNYWALHPKAGDMFGNGSFLRRSKRFK  2487
              WQNSIRHNLSLNDCFLK+ +  G+PGKGNYWALHP  GDMF NGSFLRR+KRFK
Sbjct  62    WQNSIRHNLSLNDCFLKIPRAPGRPGKGNYWALHPSCGDMFANGSFLRRAKRFK  115


>gb|ESP01313.1| hypothetical protein LOTGIDRAFT_99760, partial [Lottia gigantea]
Length=105

 Score =   191 bits (484),  Expect = 9e-53, Method: Compositional matrix adjust.
 Identities = 85/102 (83%), Positives = 93/102 (91%), Gaps = 0/102 (0%)
 Frame = -2

Query  2792  FAEVKPPYSYIALITMAIESSPTGMMTLNEIYQFIENRFPYFKENTQRWQNSIRHNLSLN  2613
              FA+VKPPYSYIALITMAIESS +GMMTLNEIY FI NRFPYFK+N QRWQNSIRHNLSLN
Sbjct  1     FADVKPPYSYIALITMAIESSTSGMMTLNEIYAFIMNRFPYFKQNQQRWQNSIRHNLSLN  60
```

```
Query  2612  DCFLKVSKNAGKPGKGNYWALHPKAGDMFGNGSFLRRSKRFK  2487
             DCF+KV +  G+PGKGNYW+LHP  GDMFGNGSFLRR+KRFK
Sbjct  61    DCFMKVPRGPGRPGKGNYWSLHPSCGDMFGNGSFLRRAKRFK  102


>ref|XP_003727565.1| PREDICTED: forkhead box protein A4 [Strongylocentrotus
purpuratus]
Length=353

 Score =   196 bits (497),  Expect = 3e-51, Method: Compositional matrix adjust.
 Identities = 84/105 (80%), Positives = 94/105 (90%), Gaps = 0/105 (0%)
 Frame = -2

Query  2801  KRRFAEVKPPYSYIALITMAIESSPTGMMTLNEIYQFIENRFPYFKENTQRWQNSIRHNL  2622
             KRRFA+VKPPYSYIALITMA+E S  GMMTLNE+YQFI ++FPYF+EN QRWQNSIRHNL
Sbjct  109   KRRFADVKPPYSYIALITMALEGSKDGMMTLNEVYQFIMDKFPYFRENQQRWQNSIRHNL  168

Query  2621  SLNDCFLKVSKNAGKPGKGNYWALHPKAGDMFGNGSFLRRSKRFK  2487
             SLNDCF+KV +  G+PGKGNYWALHP  GDMF NGSFLRR+KRFK
Sbjct  169   SLNDCFIKVPRAPGRPGKGNYWALHPSCGDMFSNGSFLRRAKRFK  213


>ref|NP_001164676.1| fork-head box A/B transcription factor [Saccoglossus kowalevskii]
 gb|ADB22667.1| fork-head box A/B transcription factor [Saccoglossus kowalevskii]
Length=312

 Score =   184 bits (466),  Expect = 1e-47, Method: Compositional matrix adjust.
 Identities = 77/109 (71%), Positives = 96/109 (88%), Gaps = 0/109 (0%)
 Frame = -2

Query  2813  ISQQKRRFAEVKPPYSYIALITMAIESSPTGMMTLNEIYQFIENRFPYFKENTQRWQNSI  2634
             ++ +KRRFA+VKPPYSYIALI M++E++  GM+TLNE+Y+FI N+FPYF+EN QRWQNSI
Sbjct  67    LTNEKRRFADVKPPYSYIALIAMSLENAQDGMLTLNEVYEFIMNKFPYFRENQQRWQNSI  126

Query  2633  RHNLSLNDCFLKVSKNAGKPGKGNYWALHPKAGDMFGNGSFLRRSKRFK  2487
             RHNLSLNDCF+K+ +  G+ GKGNYWALHP A DMF NGS+LRR+KRFK
Sbjct  127   RHNLSLNDCFVKIPRAPGRAGKGNYWALHPAARDMFANGSYLRRAKRFK  175


>ref|XP_001631592.1| predicted protein [Nematostella vectensis]
 gb|EDO39529.1| predicted protein [Nematostella vectensis]
Length=116

 Score =   171 bits (432),  Expect = 2e-45, Method: Compositional matrix adjust.
 Identities = 77/102 (75%), Positives = 88/102 (86%), Gaps = 1/102 (1%)
 Frame = -2

Query  2786  EVKPPYSYIALITMAIESSPTGMMTLNEIYQFIENRFPYFKENTQRWQNSIRHNLSLNDC  2607
             EVKPP+SYIALITM+IE+SP  M TLNEIY+FI  RFPYF++N Q+WQNSIRHNLSLNDC
Sbjct  14    EVKPPFSYIALITMSIEASPYRMRTLNEIYEFIMTRFPYFRKNQQKWQNSIRHNLSLNDC  73

Query  2606  FLKVSKNA-GKPGKGNYWALHPKAGDMFGNGSFLRRSKRFKT  2484
             F+KV ++  GKPGKGNYW LHP  GDMFG+GSFLRR KRFK
Sbjct  74    FVKVPRSIFGKPGKGNYWTLHPSCGDMFGSGSFLRRPKRFKC  115


>gb|AFP87438.1| forkhead domain protein A-B-like protein [Nematostella vectensis]
Length=312

 Score =   177 bits (449),  Expect = 2e-45, Method: Compositional matrix adjust.
 Identities = 79/110 (72%), Positives = 92/110 (84%), Gaps = 1/110 (1%)
 Frame = -2

Query  2786  EVKPPYSYIALITMAIESSPTGMMTLNEIYQFIENRFPYFKENTQRWQNSIRHNLSLNDC  2607
             EVKPP+SYIALITM+IE+SP  M TLNEIY+FI  RFPYF++N Q+WQNSIRHNLSLNDC
Sbjct  90    EVKPPFSYIALITMSIEASPYRMRTLNEIYEFIMTRFPYFRKNQQKWQNSIRHNLSLNDC  149

Query  2606  FLKVSKNA-GKPGKGNYWALHPKAGDMFGNGSFLRRSKRFKTSSPHKKDE  2460
             F+KV ++  GKPGKGNYW LHP  GDMFG+GSFLRR KRFK   P + +E
Sbjct  150   FVKVPRSIFGKPGKGNYWTLHPSCGDMFGSGSFLRRPKRFKCRMPQRPNE  199


>emb|CBY19345.1| unnamed protein product [Oikopleura dioica]
Length=360

 Score =   172 bits (437),  Expect = 3e-43, Method: Compositional matrix adjust.
 Identities = 72/103 (70%), Positives = 87/103 (84%), Gaps = 0/103 (0%)
 Frame = -2
```

```
Query  2792  FAEVKPPYSYIALITMAIESSPTGMMTLNEIYQFIENRFPYFKENTQRWQNSIRHNLSLN  2613
             + + KPPYSYIAL  MAI+S+P  MMTL EIY+FI +RFPY+++NTQRWQNS+RHNLS N
Sbjct  9     YGDQKPPYSYIALTAMAIQSAPDKMMTLAEIYKFIMDRFPYYRKNTQRWQNSLRHNLSFN  68

Query  2612  DCFLKVSKNAGKPGKGNYWALHPKAGDMFGNGSFLRRSKRFKT  2484
             DCF+K+ + A KPGKG+YW+LHP  GDMF NGSFLRR KRFKT
Sbjct  69    DCFIKIPRRADKPGKGSYWSLHPSCGDMFENGSFLRRRKRFKT  111


>gb|ESO06836.1| hypothetical protein HELRODRAFT_156792 [Helobdella robusta]
Length=123

 Score =   164 bits (415),  Expect = 5e-43, Method: Compositional matrix adjust.
 Identities = 70/102 (69%), Positives = 86/102 (84%), Gaps = 0/102 (0%)
 Frame = -2

Query  2792  FAEVKPPYSYIALITMAIESSPTGMMTLNEIYQFIENRFPYFKENTQRWQNSIRHNLSLN  2613
             +++ KPPYSYIAL  MAI+SS   +MTL+EIYQFI +RFPY++ENTQRWQNS+RHNLS N
Sbjct  9     YSDQKPPYSYIALTAMAIQSSRDKIMTLSEIYQFIMDRFPYYRENTQRWQNSLRHNLSFN  68

Query  2612  DCFLKVSKNAGKPGKGNYWALHPKAGDMFGNGSFLRRSKRFK  2487
             DCF+K+ +   +PGKG+YW LHP+ GDMF NGSFLRR KRFK
Sbjct  69    DCFVKLPRRPDRPGKGSYWTLHPQCGDMFENGSFLRRRKRFK  110


Query= NODE_440046_length_40245_cov_35.859535

Length=7200


                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value


ALIGNMENTS
Query= NODE_2802840_length_447_cov_47.935123

Length=467


                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value


ALIGNMENTS
  Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF
excluding environmental samples from WGS projects
    Posted date:  Feb 5, 2014  1:43 PM
  Number of letters in database: 12,906,967,781
  Number of sequences in database:  36,556,296

Lambda      K        H
   0.318    0.134    0.401
Gapped
Lambda      K        H
   0.267   0.0410    0.140
Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Sequences: 36556296
Number of Hits to DB: 8139247339
Number of extensions: 129717475
Number of successful extensions: 242958
Number of sequences better than 10: 47
Number of HSP's better than 10 without gapping: 0
Number of HSP's gapped: 242808
Number of HSP's successfully gapped: 127
Length of database: 12906967781
T: 12
A: 40
X1: 16 (7.3 bits)
X2: 38 (14.6 bits)
X3: 64 (24.7 bits)
S1: 41 (20.4 bits)
ka-blk-alpha gapped: 1.9
ka-blk-alpha ungapped: 0.7916
ka-blk-alpha_v gapped: 42.6028
ka-blk-alpha_v ungapped: 4.96466
ka-blk-sigma gapped: 43.6362
```