THE EVOLUTION OF METAZOAN GATA TRANSCRIPTION FACTORS

by

WILLIAM JOSEPH GILLIS

A DISSERTATION

Presented to the Department of Biology
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

September 2008

University of Oregon Graduate School

Confirmation of Approval and Acceptance of Dissertation prepared by:

William Gillis

Title:

"The evolution of metazoan GATA transcription factors."

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Biology by:

Eric Johnson, Chairperson, Biology
Bruce Bowerman, Advisor, Biology
John Postlethwait, Member, Biology
Joseph Thornton, Member, Biology
Stephan Schneider, Member, Institute of Molecular Biology
John Conery, Outside Member, Computer & Information Science

and Richard Linton, Vice President for Research and Graduate Studies/Dean of the Graduate School for the University of Oregon.

September 6, 2008

Original approval signatures are on file with the Graduate School and the University of Oregon Libraries.

An Abstract of the Dissertation of

William J. Gillis for the degree of Doctor of Philosophy

in the Department of Biology to be taken September 2008

Title: THE EVOLUTION OF METAZOAN GATA TRANSCRIPTION FACTORS

Approved: _____

Dr. Bruce Bowerman

This thesis explores the origin and evolution of animal germ layers via evolutionary-developmental analyses of the GATA family of transcription factors. GATA factors identified via a conserved dual zinc-finger domain direct early germ layer specification across a wide variety of animals. However, most of these developmental roles are characterized in invertebrate models, whose rapidly evolved sequences make it difficult to reconstruct evolutionary relationships. This study reconstructs the stepwise evolution of metazoan GATA transcription factors, defining homologous developmental roles based upon clear orthology assignments.

We identified two GATA transcription factors (*PdGATA123* and *PdGATA456*) from the marine annelid *Platynereis dumerilii* to aid comparison of protostome and deuterostome GATA factors. Our phylogenetic analyses defined these as protostome orthologs of GATA1/2/3 and GATA4/5/6 vertebrate subfamilies, while the mRNA localization of the *Platynereis* GATAs showed ectodermal versus endomesodermal germ layer restrictions, similar to their vertebrate orthologs.

To define the phylogenetic relationships of more divergent genes in the invertebrate models, we identified GATA homologs from recently sequenced protostome genomes. Molecular phylogenetic analyses, comparisons of intron/exon structure, and conserved synteny confirm all protostome GATA transcription factor genes are members of either the GATA123 or GATA456 class. These data allowed us to identify multiple protostome-specific duplications of *GATA456* homologs and reconstruct the origin and relationships of all arthropod GATA genes.

To probe GATA transcription factor evolution in deuterostomes, including vertebrates, we identified GATA factors in basal deuterostomes, including the cephalochordate *Branchiostoma floridae* and the hemichordate *Saccoglossus kowalevskii*. Phylogenetic analyses of these data independently confirmed that the ancestral deuterostome and chordate - like the bilaterian ancestor - possessed only two GATA transcription factors. This work was facilitated by a bioinformatics platform we are developing to identify gene families from preassembled genomic sequence.

We generated anti-*PdGATA* antibodies to further explore the role of *Platynereis* GATAs in germ layer formation. We identified multiple presumptive endomesodermal cells in which nuclear localization of *PdGATA456* protein first occurs and utilized *PdGATA456* protein localization to follow endomesodermal cell populations throughout development. These analyses represent some of the first cellular and molecular analyses of *Platynereis* germ layer formation.

This dissertation includes both my previously published and unpublished co-authored material.

CURRICULUM VITAE

NAME OF AUTHOR: William Joseph Gillis

PLACE OF BIRTH: Cambridge, Ontario, Canada

DATE OF BIRTH: January 2nd, 1981

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon
University of Delaware

DEGREES AWARDED:

Doctor of Philosophy, 2008, University of Oregon
Bachelor of Science, 2003, University of Delaware

AREAS OF SPECIAL INTEREST:

Evolution and Development
Molecular Phylogenetics
Cell and Developmental Biology

PROFESSIONAL EXPERIENCE:

Graduate Research Fellow, Department of Biology, University of Oregon, Eugene, 2003-2008

Graduate Teaching Fellow, Department of Biology, University of Oregon, Eugene, 2003-2004

Undergraduate Research Assistant, Carl Schmidt Lab, Dept. of Animal Science, Unversity of Delaware, Newark, 2001-2003

GRANTS, AWARDS AND HONORS:

Appointed to the NSF IGERT training grant, University of Oregon 2005-2008

NIH Predoc Training Award-Embryology Course, Marine Biological Laboratory, Woods Hole, MA, 2005

John K. Scoggin Sr. Memorial Award in Bioinformatics, University of Delaware, 2002

Undergraduate Science and Engineering Scholar, University of Delaware 2000

PUBLICATIONS:

Gillis, W.Q., Bowerman B.A., Schneider S.Q. (2008) The evolution of protostome GATA factors: molecular phylogenetics, synteny, and intron/exon structure reveal orthologous relationships. *BMC Evol. Bio.*, 8(112): 1-15

Gillis W.J., Bowerman B.A., Schneider S.Q. (2007) Ectoderm- and endomesoderm-specific GATA transcription factors in the marine annelid *Platynereis dumerilii*. *Evol. Dev.* 9 (1): 39-50.

Khan S., Makkena R., McGeary F., Decker K., Gillis W.J., and Schmidt C.J. (2003) A Multi-Agent System for the Quantitive Simulation of Biological Networks". *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems*. 385-392

Khan S., Decker K., Gillis W.J., and Schmidt C.J. (2003) A Multi-Agent System-driven AI Planning Approach to Biological Pathway Discovery. *Proceedings of the International Conference on Automated Planning*. 246-255

ACKNOWLEDGMENTS

I wish to express my many thanks to Dr. Stephan Q. Schneider for providing me a stimulating research topic and for his enthusiastic support and advisement, and to Dr. Bruce Bowerman, who allowed me to pursue a unique line of research in his laboratory and provided excellent training in scientific thinking, presentation, and writing. I would like to thank the members of my committee for their guiding me on my way through my program, and especially Eric Johnson for agreeing to head my committee and introducing me to fly pushing. Thanks to Joe Thornton for sharing his expertise in phylogenetics, John Conery for his patience in teaching computer science to biologists, and John Postlethwait for training in evolution and developmental biology. Many thanks to Chris Q. Doe, Chuck Kimmel, Phil Washbourne, and Judith Eisen, for their generous sharing of microscopes and reagents, and Dr. Tom Stevens, who in his tenure as institute head encouraged me to enter into a fruitful, if atypical, research environment. I would like to acknowledge the UO IGERT evolution and development group, and especially my fellow IGERT fellows Jason Q. Boone and Sean M. Carroll for insightful discussions, encouragement in the hard times, and sharing in the good times. Thanks to all of the biology staff who have done so much for me over the years, and especially Donna Overall, Peg Morrow, Ellen McCumsey, Sara Nash, and Lisa Williams. This work was supported by the National Science Foundation IGERT Grant DGE-0504627, and the National Institutes of Health Grant GM071478.

...and thank you Laura, for everything you do...

This work is dedicated to my parents
Leonard E. and Polly P. Gillis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

The progression of a single cell to a multicellular adult organism during animal development is truly a spectacular occurrence. This single cell must undergo numerous cell divisions, many of which are asymmetric and lead to a diversity of cell types. These divisions must occur in the correct spatial and temporal pattern, and the resulting progeny must coordinate to form distinct tissues. The distinct cells and tissues migrate and undergo dramatic morphological changes, regulated both by cell autonomous and inductive events, to produce stereotyped form and generate specialized functions. The culmination of which is the production of an adult that produces the next sperm and/or egg, to give rise to the next generation. Different species of animals undergo these functions in dramatically different ways, and yet embryologists have long recognized underlying developmental principles that appear to underlie this process across species and phyla.

Our ability to decipher the genetic code of life, via DNA sequencing technology, and the subsequent birth of comparative genetics, has revealed that these common developmental principles are directed by very similar looking molecules across great evolutionary times. From a deep-sea sponge to a desk-bound scientist, animals appear to control their development using genes which have undergone surprisingly little alteration over the last half billion years of evolutionary change ("a conserved developmental

toolkit"). In a practical sense, this means that understanding the genes that regulate development in a fly or a worm can inform us about our own development.

However, despite the remarkable similarity in the types of developmental genes present in animals, many of these ancient genes have undergone frequent duplication and subsequent modification over evolutionary time, a process that must contribute to the dramatic diversity of body plans seen across animals. Indeed, most 'conserved developmental toolkit' genes differ drastically in number and sequence across animals. When analyzed in a phylogenetic framework, one can map changes in gene number, sequence, and function onto the tree of life, and reconstruct how these toolkit genes looked in various ancestors. By comparing difference seen at these ancestral nodes, we can gain a glimpse into the overall pattern of animal evolution.

My work explores the evolution of one such gene family, the GATA-family of transcription factors, within animals, and compares the expansion and modification of this family to increasingly specialized roles in cell and tissue level specification. This work has explored the evolution of this family through a careful phylogenetic analysis to determine key nodes of animal evolution. In addition, we have performed developmental analysis of this gene family with the marine annelid *Platyneries dumerilii,* whose relatively conservative rate of evolution and phylogenetic position make it an important organism for comparative analysis. This introduction provides the necessary background in comparative developmental biology to contextualize this work.

Animal relations

In order to place the development of different animals in context, we need to briefly discuss our current perspective on the relationship between animals. We have

interpreted our work based upon a phylogenic tree that we represent in Figure I.1. This tree includes several updates from the traditional views of animal phylogeny, as recent comparisons of an ever-increasing amount of molecular sequences from new taxa have allowed us to re-evaluate previous morphological or embryological characters [Alguinado 1997, Phillipe 2005, Dunn 2008].

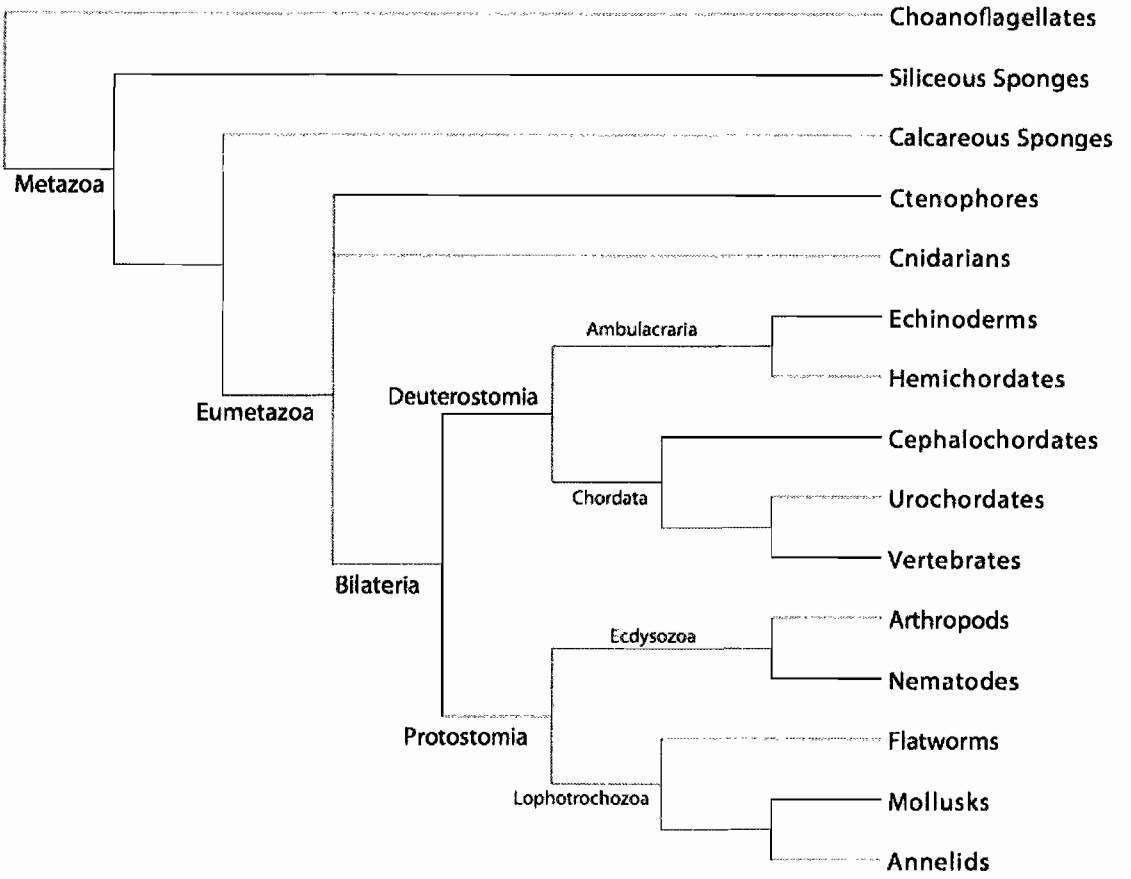

Figure I.1 Animal Phylogeny

All animals are believed to have derived from a single unicellular ancestor, with the closest outgroup being considered a group of protists called the choanoflagellates. Therefore, multicellularity in animals is thought to have originated independently of fungi, plants, and other multicellular organisms. The most basal branching group of

metazoans is the sponges, which are thought to be paraphyletic. Next we have the eumetazoa, which consist of animals with higher tissue and organ level organization, including the cnidarians (e.g. sea anemones, jellyfish, coral, hydra), the ctenophores (comb jellies), and also the bilaterians.

Bilateria, so named due to the synapomorphic character of bilateral symmetry in their left/right body axis, includes two major branches, the deuterostomes and protostomes, and has undergone several important changes recently. Nematodes and priapulids are now widely recognized as part of a monophyletic subgroup of protostomes referred to as ecdysozoans, which also includes the arthropods, tardigrades, and other molting organisms. The remaining protostomes are now united in another major monophyletic clade, the lophotrochozoans, which includes spiralians (animals which possess spiral cleavage and trochophore larvae) such as annelids and mollusks, and also lophophorates including brachiopods, and others. In deuterostomes, two groups of animals are now recognized; Ambulacraria, which include echinoderms and hemichordates, and the Chordata. Within chordates, urochordates (e.g. tunicates and larvaceans) are now considered to be the closest invertebrate group to the vertebrates, with cephalochordates representing the most basal chordate branch (see Chapter 4).

The origin and evolution of bilaterian germ layers

One of the hallmarks of eumetazoan development is organization of cells into distinct tissue layers, which embryologists refer to as germ layers. Bilaterians are considered unique from other animals in their development from three distinct germ layers; an inner endoderm, which gives rise to the midgut and related organs, an outer

ectoderm, which gives rise to skin and nervous tissue, and an intermediate mesodermal layer, which gives rise to muscle, circulatory systems including the heart, vascular system and blood, as well as bone (in vertebrates) and other organs. The evolution of a distinct mesodermal germ layer in bilaterians is thought to be an important step in the generation of increasingly complex and motile body plans in bilaterians.

We can begin to generate a picture of the evolution of the mesodermal germ layer by a comparison of the development of germ layers across bilaterians and closely related animal species. Although we are not exactly certain what other group of animals is most closely related to Bilateria, it is likely that both ctenophores and cnidarians will be important groups to study in regards to this question. Both of these groups are classically considered diploblastic, in that they only possess clear inner and outer germ layers that have been considered homologous to the bilaterian endoderm and ectoderm, respectively.

The formation of the germ layers is connected with the formation of the gastric cavity, a process called gastrulation [Technau 2003, Arendt 2004, Burton 2007]. Different groups of animals have fundamental differences in their mode of gastrulation, but the end result appears to be similar; the formation of an inner tissue layer, called the endoderm, surrounding an inner gut cavity. In bilaterians this process starts with the internalization of vegetal cells that give rise to the inner tissue layers (mesoderm and endoderm), creating an opening in the spherical embryo called the blastopore. This blastoporal space expands along the animal/vegetal axis, and eventually forms the gut cavity with mouth and anal openings at either end. Which opening the blastopore becomes varies across animals, and in fact betrays the etymological origin of the two

major bilaterian subdivisions; in protostomes ('first mouth') this blastopore becomes the mouth, and in deuterostomes ('second mouth') the anus.

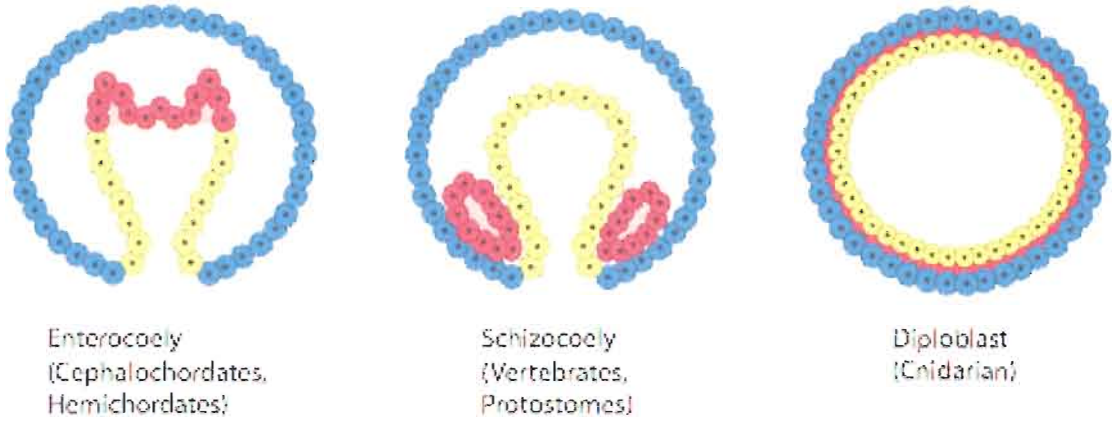

Figure I.2 Modes of Mesoderm formation

The gut cavity is surrounded by the endodermal tissue layer, but is also associated with the formation of mesoderm, which forms its own cavities refereed to as coeloms. In fact, the ontogeny of endoderm and mesoderm in most bilaterians is joint during the early phase of development as the endomesoderm, which becomes segregated during later time points. The mode in which mesoderm arises as a distinct germ layer during development can appear quite different [see Figure I.2]; in enterocoelic ('gut-coelom') animals, such as hemichordates or cephalochordates, mesoderm and endoderm internalize as one layer, and the mesoderm layer buds off as pouches from the gut cavity. However, in both vertebrates and most protostome invertebrates [Arendt 2004], mesodermal cells become a distinct inner mass of cells from the endoderm before forming an epithelial layer, which creates a new cavity via schizocoely ("split-coelom"). For instance, in spiralians, two mesodermal progenitor cells enter the embryo from the posterior edge of the blastopore,

divide and give rise to two bands of cells on either side of the embryo, which form their own coelomic spaces. Schizocoely has been suggested to be evolved due to a heterochronic shift in the timing of mesoderm formation to an earlier point of development [Arendt 2004]. In addition to mesoderm derived from endoderm, there is also a smaller portion of mesoderm which shares a common ontogeny with ectoderm in some animals; for instance, the C blastomere in *C. elegans* gives rise to only mesoderm and ectoderm, while similar ectomesodermal cells have been described in some lophotrochozoan species, as well as mesodermal derivatives of the vertebrate neural crest [Reviewed in Seipel and Schmid 1995].

However, it is unclear how the mesodermal germ layer is related within bilaterians, or how it relates to similar tissues outside of bilaterians. Some workers have suggested that the inner tissue layer of diploblasts, such as cnidarians, can be considered as bifunctional 'endomesodermal' layer [Martindale 2005]. Additionally, some workers have homologized the mesoglea, an extracellular matrix found between the diploblast germ layers which contain mesenchymal cells in some species of ctenophores and cnidarians, to the bilaterian mesoderm [Technau 2003]; however, these cells do not form a distinct germ layer in these organisms.

Platynereis dumerilii as an evo-devo model system

The marine annelid *Platynereis dumerilii*, a polychaete of the *Nereididae* family, provides a compelling model for reconstructing the ancestral developmental toolkit of bilaterian animals. As a lophotrochozoan, *Platynereis* serves as a key intermediate to help connect between our invertebrate and vertebrate model systems. Additionally,

Platynereis appears to retain many ancestral morphological and molecular features. Recent studies have demonstrated the conservative nature of *Platynereis* molecular sequence, including the retention of an ancestral gene complement, an ancestral intron-exon complexity, and generally slow mutation rates [Tessmar-Raible and Arendt 2003; Fischer and Dorresteijn 2004; Raible et al. 2005].

As a laboratory organism, *Platynereis* is amenable to embryological assays; easy to culture, and external fertilization allows for the collection of thousands of synchronously developing embryos. The seminal work of E.B.Wilson established the early cell lineage of a closely related nereid species as a model for the mosaic development of spiral-cleaving animals [Wilson 1892], that appears to be nearly identical to the cell lineage described for *Platynereis* (Figure 3) [Dorresteijn 1990, Schneider 1992, Ackerman 2005]. Similar to the nematode *C. elegans,* early asymmetrical and stereotypic cell cleavages allow for the reproducible identification of every cell in the early embryo, and to determine their invariant contribution to adult tissues. Two rounds of asymmetric meridional (through animal and vegetal pole) cleavage first result in the formation of four macromeres representing different quadrants of the embryo (A-D). These macromeres undergo four rounds of latitudinal asymmetric cleavage, giving rise to four 'quartets' of smaller animal micromeres (1a-d, 2a-d, 3a-d, 4a-d), versus large vegetal macromeres (1A-D, 2A-D, 3A-D, 4A-D) that appear to become quiescent during early gastrulation. These macromeres (4A-D) become internalized via epibolic gastrulation as continually dividing smaller micromeres spread over the vegetal macromeres. By 24 hours post fertilization (HPF), the embryo develops into a trochophore larva that swims via ciliary beating of the prototroch, a pre-oral equatorial band of ciliated cells.

Trochophore larvae possess a pair of larval eyes, and become positively photo-tactic at around 36 HPF. At 54 HPF, the larva clearly has three trunk segments, as indicated by the presence of three paired chaetal sacs, and at around three days exhibits many of the features of a juvenile worm.

Cells within the D (dorsal) quadrant appear to be key organizing centers in most spiralians, including *Platynereis*, influencing many early fate decisions of the early embryo [Dorresteijn 2005, Lambert 2008]. One of the D-descendants, the somatoblast 2d, is distinctly larger then other second quartet micromeres, and its descendant $2d^{112}$ undergoes the first bilateral cleavage to break the early spiral cleavage pattern giving rise to the progeny that will form the ventral plate epithelium. These paired dorsal somatoblasts ($2d^{1121}$, $2d^{1122}$) undergo rapid proliferation relative to other cell lineages and expand ventrally to gives rise to the entire trunk ectoderm, including the chaetal sacs, epidermis, and ventral plate neuroectoderm [Steinmetz 2007]. The mesentoblast 4d undergoes the second bilateral cleavage, and descendants of this cell give rise to the *Platynereis* paired mesodermal bands as well as a portion around the posterior end of the midgut [Wilson 1892, Rebscher 2007, Ackerman 2005]. These mesodermal bands gives rise to the paired segmental coelomic body cavities, lined via a ciliated mesothelium, and elaborate trunk muscular system, which includes paired dorsal-lateral and ventral-lateral longitudinal muscle bands, oblique muscles which overlie the ventral-lateral muscles, partial-circular muscle, and segmental muscles which include the parapodial and chaetal muscle complexes. [Rouse 20001, Tzetlin 2005, Ackerman 2005, Berger 2008].

However, some muscle/mesoderm appears to be non-4d derived. Comparisons of lineage-tracers analyses of 4d and muscle, and the distribution of mesodermal and muscle

markers such as F-actin [Ackermann 2005], *Pdu-fgfr*, and *Pdu-twist* [Steinmetz 2007, Schaub 2007] suggest that the progeny of 4d gives rise to the entirety of the trunk body-wall musculature, but also reveal additional non-4d derived 'ectomesoderm' tissues in the anterior half of the embryo. Ectomesodermal 'envelopes' appear on either side of the stomodeum [Ackerman 2005] which likely give rise to pharyngeal-related muscle, and are thought be to arise from the 3a, 3b, and/or 2b2 micromeres [Ackerman 2005, Steinmartz 2007]. Additionally, a population of non-4d anterior ectomesodermal tissue creates four horn-like paired strands of muscle at the base of the antennae, which have been suggested to arise from 3c and 3d micromeres. Therefore, it appears that anterior mesoderm has an ectomesodermal origin from multiple micromeres, in contrast to the trunk mesoderm that is derived solely from a single micromere (4d) in *Platyneries.*

Molecular control of germ-layer formation

Despite the wide variety in the mechanisms of germ-layer formation, a number of genes are involved in the patterning of endoderm and mesoderm in both protostome and deuterostome animals, including transcription factors of the *GATA*, *twist*, *snail*, and *brachyury* families (reviewed in [Technau 2003, Leptin 2005]). In *Drosophila*, *twist* and *snail* expression demarcate the mesodermal primordia, and are both transcriptionally activated by nuclear Dorsal protein [Ip 1992]. *Twist,* a bHLH transcription factor, upregulates a number of mesodermal patterning and differentiation genes, such as *tinman* and *mef2*. *Drosophila twist* mutants lack mesoderm, and also fail to gastrulate, suggesting it also plays a morphogenetic role. In vertebrates, *twist* also appears to be broadly expressed in the neuroectoderm, but is thought to be involved in the repression,

not an activation, of myogenic differentiation. Snail, a zinc finger transcription factor, is a transcriptional repressor of neuroectoderm fate in the invaginating mesoderm, and like *twist*, is also required for gastrulation movements [Nieto 2002]. However, in vertebrates, both *twist* and *snail* appear to be utilized later during morphogenesis and differentiation of the mesoderm [Leptin 2005, Technau 2003], as well as roles in regulating epithelial-mesenchymal transitions in other cell types, such as the vertebrate neural crest [Nieto 2002] . Indeed, the late expression of *twist* mRNA in the crustacean *Paryhale hawaiiensis* is suggestive of a more general later role in mesoderm differentiation, and the authors interpret the early expression in mesoderm precursors in *Drosophila* as the results of an adaptive requirement for rapid embryogenesis in this system [Price and Patel, 2007]. In vertebrates, Nodal establishes the expression of *brachyury* in vertebrate mesodermal primordial; however, Nodal is absent in both fruitflies and nematodes. In most bilaterians, *brachyury* is expressed throughout the blastopore and it's derivatives, the foregut and hindgut. However, *brachyury* is neither expressed pan-mesodermally in most invertebrates, nor in vertebrate neural crest-derived mesoderm [Marcellini 2006]. In many spiralians, activation of Erk1/2 MAP kinase (MAPK) via tyrosine diphosphorylation appears to be involved in the specification of the primary mesoderm cell (4d) or it's precursor (3D), and inhibition of this signaling has been shown to cause a loss of mesodermal derivatives in some mollusks [Lambert 2001, 2003, 2008].

Many of these mesoderm-related genes have been characterized in the cnidarians [Martindale 2005, Seipel 2005], where most (with the exception of *Mef2)* appear to be associated with the endodermal germ layer.

The role of GATA factors in germ layer patterning

The GATA family of transcription factors has been considered a 'key mesodermal gene', but appears to play many roles in germ-layer specification throughout Bilateria. . GATA transcription factors are named for their ability to bind DNA with a canonical (A/T)GATA(A/G) binding sequence [Koh 1993]. Although the GATA-type DNA-binding zinc finger has been found throughout eukaryotes, animals GATA factors appear to be uniquely defined by the presence of a conserved dual zinc finger DNA-binding domain. In contrast to the aforementioned mesodermal genes, GATA homologs have been implicated in the early specification of mesendodermal lineages in both deuterostomes and protostomes. Additionally, the GATA family has undergone significant expansion in many bilaterians, and the exact relationships of these gene duplicates in different species has been poorly understood.

The role of GATA factors in endoderm and endomesoderm specification is best understood in the nematode *C. elegans*, where six of eleven identified GATA factors are involved at some point in this process (reviewed in [Maduro 2006]). Paralogous GATAs *med-1/2* are activated by SKN-1 at the four cell stage in the EMS cell, the endomesodermal progenitor cell; disruption of zygotic MED-1/2 causes a loss of all MS-derived pharynx and body-wall muscle, whereas removal of maternal MED1/2 results in partial defects in endoderm. MED-1 can directly activate downstream GATA paralogs *end-1* and *end-3,* and disruption of END-1 and END-3 results in a transformation of the endodermal cell (E) into an ectomesodermal (C) cell fate.

A similar role for GATA-4,-5,-6 homologs in vertebrate endomesoderm and endodermal patterning has lead to the suggestion that GATAs are playing a common role

in the specification of the mesendoderm [Rodaway and Patient 1999], and that the recruitment of GATA factors into this role constitutes a key evolutionary step towards the formation of this ancient germ layer.

However, other GATA factors appear to play a role in ectodermal patterning, distinct from endodermal GATA genes. Five GATA genes in nematodes show distinct roles in ectodermal patterning. *Elt-1*, the most conserved GATA gene in nematodes, is first expressed in hypodermal precursors in the 28-cell stage [Page 1997], and is required for proper ventral motor neuron activity [Smith, McGarr 2005]. The remaining factors, including *elt-5, -6* [Koh, 2001] and *elt-1, -3* [Gilleard 2001] appear to be involved in functions suggestive of multiple germ layer roles for independent GATAs. Likewise, the vertebrates GATA –1, -2, and –3 factors appear to have little role in the early specification of endomesoderm. Although some of these factors are playing roles in mesoderm-progenitors during blood-cell differentiation [Patient 2002], GATAs –2 and -3 also show prominent early expression in the CNS, and appear to be required during the development of several neuronal populations [van Doorninck 1999, Karis 2001, van der Wees 2004, Tsarovina 2004, Pattyn 2004].

However, despite some functional similarity of individual homologs, the ancestrally conserved roles for GATA transcription factors remain ill defined. This problem stems in part from the uncertain evolutionary history of GATA factors in different bilaterian lineages; for instance, it has been unclear how six vertebrate, five fruitfly, and eleven nematode GATA genes are evolutionary related across animals [Lowry 2000, Patient 2002], and whether they originated from one or multiple GATA genes in the last common bilaterian ancestor. A large part of the current uncertainty is

due to rapid rate of evolution of both nematodes and fruitfly GATAs, which make phylogenetic reconstruction based upon their conserved sequence difficult. This ambiguity made it hard to define the ancestral roles of GATA genes and to compare their functions in mesendodermal and/or other germ layer patterning between animals.

In order to overcome the current uncertainty in the relationships between bilaterian GATAs, we describe the identification of GATA factors from many previously unexplored genomes Chapters II-IV, via PCR screening and in-silico identification in recently sequenced genomes. We have tried to select organisms for detailed analyses based both on their phylogenetic position, but also those that appear to have slow rate of evolution relative to their node. We have used these newly identified GATA factors to reconstruct the evolutionary steps leading to the expansion of this family. In Chapter II, we have examined novel genes from the polychaete *Platynereis dumerilii*, and were able to reconstruct that the last common bilaterian ancestor – the so called urbilaterian – had at least two GATA factors, similar to the vertebrate GATA1/2/3 class and the GATA4/5/6 class. In chapter III, the identification of GATA factors from many new and uncharacterized protostome genomes allowed us to identify expansions of GATA4/5/6-type genes in both lophotrochozoans and arthropods, and to reconstruct the relationships between identified insect and arthropod GATA genes. Both Chapter II and III have been previously published with my advisors Stephan Schneider and Bruce Bowerman. Finally, in chapter IV, we describe the identification of two well-conserved GATAs in a hemichordate and a cephalochordate, providing strong independent support that the basal deuterostome and chordate had two and only two GATA factors.

Additionally, we have set out to describe and compare the developmental role of GATA genes in systems that appear to have a more conservative rate of evolution. To help address the role of GATA factors within Bilateria, in Chapter III, I characterized mesodermal versus ectodermal germ layer restricted mRNA expression of GATA123 and GATA456 orthologs in *Platynereis*. I have used these genes to help establish phylogenetic links across protostome and deuterostome GATA factors, and to identify two clear classes of GATAs as common across Bilateria. In addition, we have begun the developmental characterization of two *Platynereis* GATA homologs, which appear to exhibit different germ layer restrictions in spatial mRNA expression. Furthermore, we have generated *Platynereis* specific antibodies to perform a detailed spatial and temporal analysis of *PdGATA* nuclear localization during gastrulation, representing some of the first detailed cellular analysis of this process in this organism.

CHAPTER II

ECTODERM- AND ENDOMESODERM- SPECIFIC GATA TRANSCRIPTION FACTORS IN THE MARINE POLYCHAETE *PLATYNEREIS DUMERILII*

Contributors

This work was published in Volume 9 of the journal Evolution and Development in January 2008. I would like to acknowledge to work of the coauthors of this material. I performed the experiments, and conducted sequence and evolutionary analyses. With the help of Stephan Q. Schneider, I designed the study and analyzed the data. Bruce Bowerman and Stephan Q. Schneider conceived and supervised the study. With the help of Bruce Bowerman and Stephan Q Schneider, I drafted the manuscript. All the authors read and approved the final manuscript.

Introduction

The ability of embryonic tissue to form separate cell layers is crucial to the evolution of metazoan animals. Cnidarians, including jellyfish, anemones, and corals, possess two distinct germ layers, separated into a bifunctional inner layer, the endomesoderm, and the outer ectoderm. The vast majority of animal phyla possess three distinct germ layers, the outer ectoderm, inner endoderm, and the intermediate mesoderm (Martindale 2005). It is still unclear to what degree these layers are to be considered homologous across phyla, especially due to great variation in the modes of gastrulation (Technau and Scholz 2003). Many gene-families have been implicated as playing similar

roles in germ layer patterning across animal phyla (Martindale 2005). However, due to the high rates of gene duplication, gene loss, and sequence divergence within invertebrate model organism genomes, it has been difficult to accurately reconstruct the evolutionary history of these gene families.

The zinc finger transcription factors called GATA factors are one class of developmental regulatory genes that appear to share conserved roles in germ layer patterning during early embryogenesis in many animal phyla (Rehorn et al. 1996; Shoichet et al. 2000; Patient 2002; Martindale et al. 2004). Among vertebrates, molecular phylogenetic analysis clearly places these factors into either a GATA1/2/3 class or a GATA4/5/6 (Lowry and Atchley 2000; Molkentin 2000; Patient 2002). These two classes have been implicated in distinct developmental processes, such as erythroid and neural specification by GATA1/2/3 factors, and cardiac mesoderm and endoderm specification by GATA4/5/6 family members (Patient 2002). In both mice and zebrafish, GATA1/2/3 genes are expressed in early ectodermal lineages (Xu et al. 1997; Nardelli et al. 1999; Thisse et al. 2001; Tsarovina et al. 2004), while GATA4/5/6 genes are expressed in endomesodermal lineages (Molkentin 2000; Patient 2002; Loose and Patient 2004; Watt et al. 2004; Holtzinger and Evans 2005).

Similar largely germ layer-specific requirements have been identified for invertebrate GATA factors. Six of the eleven *C. elegans* GATA factors are required for the specification of mesodermal and endodermal lineages (Shoichet et al. 2000; Maduro et al. 2001; Maduro and Rothman 2002; Coroian et al. 2006). The remaining five influence ecto- and neuroectodermal lineages, including the epidermis and motor neurons (Page et al. 1997; Smith et al. 2005; Woollard 2005). The *Drosophila* GATA factor

pannier appears related to the vertebrate GATA4/5/6 class, as it is expressed in dorsal mesoderm and required for cardiac development, while the *Drosophila* GATA factor *grain (GATAc)* appears to be related to the vertebrate GATA1/2/3 class, as it is required for the specification of ectodermal and neural tissues (Brown and Castelli-Gair Hombria 2000). Of the remaining *Drosophila* GATA factors, *serpent (GATAb)* and *GATAe* are involved, respectively, in the early specification (Reuter 1994) and later differentiation (Okumura et al. 2005) of endoderm tissues, while the fifth (*GATAd*) does not appear to be expressed in embryos (Murakami et al. 2005; Okumura et al. 2005).

Despite the documented requirements for both vertebrate and invertebrate GATA factors in the development of either endomesodermal or ectodermal derivatives, their phylogenetic relationships remain unclear, and some exceptions to the germ layer specific requirements have been documented. For example, in addition to being required for mesodermal fates early in development, *pannier* also regulates dorsal closure of the epidermis, suggesting that some GATA factors may have acquired additional roles in other germ layer derivatives. Similarly, *Drosophila serpent* mutants are defective in both endoderm and blood cell fate specification, leading to suggestions that *serpent* is related to both the vertebrate GATA1/2/3 and 4/5/6 classes (Reuter 1994; Rehorn et al. 1996).

To explore the evolutionary conservation of GATA factors and their roles in germ layer development, we have isolated GATA transcription factors genes and have characterized their transcriptional expression patterns in a marine annelid, the polychaete *Platynereis dumerilii* (Fischer and Dorresteijn 2004). *Platynereis* appears to have retained primitive morphological and genetic features, and thus may be well suited for reconstructing the ancestral composition of early developmental programs (Tessmar-

Raible and Arendt 2003; Fischer and Dorresteijn 2004). The *Platynereis* body plan appears similar to fossils of the early Cambrian (>530mya), and has been referred to as a 'living fossil' (Tessmar-Raible and Arendt 2003). The life history, gene complement and composition, and genomic intron-exon structure of *Platynereis* suggest it has diverged significantly less from the last common protostome-deuterostome ancestor, the so-called *Urbilaterian*, than many other extant bilaterians (Tessmar-Raible and Arendt 2003; Fischer and Dorresteijn 2004; Raible et al. 2005).

Here, we report the isolation and sequence of two *Platynereis* GATA factor genes. Our phylogenetic analysis indicates that the two *Platynereis* GATA factors are highly conserved orthologs of the two vertebrate classes of GATA factors. The *Platynereis* gene sequences make it possible to place highly divergent invertebrate GATA factors within either of the two vertebrate classes, and to identify GATA1/2/3 and GATA4/5/6 specific sequence motifs. Consistent with their phylogenetic relationships, *PdGATA123* is expressed during embryogenesis in ectodermal lineages and *PdGATA456* in endomesodermal lineages. Finally, we found only one copy of a GATA factor gene present within the databases for two sequenced cnidarian genomes. We conclude that the duplication of GATA factors, with a subsequent divergence of ectodermal and endomesodermal roles, occurred after the split of cnidarian and bilaterian animals and before the protostome/deuterostome divergence.

Materials and Methods

Gene Isolation

Degenerate primers were designed to the most highly conserved regions of bilaterian GATA factor homologs (Martindale et al. 2004). Gene fragments were

obtained by PCR amplification from an embryonic cDNA library (kindly provided by D. Arendt and A. Dorresteijn). PCR fragments were cloned in the pGEM-T Easy Vector (Promega) and sequenced at the University of Oregon Molecular Biology Sequencing Facility. Sequences from authenticated clones were used to design nested sets of non-degenerate primers with annealing temperatures between 68-70°C for RACE (rapid amplification of cDNA ends), using external primers designed to either the cDNA library vector, or in amplification of the cDNA pools. Additional sequence was obtained for *PdGATA123* by sequencing clones obtained from a bacteriophage lambda-packaged *Platynereis* cDNA library. The radiolabeled probe was generated using random primed labeling of a PCR fragment from authenticated RACE clones, and a REDIprime random prime labeling kit (Amersham). Phage plaques were grown to high density in NZY-top agar with a bacterial lawn, transferred to a Hybond-N+ nylon membrane, and hybridized with probe under stringent conditions according to protocols from the manufacturer (Amersham).

Sequence Analysis

GATA sequences were collected using the NCBI PubMed database (see Supplemental Figure 1 for complete sequence list and accession numbers), and aligned using the Satchmo multiple sequence alignment program (see Supplemental Figure 2) (Edgar and Sjolander 2003). Alignments were also compared with T-coffee, clustalx, and muscle, with little improvement of the overall local alignments of distantly related orthologs. Phylogenetic reconstruction was generated using Bayesian Inference, using the software Mr. Bayes (Huelsenbeck and Ronquist 2001). An initial analysis (see Supplemental Figure 3) of 500,000 generate rations was conducted on the conserved

domain of all of the GATA factors listed in the sequence list (Supplemental Figure 1). The final tree was generated using two converged runs, each of 1,000,000 generations, and a consensus tree was generated between the two runs, analyzed using a 1000k burn-in.

Motif analysis

To improve the detection of class specific GATA motifs, we used the Blockmaker suite of programs (D'Ambrosio et al. 2003). Two sequence alignments of vertebrate representatives from either the GATA1/2/3 or GATA4/5/6 classes were generated using Satchmo (Edgar and Sjolander 2003). Blocks were carved from these alignments using Multiple Sequence Processor (D'Ambrosio et al. 2003), and then used to screen for class-specific motifs from a variety of vertebrates (*H. sapiens, M. musculus, G.gallus, X.laevis, D. rerio, R. eglanteria),* a tunicate (*C. intestinalis),* an echinoderm *(S. purpuratus),* a fruitfly *(D. melanogaster),* a nematode *(C. elegans*), and a cnidarian *(N. vectensis*). Each GATA factor sequence was screened against both classes of motifs using the Do-it-yourself (DIY) feature of Blockmaker. These automatically assigned blocks were used as 'primers' to generate larger manually improved alignments, including invertebrate GATA sequences, containing these blocks (see Supplemental Data 4-6). A motif was scored as retained if it shared at least a 20% pairwise identity with another example of the motif, as determined by the complete pairwise distance matrix for the individual motif sequence alignment.

Trace Archive Search

Gene sequences for the Human GATA-3 (AY888592) and GATA-4 (NM_002052) were used as bait sequences for blast analysis using the NCBI

discontinuous MegaBLAST server. Results for these appear to be mostly overlapping, due to the high conservation of the dual-zinc finger in all bilaterian GATA homologs. To ensure an exhaustive list of clones, these lists were checked using divergent GATA sequences, including the *Drosophila serpent* gene, and the *C.elegans med-1* gene. Copy number of GATA factors in each genome was determined by grouping redundant traces. Orthology and frame determinations were assessed using a reverse BlastX analysis, and used to aid translation of sequence fragments in the MacVector program.

Polychaete culture, embryo fixation, and whole mount in situ analysis

Fertilizations were induced via the introduction of free-swimming mature adults of opposite sexes to a single dish of filtered sea water (FSW) (Fischer and Dorresteijn 2004). Jelly coats were removed via a rinse in acidified seawater, followed by several rinses in FSW. To penetrate the vitelline membrane of embryos younger than 24 hours, embryos were treated for 8-15 minutes in a solution of TCMFSW (50 mM Tris, 495 mM NaCl, 9.6 mM KCl, 27.6 mM Na_2SO_4, 2.3 mM $NaHCO_3$, 6.4 mM EDTA, pH 8.0) (personal communication, R. Kostyuchenko). Embryos were fixed in 4% paraformaldehyde/2X PBS-Tween for 1 to 4 hours at room temperature. Probe generation, hybridization, and detection were performed as previously described (Tessmar-Raible et al. 2005)

Results

Molecular cloning and phylogenetic analysis of two polychaete GATA factor gene sequences: PdGATA123 and PdGATA456

To investigate the evolutionary origins of the two vertebrate GATA factor classes, we used degenerate PCR to identify GATA factor gene sequences in cDNA pools and

libraries made from *Platynereis dumerilii* embryonic and larval extracts (see Materials and Methods). We isolated cDNA fragments that encode highly conserved dual zinc finger DNA binding domains and correspond to two different but highly conserved *Platynereis* GATA factors genes. Additional sequences were obtained using RACE PCR with gene specific and cDNA vector primers, and through screening bacteriophage cDNA libraries with radiolabeled probes (see Materials and Methods). We have named these two genes *PdGATA123* and *PdGATA456,* due to their clear relationships to the two vertebrate GATA factor classes GATA1/2/3 and GATA4/5/6, respectively. For *PdGATA456*, we have isolated sequences that encode 365 amino acids and include a 5'UTR and a probable start codon but lack a stop codon and a 3'UTR. For *PdGATA123*, we have identified sequences encoding 525 amino acids, including putative translational start and stop codons, and a complete 3' UTR.

For a phylogenetic analysis of *PdGATA123* and *PdGATA456*, we assembled a dataset of GATA factor gene sequences, supplementing the two polychaete sequences with representative GATA factor gene sequences from online databases (see Materials and Methods). GATA factor gene sequences can be identified based upon the strong conservation of a roughly 134 amino acid region with two GATA zinc-finger domains. These zinc-finger domains have been shown in some cases to bind to a canonical (A/G)GATA(A/G) DNA sequence found in the promoter regions of target genes regulated by GATA factors (Lowry and Atchley 2000). An alignment of this conserved domain and molecular phylogenetic analysis was used to reconstruct the origin of the duplication events that have led to the known set of contemporary GATA factors (Figure II.1, Supplemental Figures 1, 2).

By restricting our data set to include only echinoderm, vertebrate, and polychaete GATA factors, we consistently resolved 2 GATA factor classes with high support (Figure II.1, data not shown). Addition to the dataset of more divergent GATA factors from tunicates, fruit flies, and nematodes occasionally resulted in the placement of these long-branch family members as outgroups (Supplemental Figure 3). However, these errors are likely due to Long Branch Attraction (LBA), as more rapidly evolving sequences will lose more shared ancestral characters than slowly evolving sequences (Philippe et al. 2005). Using a Bayesian analysis that incorporates species-specific evolutionary rates to reduce LBA errors (Figure II.1), we were consistently able to resolve homologs for each of the two GATA factor classes, even within the faster evolving bilaterian taxa, including the fruit fly, nematode, and tunicate. For the GATA1/2/3 class, this includes *Ciona GATAb, Drosophila grain/GATAc,* and *C. elegans elt-1*. Furthermore, we have resolved *Drosophila pannier, Ciona GATAb,* and *C. elegans end-1* as GATA4/5/6 homologs. However, many of the remaining fly and nematode GATA factors fail to resolve in either GATA class and remain as long branches outside of either tree (Supplemental Figure 3). As these additional GATA factor genes cannot be found outside insects or nematodes, respectively, they are likely to be independent and highly divergent duplications within these lineages.

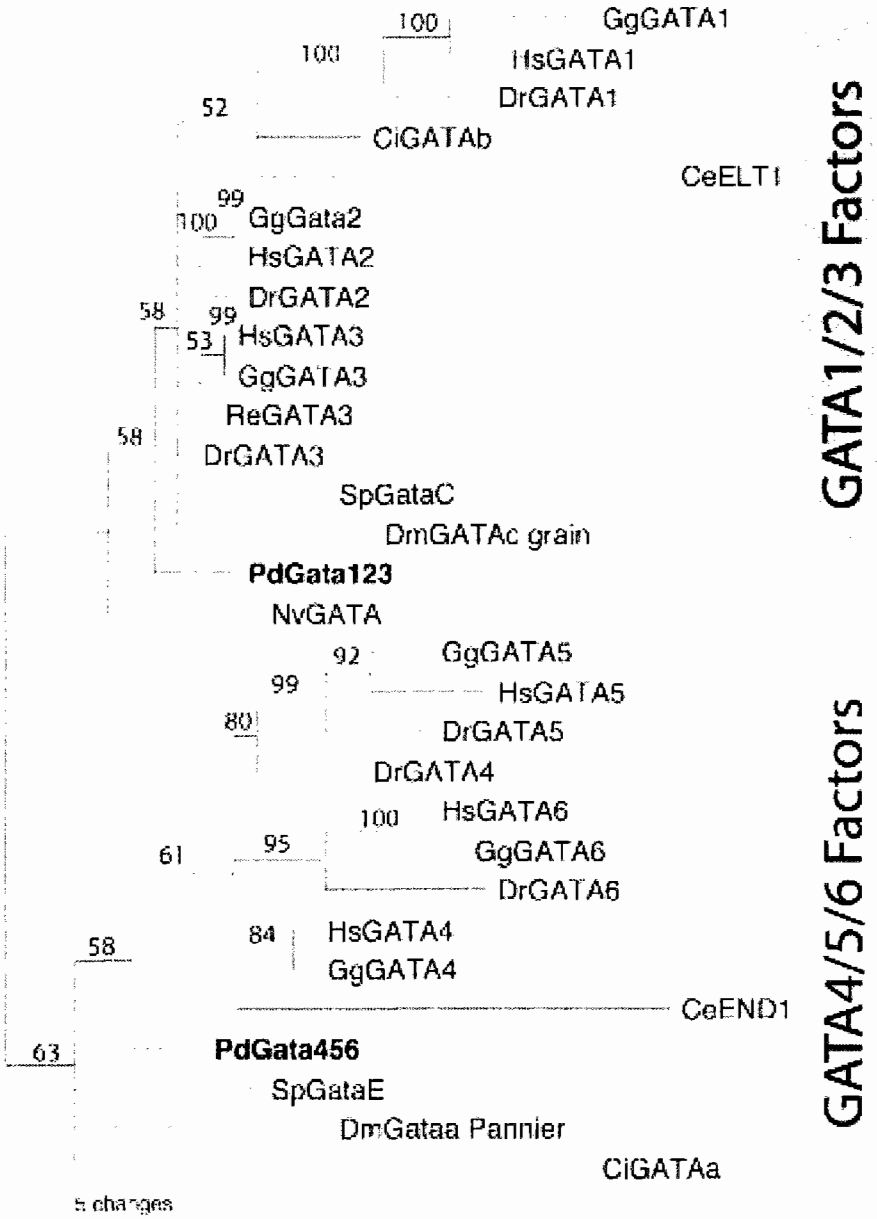

Figure II.1 Phylogeny of eumetazoan GATA transcription factors, as inferred by Bayesian analysis.

Representative GATA factors from bilaterian animals form two classes, a GATA1/2/3 class and a GATA4/5/6 class, and include the two Platynereis factors (bold). The sole cnidarian GATA, NvGATA (not shaded), appears to be a chimera of these two bilaterian classes, but shows more overall similarity to the GATA1/2/3 class. Pd, Polychaete; Hs, Human; Gg, Chicken; Dr, Zebrafish; Re, Skate; Sp, Sea Urchin (echinoderm); Dm, Fruitfly; Ci, Tunicate (urochordate); Ce, Nematode; Nv, Sea Anemone (cnidarian). See supplementary Figure S1 for accession numbers. Node labels above a branch show Bayesian posterior probabilities.

Non-bilaterian GATA factors

We also searched for GATA factor gene sequences outside of bilaterian animals, using BLAST searches of available genomic trace archives (see Materials and Methods). In addition to the sole GATA factor gene reported for the sea anemone, *Nematostella vectensis* (Martindale et al. 2004), we identified a similar ortholog from a distantly related cnidarian, *Hydra magnipapillata*. Both of these cnidarian GATA factor genes are represented at high coverage (5-8x), but we were unable to identify any additional GATA factor genes in either genome. We could not distinguish between the *Nematostella* GATA factor being part of the GATA1/2/3 class, or an outgroup to both bilaterian GATA factor classes (see Materials and Methods, Figure II.1). In an even more distantly related animal, the sponge *Reniera,* we were unable to identify any GATA factor gene sequences.

Identification of Class Specific GATA motifs

Having identified two bilaterian classes of GATA factors based on sequence alignments within the zinc finger domains, we next searched for any conservation of class-specific sequence features outside of the zinc fingers (see Materials and Methods, Figure II.2, Table II.1, Supplemental Figures 4-6). Seven motifs were identified for the GATA1/2/3 class, and four for the GATA4/5/6 class, among metazoans. The motifs range from 14-44 AA in length, and are described below with a subset of signature amino acids. The order and positioning of these motifs within open reading frames are conserved. With one exception, all class-specific motifs are conserved in the *Platynereis* GATA factors: only the first N-terminal motif in the GATA1/2/3 class is missing. Both the *PdGATA123:N3* (GxQVCRPH) and the *PdGATA123:N4* (LFxFPPTPPK) motifs

share over a fifty-five percent identity between *Platynereis* and vertebrates GATA1/2/3 orthologs. In the GATA456 class, both the *Platynereis GATA456:N2* (SPVYVPT) and the *GATA456:N3* (HPPxxxFSxxSPP) motifs were sixty percent identical to vertebrate counterparts.

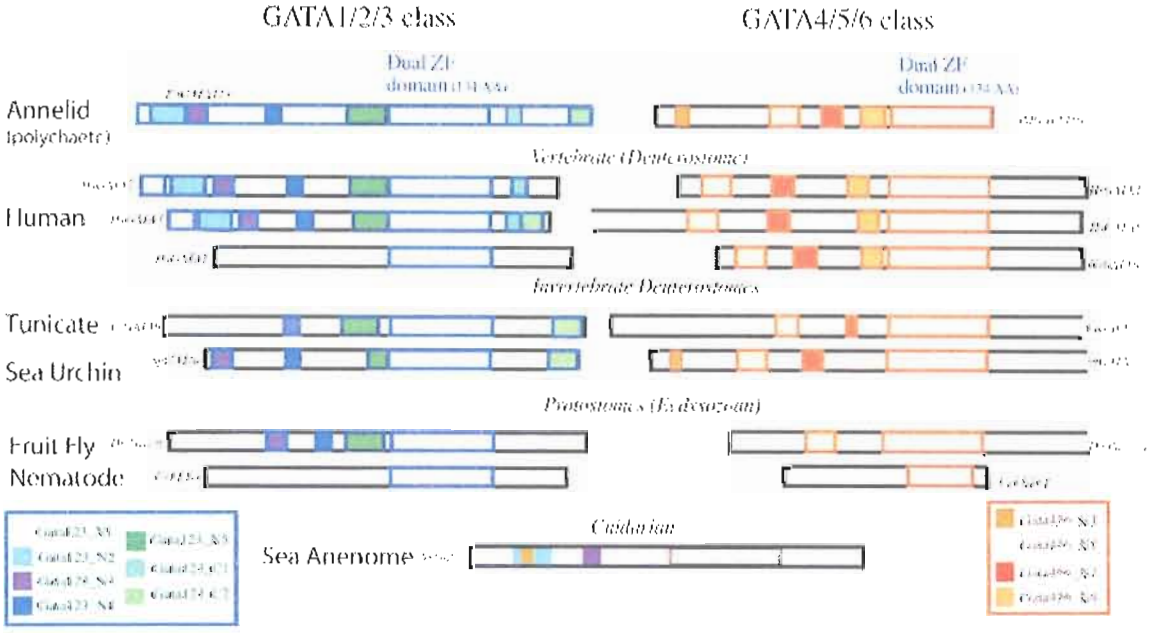

Figure II.2 Gene structures of GATA1/2/3 (left) and GATA4/5/6 factors (right). Sequences were aligned using the conserved dual-zinc finger domains (grey boxes). Class-specific conserved sequence motifs (see supplementary Figs. S4–S6) are indicated by colored boxes; transparent boxes refer to weakly conserved motifs (<20% conserved with any other example of that motif). Blue outlines indicate identification within the GATA1/2/3 class; red outlines indicate identification within the GATA4/5/6 class. For GATA1/2/3, 5 N-terminal and 2 C-terminal motifs were identified, as shown. For GATA4/5/6, we identified 4 N-terminal motifs. The 5' end of the HsGATA5, and the 3' ends of SpGATAc, CiGATAa, and DmPannier are not shown.

GATA factors from animals that possessed highly divergent dual-zinc finger domains also showed a significant divergence in these external motifs, using a cutoff of twenty percent identity with at least one other example of that motif. All of the percent

Table II.1 Conservation of GATA motifs: Percent Identity between representative GATA motifs

Motif	*Platynereis* to Vertebrates	*Platynereis* to *Drosophila*	*Platynereis* to *Nematostella*	*Nematostella to* Vertebrates
Gata123:N1	-	-	-	22-36%
Gata123:N2	20-39%	23%	27%	27-50%
Gata123:N3	28-56%	60%	52%	30-41%
Gata123:N4	48-64%	64%	64%	40-64%
Gata123:N5	31-44%	12%	11%	8-12%
Gata123:C1	29-47%	17%	11%	5-17%
Gata123:C2	17-33%	13%	-	16-25%
Gata456:N1	-	-	53%	-
Gata456:N2	35-60%	24%	21%	12-18%
Gata456:N3	43-60%	-	21%	9-23%
Gata456:N4	13-27%	-	-	-

identities of the motifs described below refer to the next-closest example of that motif in another phyla . No nematode GATA factor sequences possess motifs at this cutoff; however, we included only the most conserved members in our diagram for comparison (Figure II.2). We were able to find motifs that met the twenty percent identity cutoff in only two *Drosophila* GATA factors, with *DmGrain/GATAc* possessing three GATA1/2/3 motifs [GATA123:N2 (23%), N3 (60%), N4 (64%)] and *DmPannier* possessing one GATA4/5/6 motif [GATA4/5/6:N2 (24%)]. Despite being taxonomically closer, individual *Platynereis* GATA motifs were more conserved with vertebrates than with fruitflies (Table II.1). Of the ten motifs common between *Platynereis* and vertebrates, only the *PdGATA123:N3* shares more identity with a fruitfly motif [*DmGrain:N3 (60%)*], than with the next closest vertebrate sequence [*DrGATA2:N3 (56%)*]. Similarly, the urochordate *Ciona* possesses only three GATA1/2/3 motifs [GATA123:N4 (44%), N5

(21%), C2 (33%)] and two GATA4/5/6 motifs [GATA456:N2 (21%), N3 (21%)]. In comparison, the echinoderm *S. purpuratus* possesses four GATA1/2/3 motifs [GATA123:N3 (56%), N4 (70%), N5 (21%), C2 (25%)] and three GATA4/5/6 motifs [GATA456:N1 (60%), N2 (46%), N3 (46%)]. Thus the *S. purpuratus* GATA factors appear to be more conserved than those in *Ciona.*

We also identified class specific motifs from both classes within the sole *Nematostella* GATA factor. Five of the motifs from the GATA1/2/3 class, and two from the GATA4/5/6 class, can be identified in *NvGATA* at greater than twenty percent identity to any bilaterian motif. Additionally, this allowed for the identification of overlap between two sets of motifs between each class (GATA1/2/3:N2 and GATA4/5/6:N1; GATA1/2/3:N4 and GATA4/5/6:N3), suggesting that these may have arisen from a common motif (Figure II.2, Table II.1, see Supplemental Figure 6).

Ectoderm and mesoderm development in Platynereis

To determine if the *Platynereis* GATA factor genes might function within distinct germ layers (see Introduction), we next examined their transcriptional expression during embryogenesis. Before describing these results, we first briefly summarize the embryonic development of polychaete mesodermal and ectodermal germ layers.

The early *Platynereis* embryo undergoes a stereotyped sequence of asymmetric cell divisions called holoblastic spiral cleavages. An initial sequence of asymmetric cleavages unequally distribute clear and yolky cytoplasm, with the dorsal cell D being the largest and having the most clear cytoplasm (Dorresteijn 1990). The clear cytoplasm has been suggested to contain undetermined cellular determinants that specify mesodermal fates (Dorresteijn and Eich 1991; Schneider et al. 1992; Dorresteijn 2005). Starting at the

third cleavage, after the birth of the D cell, spindle axes in subsequent divisions form orthogonal to the axes of the previous divisions, resulting in a spiral arrangement of daughter cells (Wilson 1892). The bilateral cleavages of two conspicuously large progeny of the D macromere, the mesentoblast 4d that produces trunk mesoderm, and a descendant of the first somatoblast 2d, $2d^{112}$, that produces trunk ectoderm, mark the transition to the bilateral symmetry of the larva.

Most if not all of the polychaete trunk mesoderm descends from a single cell, the mesentoblast 4d, produced at the sixth round of cell cleavages (~64 cell stage). This endomesodermal cell lineage has been previously described, both through fixed stage analysis (Wilson 1892), and more recently after injection of a fluorescent lineage tracer into the 4d blastomere (Ackermann et al. 2005). At the 7^{th} round of mitotic cell division, the spiral cleavage pattern is broken with a transverse division of 4d. This division leads to two bilaterally symmetrical daughter cells, $4d^1$ and $4d^2$, that are located dorsally, above the vegetal pole. These cells are then internalized during epiboly and blastopore closure. These two daughters subsequently undergo rapid proliferation to form two paired mesodermal bands that extend anteriorly, as seen in the 24-hour post-fertilization (hpf) trochophore larva (Figure II.3, middle panel)(Wilson 1892). During metamorphosis, these mesodermal bands form segmental blocks, which split to form the coelomic tissues. The coelomic epithelium later differentiates to form body wall muscle and other mesodermal tissues.

The first somatoblast 2d forms as a relatively large blastomere descended from the D macromere at the fifth round of cleavage, and contributes to the majority of the trunk ectoderm. After the bilateral cleavage of the 4d mesentoblast, the 2d descendant

$2d^{112}$ divides with bilateral symmetry at the dorsal midline to form progenitors of trunk ectoderm. The $2d^{112}$ daughter cells undergo rapid proliferation to form two bilaterally symmetrical ectodermal fields that move ventrally around the posterior pole by the 24 hpf trochophore larval stage (Figure II.3, middle panel). These fields converge at the ventral midline during metamorphosis to form the trunk ectoderm including the ventral neural plate. By 72 hpf (Figure II.3, left panel), the ventral plate including the neuroectoderm is conspicuous as a multi-layered epithelial sheet on the ventral surface of the trunk that converges and extends along the anterior-posterior axis.

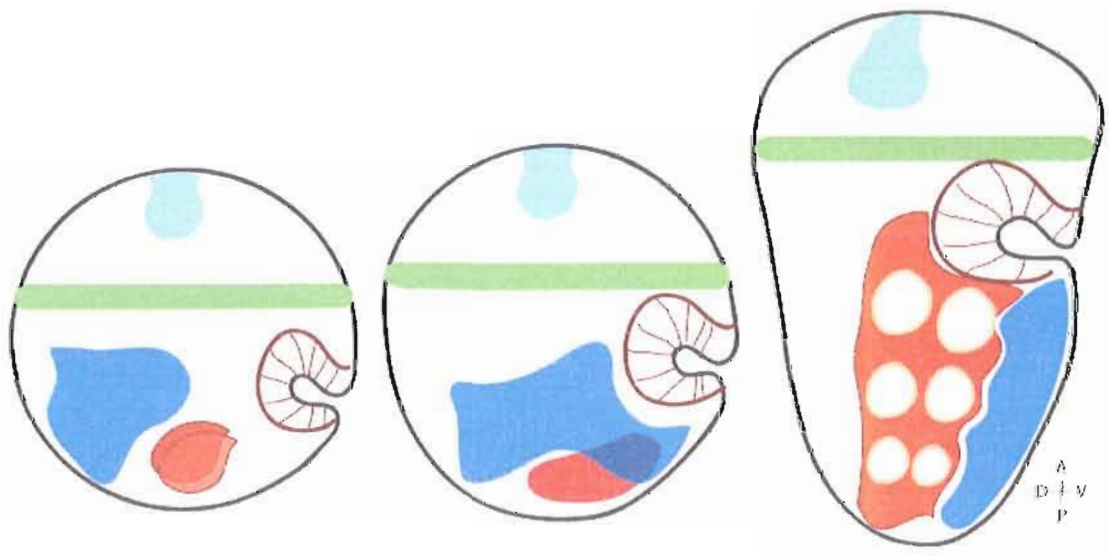

Figure II.3 Development of *Platynereis* from an early trochophore larva (left) to a three-segmented juvenile worm (right) adapted from Wilson (1892).
These lateral views depict only one side of the bilaterally symmetrical embryo/larva. Cells that give rise to the ventral neural plate ectoderm (dark blue) are first located on the posterior/vegetal dorsal surface. These cells then move ventrally around the posterior pole and converge at the ventral midline to form the definitive ventral neural plate. The mesoderm precursors (light red) are internalized at gastrulation. They proliferate and extend anteriorly to form mesodermal tissues, including the trunk musculature that surrounds the setal sacs (tan). These lateral views also depict the development of the anterior neuroectoderm (light blue), stomodaeum (dark red), and prototroch (green). The juvenile worm is labeled with anterior (A), posterior (P), dorsal (D), and ventral (V) axes.

Expression of *Platynereis* GATA factor mRNAs

We used in situ hybridization to detect *PdGATA123* mRNA expression in fixed *Platynereis* embryos (see Materials and Methods). We detected embryonic expression of *PdGATA123* within ectodermal lineages of the ventral and anterior neural plates (Figure II.4). In the early trochophore larva (Figure II.4a-c), expression of *PdGATA123* was restricted to two fields of dorsal ectoderm (Figure II.4b, c), which we presume to be the descendants of the trunk ectodermal stem cell, $2d^{112}$, with the strongest staining localized to two paired fields of dorsal-vegetal ectoderm. In staged embryos from 24-48 hpf, we detected *PdGATA123* expression in paired ectodermal fields that wrap around the posterior pole. In late larvae (48 hpf, Figure II.4g-i), the two fields of *PdGATA123* were fused at the ventral midline (Figure II.4h), with the strongest expression in the deeper layers of the ventral plate (Figure II.4g), the prospective neuroectoderm. This strong expression coincides with the formation of several neural structures, including the paired ventral nerve cords, circumesophogael ganglia, and segmental ganglia (Figure II.4g,h) (Wilson 1892; Rouse and Pleijl 2001; Ackermann et al. 2005; Dorresteijn 2005) . Similar ectodermal expression of *PdGATA123* was detected throughout development to the juvenile worm stage (Figure II.4j). In addition to the ventral ectoderm, faint but distinct staining of *PdGATA123* expression was detected in paired patches of anterior ectodermal cells just dorsal and below the surface of the animal pole (Figure II.4a,b). This expression appeared to be continuous with expression seen in later embryos, although these cells appeared to occupy less space and were shifted dorsally by later stages (Figure II.4a, d).

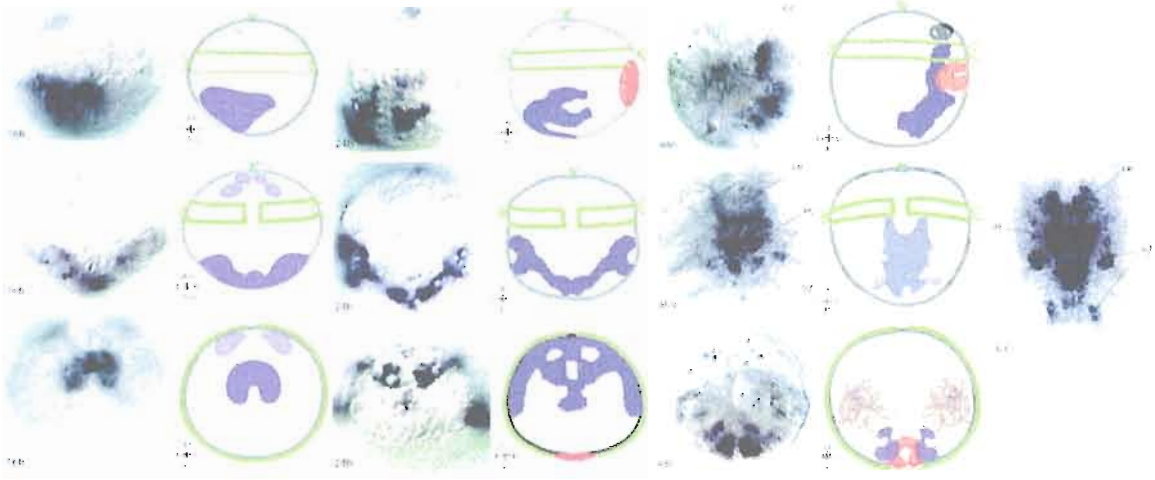

Figure II.4 Expression of *PdGATA123* mRNA shown in lateral (top), dorsal (middle), and posterior (bottom) optical sections at 16, 24, and 48 h (H), and at 4 days (D) postfertilization (pf).
Schematic views depict expression (shown in grey/purple shading) at each stage: the prototroch (green), apical tuft (blue), stomodaeum (red), setal sacs (brown), and frontal bodies (grey) are included as landmarks. The paired ventral nerve cords (vc), circumesophogael ganglia (cg), and segmental ganglia (sg) are also indicated. Expression was detected in developing ectoderm and neuroectoderm progenitors by 16 h postfertilization and continues up to 4 days postfertilization (ventral view; far right panel). Orientations are indicated: An, animal; Vg, vegetal; A, anterior; P, posterior; D, dorsal; V, ventral; L, left; R, right.

We next used in situ hybridization to examine the expression of *Platynereis PdGATA456* transcripts. *PdGATA456* mRNA localization appeared to be restricted to mesodermal lineages and derivatives (Figure II.5). In support of this conclusion, *PdGATA456* expression closely resembled that of the conserved mesoderm factor *PdTwist* and, in later stages, a muscle actin (Stephan Schneider and Khoa Tran, unpublished results). Additionally, early expression of *PdGATA456* appeared to be purely endomesodermal, which in spiralians is all produced by the 4d micromere. The 4d progeny have been previously traced via intracellular injection, (Ackerman et. al. 2005), and by comparison appears identical to the progeny containing *PdGATA456* expression.

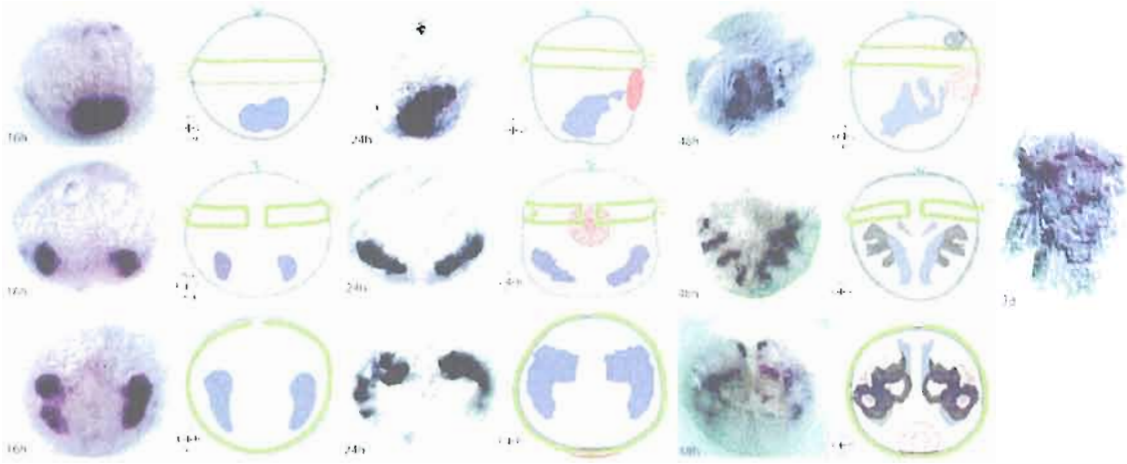

Figure II.5 Expression of *PdGATA456* mRNA shown in lateral (top), dorsal (middle), and posterior (bottom) optical sections at 16, 24, and 48 h postfertilization. Schematic views depict expression (shown in grey/purple shading) at each stage. The prototroch (green), apical tuft (blue), stomodaeum (red), setal sacs (brown), and frontal bodies (grey) are included as landmarks. Expression can be detected in developing mesoderm up to 3 days postfertilization (ventral view, far right panel). Orientations are indicated: An, animal; Vg, vegetal; A, anterior; P, posterior; D, dorsal; V, ventral; L, left; R, right.

Interestingly, later stages (Figure II.5h, j, data not shown) also showed stomodeal *PdGATA456* expression, consistent with previous descriptions of ectomesoderm that is likely derived from progenitors of the third micromere quartet (Ackermann et al. 2005).

At 16 hpf, we detected *PdGATA456* expression in paired lateral bands (Figure II.5b, c) underneath the surface ectoderm near the vegetal pole of the embryo (Figure II.5a), most likely in descendants of the mesodermal progenitor cells $4d^1$ and $4d^2$. By 24 hpf (Figure II.5d, e), these bands were expanded anteriorly nearly to the prototroch, which serves as a dividing line between trunk and anterior structures. By 48 hpf, the expression pattern became segmental (Figure II.5h) and was associated with the involuted setal sacs that generate the bristles (Figure II.5h, i), or chaetae, used for locomotion. *PdGata456* expression was highest near these sacs but was detectable throughout the

entire mesodermal band. By the 3-day juvenile stage (Figure II.5j), only a few small patches of expression were detectable in what appears to be trunk musculature associated with the setal sacs.

To summarize, the two *Platynereis* GATA factor genes exhibit germ layer restrictions in transcriptional expression that also have been observed for their vertebrate orthologs. *PdGATA123* was expressed in ectodermal and *PdGATA456* in endomesodermal derivatives, although we cannot rule out some overlap or lack of germ layer specificity in later stages.

Discussion

Evolutionary conservation of two bilaterian GATA factors

Our analysis of two *Platynereis* GATA factor gene sequences and their transcriptional expression provides the first definitive phylogenetic support for protostome invertebrates possessing orthologs of both vertebrate GATA factor classes. The *Platynereis* GATA factors are transcribed at detectable levels with similar germ layer restrictions that have been documented for their respective vertebrate orthologs. *PdGATA123* was expressed in ectodermal lineages, while *PdGATA456* was restricted to mesoderm.

Vertebrate orthologs of *PdGATA456* have prominent roles in endoderm specification, but we did not observe any endoderm expression for either of the two GATA factor genes we isolated from *Platynereis*. We speculate that either *PdGATA456* lost an ancestral endodermal role, or we missed an additional *GATA456* gene with endodermal functions in our degenerate PCR screens. Nonetheless, we found a clear phylogenetic relationship of the *Platynereis* GATA factor genes to the two vertebrate

GATA factor classes, and a similar restricted germ layer expression pattern. These results suggest that protostomes and vertebrates likely descended from a common 'urbilaterian' ancestor that possessed two GATA factor orthologs, with separate ectodermal and endomesodermal expression domains (Figure II.6).

The evolution of vertebrate GATA factors

So how and when did vertebrates acquire six GATA factors? We have found that all examined invertebrate deuterostomes possess only the ancestral complement of two GATA factors. Our genomic analyses (see Supplemental Figure 1) detected the presence of only a single GATA1/2/3 and a single GATA4/5/6 ortholog for basal chordates, in a cephalochordate (*Branchiostoma floridae)* and in a urochordate tunicate (*Ciona intestinalis).* We also detected only a single member of each class in two echinoderms, a sea star (*Asterina miniata*) and a sea urchin (*Strongylocentrotus purpuratus).* The expansion of six mammalian GATA factors presumably corresponds to the two rounds of whole-genome duplication that occurred during the evolution of higher vertebrates (Dehal and Boore 2005). We predict that six factors were likely the result of retention of three duplicates from each GATA class, and the loss of a single duplicate after the second round. These results can be tested with upcoming genome projects in a marine lamprey (*Petromyzon marinus*), and the tiny skate (*Raja erinacea*), which are thought to have diverged prior the first and second rounds of vertebrate genome duplication, respectively.

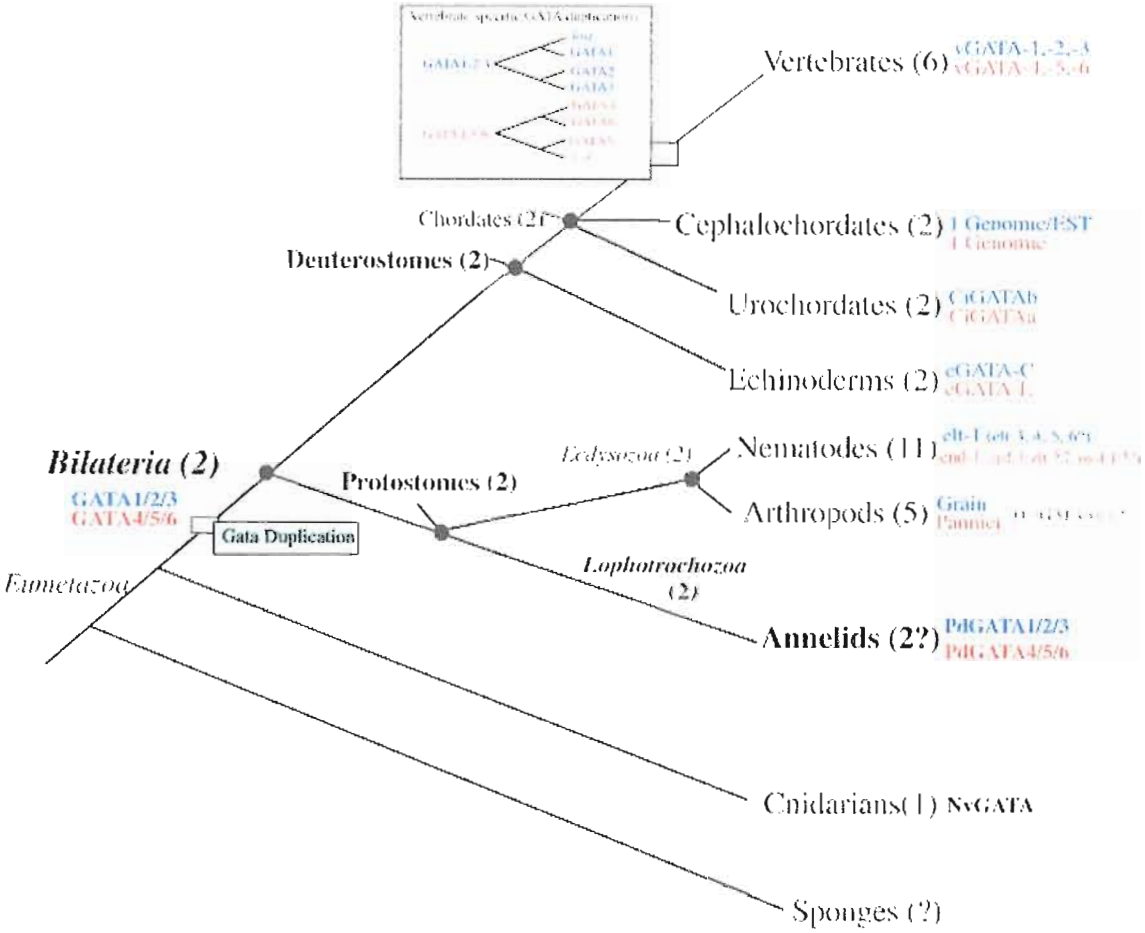

Figure II.6 Evolutionary scenario of GATA factors in metazoan animals.
A duplication of a single ancestral GATA factor occurred after the divergence of the cnidarian and bilaterian lineages, leading to two distinct GATA factors in the last-common bilaterian ancestor, the Urbilaterian. This duplication coincides with the emergence of a mesodermal germ layer. These two urbilaterian GATA factors likely possessed restricted germ layer roles in the ectoderm for GATA1/2/3 (blue), and the mesendoderm for GATA4/5/6 (red). Vertebrates appear to have expanded these to a total of six in the last-common vertebrate ancestor, possibly during the 2R genome duplications (Dehal and Boore 2005). A gene tree for vertebrate GATA factors is depicted in the grey box, showing the likely locations of two lost GATA factors.

Basal deuterostome GATA factors also appear to share some of the class-specific roles described for vertebrates (see Introduction). In the urochordate *Ciona intestinalis* two GATA factors have been described (Yamada et al. 2003). These two genes, called *CiGATAb* and *CiGATAa*, appear to be orthologs of the vertebrate GATA1/2/3 and

GATA4/5/6 classes, respectively. *CiGATAa* expression was initially reported within the developing central nervous system and in blood cells (D'Ambrosio et al. 2003). In contrast, a recent large scale in situ screen of *Ciona* transcription factors described the expression of *CiGATAa* within endoderm, and *CiGATAb* within the developing nervous system and mesenchyme (Imai et al. 2004). These two *Ciona* GATA factor genes are highly diverged, and our phylogenetic analysis only weakly supports their placement within the two vertebrate classes. In addition to tunicates, orthologs of both bilaterian GATA factor classes have been identified in echinoderms (Davidson et al. 2002),. In both a starfish and a sea urchin, the GATA456 homologs are expressed in the developing endomesoderm, and appears to be major activators of the endomesodermal gene network (Hinman and Davidson 2003). Transcripts of the sea urchin GATA1/2/3 homolog have been detected in coelomocytes (Pancer et al. 1999), which may relate to roles in erythroid specification for the vertebrate GATA1/2/3 factors, but early germ layer expression has not been reported. We are currently testing our hypothesis of germ layer restricted expression of GATA factors within *Branchiostoma*, which appears to be a less diverged basal deuterostome.

Phylogenetic relationships of protostome invertebrate GATA factors

Our analysis also suggests that the 11 *C. elegans* and *5 Drosophila* GATA factors arose independently from the same 2 ancestral bilaterian GATA factors that were independently duplicated within the vertebrate lineages. Consistent with our phylogenetic conclusions, the 11 *C. elegans* GATA factors have been shown to be involved in either endomesendoderm or epidermal lineages. These requirements correspond fully to their phylogenetic classification as either GATA1/2/3 ectoderm or

GATA4/5/6 endomesoderm class genes. The GATA4/5/6 class single zinc-fingered GATA factors *med-1* and *-2* are required for both mesoderm and endoderm fate (Maduro et al. 2001), while the GATA4/5/6 factor genes *end-1, end-3, elt-2,* and *elt-7* are required for endoderm development (Coroian et al. 2006). The most conserved nematode GATA factor gene is the GATA1/2/3 class gene *elt-1*, which is required for epidermal and ventral motor neuron fates. Closely related duplicates of *elt-1*, called *elt -3, -4, -5,* and *–6,* act later in epidermal and ventral motor neuroblast fate specification (Smith et al. 2005).

We found that the two most conserved *Drosophila* GATA factor genes are *grain/GATAc,* a GATA1/2/3 ortholog, and *pannier,* a GATA4/5/6 ortholog. Consistent with this classification, *grain* is required for motor neuron fate specification (Garces and Thor 2006), and *pannier* is required for specification of cardiac mesoderm (Klinedinst and Bodmer 2003). However, *Pannier* also is required for dorsal closure of the epidermis (Heitzler et al. 1996), and appears to be more generally used to subdivide the body along the dorsal-ventral axis (Herranz and Morata 2001).

Some outstanding questions still remain in the placement of the remaining *Drosophila* GATA factors. *Serpent* is one of the most divergent GATA factors in our analysis. Although it is one of the largest GATA factors (780 AAs), *serpent*'s relationship to other GATA factors is based only upon its possession of half of the dual zinc finger domain. As discussed in the Introduction, *serpent* mutants display defects in both endomesoderm and hematopoietic fate specification (Rehorn et al. 1996; Patient 2002). Based upon our division using ectoderm and endomesoderm germ layer roles, we would define *serpent* as a GATA4/5/6-type factor. However we cannot deny the

similarity between the blood-cell specification role of GATA1/2/3 genes in vertebrates and the role of *serpent* in the specification in the blood-like crystal cells. Although, it is difficult to distinguish between these two possibilities, the clear homology of *Grain* and *Pannier* has ruled out *serpent* as being an ortholog to the entire GATA family, as previously suggested (Rehorn et al. 1996). It is possible that GATA factors may have independently become involved in erythroid specification, but a more likely view is that both fly and vertebrate GATA factors have conserved additional ancestral features distinct from earlier roles in germ layer patterning, such as in specification of erythroid lineages. A resolution to the homology of *serpent* will likely require the examination of additional protostomes, including additional arthropod GATA gene family complements, which may resolve its evolutionary origin. We are currently examining the later expression of *Platynereis* GATA factors to address potential roles for GATA factors in polychaete erythroid specification.

An early GATA factor gene duplication

In our searches of two distantly related cnidarian genomes, *Nematostella vectensis* and *Hydra magnipapillata,* we found only a sole GATA factor ortholog in each genome, one of which had been previously reported from *Nematostella* (Martindale et al. 2004). While we cannot rule out the loss of a second GATA factor gene during the evolution of cnidarians, we believe it is more likely that the single-copy cnidarian GATA factors are extant representatives of the sole ancestral ortholog to both classes of bilaterian GATA factors. We speculate that a subsequent gene duplication after the divergence of the cnidarian-bilaterian ancestor led to two bilaterian GATA factors. In support of this conclusion, *Nematostella* is similar to *Platynereis* in that it appears to possess a highly

conserved genome relative to other invertebrates (Darling et al. 2005; Miller et al. 2005). Indeed, *NvGATA* is a slowly-evolving gene based on our molecular phylogenetic analysis. Furthermore, our motif analysis suggests that *NvGATA*, unlike any other bilaterian GATA factors, is chimaeric with class-specific motifs from both classes of bilaterian GATA factors. Moreover, the expression of *NvGATA* is suggestive of dual endodermal and ectodermal roles. *NvGATA* mRNA expression was detected mostly within a population of ingressing endodermal cells that later surround the coelenteron, a cavity described as coelom-like and thus possibly related to the mesodermally derived coelom in bilaterian animals (Martindale et al. 2004). In addition, a distinct ectodermal field of *NvGATA* expression was detected around the base of the tentacles. The most parsimonious interpretation is that one single GATA factor gene present in the common ancestor to cnidarians and bilaterian animals underwent a duplication and subsequent segregation of function to either ectoderm or endomesoderm within the urbilaterian lineage. Subsequent exceptions to these ancestral germ layer restrictions appear to have occurred in some animal lineages, but a general pattern of restriction is still observed for most GATA factors.

Gene duplication and the origins of germ layers

Our phylogenetic model for GATA factor evolution follows the duplication-degeneration-complementation (DDC) model (Force et al. 1999; Force et al. 2005), which suggests that gene duplicates are preserved by unequal segregation of previous subfunctions. The subfunctions we address here are ectodermal or endomesodermal domains of germ layer specification. Presumably such a divergence in expression and function could have resulted from the divergence of preexisting regulatory sequences and

a subsequent divergence of preexisting coding features into either of the GATA duplicates. Our discovery of conserved sequence motifs for two distinct bilaterian classes of GATA factors demonstrates the unequal conservation or degeneration of coding-level features. Our finding of conserved distinct bilaterian expression domains for these two GATA classes suggests changes in the cis-regulatory regions of both factors. Our model predicts that cis-regulatory elements responsible for the ecto- and endomesodermal expression of the sole ancestral GATA factor have been complementarily conserved or lost to regulate bilaterian GATA1/2/3 and GATA4/5/6 expression, respectively. This scenario can now be tested in *Nematostella* and *Platynereis* embryos by dissecting their promoter cis-regulatory sequences.

It is interesting that the evolutionary appearance of two separate GATA factor classes coincides with the evolution of a distinct third germ layer, the mesoderm, possibly originating from a subdivision of the endoderm. While the emergence of endomesoderm likely required many discrete steps, it may prove useful to focus on ancient gene duplications that occurred prior to the emergence of bilaterian animals, as appears to be true for the GATA factor gene expansion. The evolutionary appearance of the specialized GATA4/5/6 gene, critical for the formation of endomesoderm derivatives in many bilaterians, appears to have occurred after the split from the cnidarian lineage, but prior to the appearance of the original "urbilaterian" ancestor. Moreover, according to our analysis GATA4/5/6 genes appear to be the more divergent members of the two bilaterian GATA factors, supporting a newly emerging role for this gene on the base of Bilateria. We propose that the duplication of the progenitor to the two GATA factor gene classes played an important if partial evolutionary role in the subdivision of germ layers,

including the emergence of mesoderm and thus the evolution of diverse bilaterian animal body plans. Presumably additional ancient gene duplications of conserved regulatory factors will be identified that also correlate with the evolution of germ layer subdivisions, further advancing our understanding of this key step in animal evolution.

CHAPTER III

THE EVOLUTION OF PROTOSTOME GATA FACTORS: MOLECULAR PHYLOGENETICS, SYNTENY, AND INTRON/EXON STRUCTURE REVEAL ORTHOLOGOUS RELATIONSHIPS

Contributors

This work was published in Volume 8 of the journal BMC Evolutionary Biology in April 2008. I would like to acknowledge to work of the coauthors of this material. I performed sequence and evolutionary analyses. With the help of Stephan Q. Schneider, I designed the study and analyzed the data. Bruce Bowerman and Stephan Q. Schneider conceived and supervised the study. With the help of Bruce Bowerman and Stephan Q Schneider, I drafted the manuscript. All the authors read and approved the final manuscript.

Background

GATA transcription factors perform conserved and essential roles during animal development, including germ layer specification, hematopoiesis, and cardiogenesis (Patient 2002). Nevertheless, homologs in the GATA gene family have undergone significant divergence in both sequence and gene number in different animal phyla, making it difficult to resolve orthologous relationships of individual family members (Gillis et al. 2007, Lowry and Atchley 2000). For example, the number of GATA

paralogs--homologs within an individual genome--varies substantially between protostomes and deuterostomes. Most vertebrate genomes possess six GATA paralogs, whereas the fruitfly *Drosophila melanogaster* has only five, and the nematode/roundworm *Caenorhabditis elegans* eleven. Reconstructing the evolution and the ancestral developmental roles of these genes requires a framework of orthologous relationships among GATA homologs.

Previous studies have identified two classes of GATA homologs within deuterostomes (Gillis et al. 2007, Lowry and Atchley 2000). Basal invertebrate deuterostomes, including echinoderms, urochordates, and cephalochordates, possess only single GATA123 and GATA456 orthologs. Most vertebrates possess three paralogs from each class, likely from two whole genome duplication events that occurred during the evolution of jawed vertebrates. Within the three vertebrate GATA123 paralogs, the vertebrate GATA-2 and -3 genes are more closely related to each other than to the GATA-1 gene. Likewise, the vertebrate GATA-4 and -6 genes are both more closely related to each other than to the GATA-5 gene (Lowry and Atchley 2000). Thus two genome duplications, together with the losses of one GATA-1 like paralog and one GATA-5 like paralog, can account for the number of genes in each vertebrate GATA class.

While the evolution of GATA factors within the deeper branches of the deuterostome phylogeny is well understood, it has been more difficult to reconstruct the evolution of protostome GATA factors. We recently published data suggesting that the last common protostome/deuterostome ancestor had at least two GATA factors with distinct roles in early germ layer development: an endomesodermal GATA456 gene and

an ectodermal GATA123 gene (Gillis et al. 2007). In this analysis, at least one representative was identified from each class in multiple protostome genomes, and the germ layer specific expression for each class was documented in a basal lophotrochozoan, the polychaete annelid *Platynereis dumerilii*. However, orthologous relationships for the more degenerate *C. elegans* and *Drosophila* GATA transcription factors remained unclear.

Here, we report an analysis of the complete complement of GATA factors from several newly available protostome genomes. We have identified GATA factors from nine diverse protostomes by directly searching databases from recently conducted whole genome sequencing efforts. We have conducted phylogenetic analyses using predicted protein sequences, conserved chromosomal gene order, and conserved intron/exon boundaries to better understand the evolution of protostome GATA factors. Our results provide evidence for protostome-specific expansions of GATA456 paralogs and enable us to infer the evolutionary relationships of even the most divergent *Drosophila* GATA factors.

Results

The complement of GATA transcription factors from newly sequenced protostome genomes

To further investigate the evolution of GATA transcription factors within protostomes, we obtained GATA gene sequences from nine newly sequenced and phylogenetically informative protostome genomes (see Materials and Methods). These include five arthropods [*Ixodes scapularis* (tick), *Daphnia pulex* (water flea), *Tribolium*

castaneum (beetle), *Apis mellifera* (bee), and *Anopheles gambiae* (mosquito)], one nematode (*Caenorhabditis briggsae*), and three lophotrochozoan [*Lottia gigantea* (limpet), *Capitella capitata* (polychaete), *Schmidtea mediterranea* (flatworm)] genomes. For almost all of these collected GATA transcription factor genes we identified and assembled the complete dual-zinc finger domain for further analyses. We believe that these retrieved GATA genes represent the complete GATA gene complement within each analyzed genome (see Materials and Methods, Additional File 1).

Each ortholog was initially assigned to either the ancestral bilaterian GATA123 or GATA456 class (see Introduction), based upon reciprocal best hit BLAST analysis (see Methods). With the exclusion of the nematode *Caenorhabditis briggsae* (discussed below), each of the additional protostome genomes appeared to possess a single GATA123 ortholog and three GATA456 paralogs. In the four insect genomes, as well as in the single annelid (*Capitella capitata)*, a fourth highly divergent GATA456 paralog was detected. Thus, our initial genome wide search indicated the existence of one single copy GATA123 gene and multiple copies of GATA456 genes within these additional protostome species.

Molecular phylogenetic analysis defines multiple distinct GATA456 clades within arthropods.

Resolving the phylogenetic relationships of GATA transcription factors present in *Drosophila* has been difficult due to their highly divergent gene sequences. Previous work had shown that *Drosophila* possesses an obvious GATA123 ortholog, *grain*, as well as an unambiguous GATA456 ortholog, *pannier* (Lowry and Atchley 2000). However, the placement of the three remaining *Drosophila* GATAs (*serpent*, *GATAd*, and *GATAe)*

has been uncertain due to extensive sequence divergence within the generally well-conserved dual zinc finger domain. The *Drosophila* GATA genes *serpent* and *GATAe* have been proposed to be derived GATA456 orthologs due to their roles in endoderm and/or mesoderm development (Gillis et al. 2007). However, like some vertebrate GATA123 genes, *serpent* also has roles in blood development suggesting that *serpent* may be orthologous to all vertebrate GATA genes (Reuter 1994, Rehorn et al. 1996).

We now can resolve the uncertain *Drosophila* GATA factor relationships by including sequences from additional arthropod genomes. The phylogenetic tree in Figure III.1 represents the combined results of maximum likelihood (ML), Bayesian inference (MB), and distance based (NJ) analyses (see Materials and Methods). This tree was rooted with the sole GATA transcription factor in the cnidarian *Nematostella vectensis* (NvGATA), which appears to be equally related to GATA123 and GATA456 genes (Gillis et al. 2007). This analysis confirms the existence of two separate branches of bilaterian (protostome and deuterostome) GATA factors, a clade of GATA123 genes and a clade of GATA456 genes. It replicates previous results (Gillis et al. 2007), but now with substantially increased support due to the presence of additional and more conserved GATA sequences. With the addition of multiple arthropod sequences, the diverged *Drosophila serpent, GATAe,* and *GATAd* now unambiguously group within the larger GATA456 clade. The GATA123 versus GATA456 groupings are well supported by both ML and MB analyses, though not by NJ analysis, which shows lower bootstrap support presume this is a possible short-branch attraction artifact, due to the relatively low

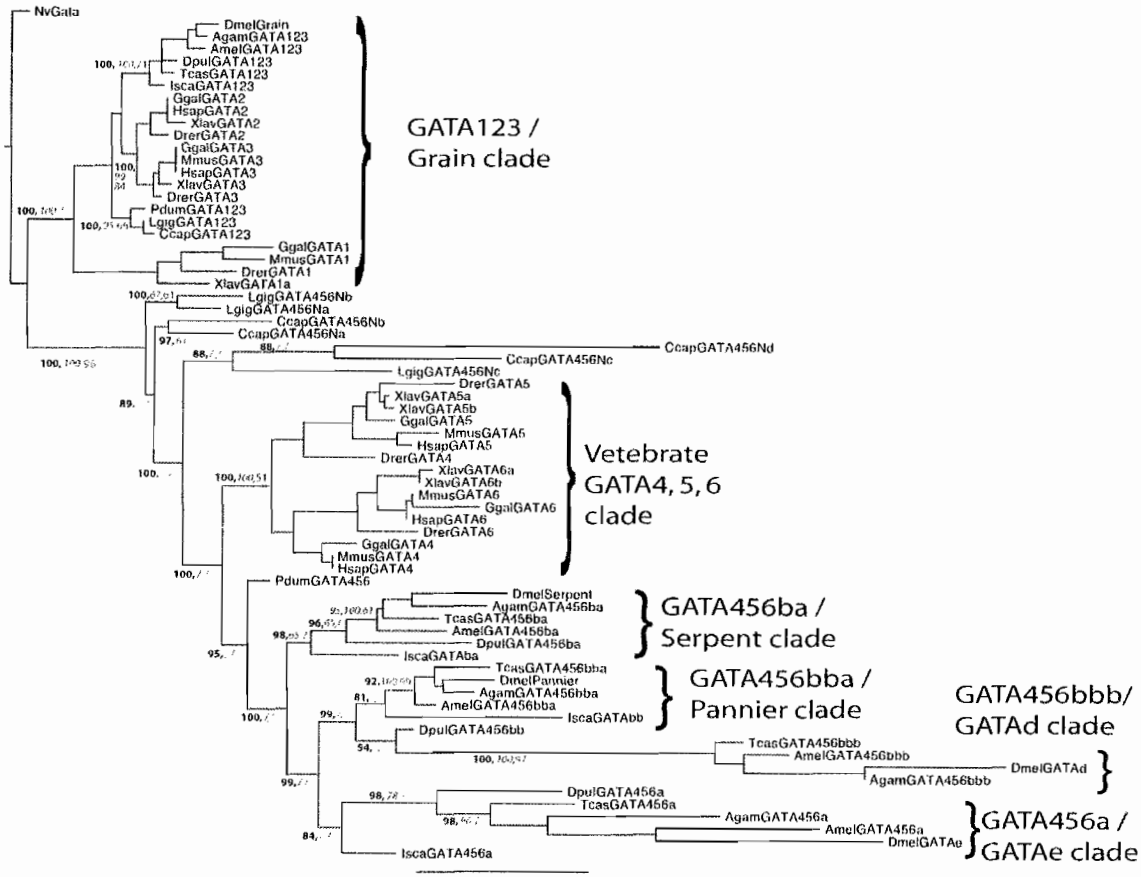

Figure III.1 Phylogenetic analysis of GATA Transcription Factors.
Gene phylogeny based on a combined molecular phylogenetic analysis using Maximum Likelihood (ML), Bayesian Analysis (MB), and Neighbour Joining (NJ) methods. Genes are prefixed by a short abbreviation for the organism (1 letter for genus, three for species). Topology and branch lengths were generated using the PhyML-aLRT program, and branch support for key nodes is shown in bold (ML, SH-like aLRT statistic), italics (MB, posterior probabilities), and plain text (NJ, bootstrap percentiles). Inferred arthropod and vertebrate clades are marked by brackets to the right.

due to the occasional grouping of the NvGATA within the GATA123 clade. We sequence divergence for the GATA123 and NvGATA genes. However, consistent with our results from the reciprocal best hit BLAST analysis (see above), all of the arthropod genomes encode sole GATA123/Grain like orthologs but multiple GATA456 like paralogs.

Our more inclusive phylogenetic analysis also reveals orthologous groups among the arthropod GATA456 paralogs. The four *Drosophila* GATA456-like transcription factors are co-orthologs of the vertebrate GATA456 family, and appear to have independently expanded early in arthropod evolution. The *Drosophila* GATA456 paralogs are members of four distinct clades, each containing a single paralog from every analyzed insect genome. Furthermore, three of these insect GATA456 paralog groups appear to be conserved within the arthropods, as one paralog for each of the three clades are found in the crustacean *Daphnia,* and in the chelicerate/tick, *Ixodes.* We have named the three common arthropod clades as GATA456a/GATAe, GATA456ba/serpent, and GATA456bb, using a nomenclature discussed below (see Material and Methods). GATA456bb appears to have undergone an insect-specific duplication (see Discussion), resulting in the GATA456bba/pannier and GATA456bbb/GATAd clades in insects, and hence explaining the fourth insect GATA456 paralog.

While our results support four distinct clades within the arthropod GATA456 subfamily, the deeper evolutionary relationships among these clades remain uncertain. NJ and MB analyses represent the internal relationships between the four GATA456 paralogous clades as an unresolved basal polytomy, despite the resolution of four external GATA456 clades. However, the ML analysis suggests additional interclade relationships, as shown in the tree in Figure 1. All of the arthropod GATA456 paralogs appear to form a distinct and well supported clade. The insect specific GATA456bbb clade groups closely with GATA456bba clade, and therefore appear to be duplicates from a common GATA456bb gene. This topology also suggests that theGATA456bb factors

appear more closely related to GATA456a orthologs than to the GATA456ba ortholog group, but we have found no additional evidence to support this relationship.

An arthropod GATA456 paralog cluster: Synteny reveals orthologous relationships

To better understand the evolutionary relationships of arthropod GATA456 paralogs, we examined the syntenic relationships among the different arthropod GATA factors and discovered a conserved linkage of GATA456 paralogs. As shown in Figure

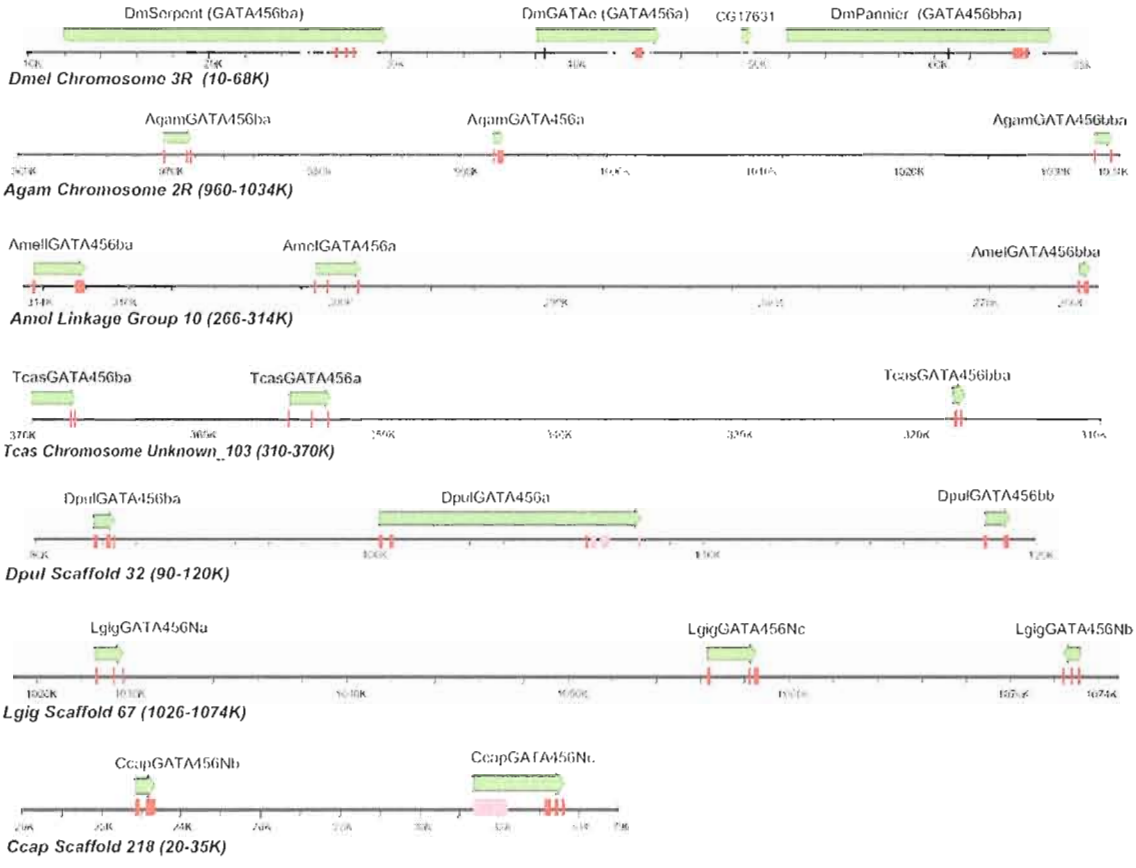

Figure III.2 Synteny of GATA456 paralogs in arthropods and lophotrochozoans. Linked GATA456 orthologs are shown within genomic regions ranging from 30kB to 75kb in length. Green arrows indicate the transcribed GATA456 gene regions and transcriptional direction. The full extent of the transcribed region is known only for well-characterized genomes like *Drosophila*. Predicted coding sequences (CDS) are shown in pink; the conserved dual-zinc finger domain in red.

III.2, we found that three of the *Drosophila* genes--*serpent, GATAe*, and *pannier*--are clustered within a 47 KB region on the *Drosophila* 2R chromosome. We have identified a similar cluster of three tightly linked GATA456 paralogs in other arthropod genomes, including additional insects and a crustacean (see Figure III.2). Gene orientation within the cluster is fully conserved, and the gene order follows the predicted orthology suggested by the clades in our molecular phylogenetic analysis (see above). As this cluster is conserved in all analyzed insects an=d a crustacean, we infer that this cluster arose at least as early as the pancrustacean ancestor, some 420 million years ago. No additional syntenic relationships were found when comparing the nearest upstream and downstream genes from each of the five assembled arthropod genomes, suggesting that gene order within this GATA456 paralog cluster is more conserved than within surrounding chromosomal regions.

The conserved linkage suggests an origin from two tandem duplication events of a single ancestral GATA456 transcription factor. This cluster includes an unambiguous GATA456 ortholog (*pannier*) (Gillis et al. 2007, Lowry and Atchley 2000), further supporting our phylogenetic inference that these three homologs are all GATA456 paralogs. Furthermore, the weak homology of the three identified tick GATA transcription factors to each of these paralogs suggests that this three-gene cluster may have existed in the last common arthropod ancestor. However, our initial attempts at local contig assembly (see Materials and Methods) have failed to find linkage for the tick GATA genes. GATA gene linkage in the tick should soon be resolved, pending assembly of the whole tick genome.

A unique intron/exon structure for each of the three arthropod GATA456 clades.

The ~135 AA dual-zinc finger domain that defines the broader GATA transcription factor gene family, including both GATA123 and GATA456 homologs (Lowry and Atchley 2000), is encoded by three exons with similar intron/exon boundaries that are found in the sole cnidarian GATA gene, and in all deuterostome GATA genes we have examined (data not shown). We infer that the ancestral GATA transcription factor gene contained these three exons (see Figure III.3, Additional File 1). An N-terminal exon (ZF1, ~50 AA) encodes the first zinc-finger, and a middle exon (ZF2, ~54AA) encodes the second DNA binding zinc-finger domain. The C-terminal exon (3'CD) encodes a conserved stretch of ~30 AA, after which conservation sharply drops.

Although the exon structure of this dual zinc finger domain is well conserved in most GATA homologs, many GATA genes in the fruitfly *Drosophila melanogaster* appear to lack the first zinc finger. In *Drosophila*, the first *GATAe* zinc finger is highly divergent, and *GATAd* lacks any sign of the first zinc finger. *Serpent* was initially thought to be a single zinc-fingered protein (Rehorn et al. 1996)--two of the three splice variants for *serpent* lack the first zinc finger—but a more complete analysis identified the *serpent* isoform B with a complete conserved domain (Waltzer et al. 2002). A lack of a first zinc finger in many GATA factors within *Drosophila*, as well as *C. elegans* (see below), could suggest that these GATA factors evolved from an ancestral sequence encoded by a single zinc finger (Lowry and Atchely 2000, Rehorn et. al. 1996). However, our examination of additional arthropod GATAs indicates that

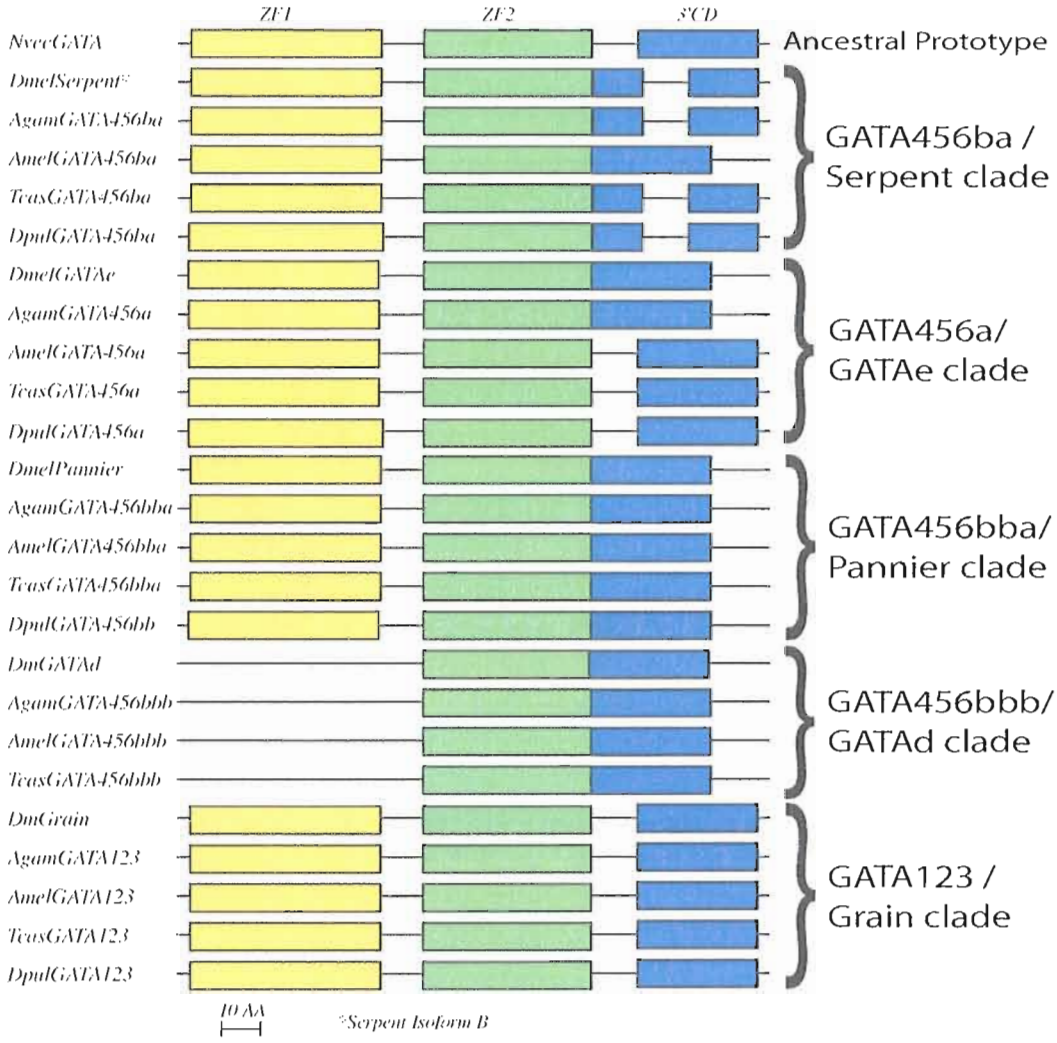

Figure III.3 Intron/exon Structure of arthropod GATA conserved domains. Schematics of the exon structure for the conserved dual-zinc finger domain of the arthropods GATA transcription factors. The ancestral prototype for this conserved domain consists of three exons with well-conserved boundaries, represented here by the sole *Nematostella vectensis NvecGATA*. The first exon (ZF1 in yellow) consists of the first DNA binding zinc-finger, whereas the second (ZF2 in green) contains the second DNA binding zinc-finger domain. The third exon (3'CD in blue) contains the GATA N-terminal activation domain described for some species.

independent losses of the first zinc-finger have occurred. Both zinc-finger exons appear intact for the three ancestral arthropod GATA456 genes in the additional arthropod species. Only the insect specific GATA456bbb/*GATAd* orthologs consistently lack the

first zinc finger exon; however, our analysis shows these are relatively recent duplicates, and hence must have arisen from dual-zinc fingered GATA456 factors. Therefore, we conclude that the absence of the first exon and zinc finger in some *Drosophila* GATA genes is a derived rather than an ancestral trait.

Our analysis of GATA gene structure also suggests a derived loss or modification of the second (2ZF) and third (3'CD) exon boundary in many of the arthropod GATA homologs (Figure III.3). 16 of the 24 identified insect and crustacean (pancrustacean) GATA homologs have undergone a modification of this exon boundary compared to the inferred ancestral sequence. The second and third exons in all of the GATA456bbb/GATAd and GATA456bba/pannier orthologs are fused. The GATA456a/GATAe genes have retained the ancestral state for the intron/exon domains, though they have also fused their second and third exons in dipteran insects. The GATA456ba/serpent genes (except the honey bee *Apis)* also contain a second intron in the conserved domain, yet the boundaries of this second intron do not appear conserved.

The unusual intron/exon structure of the GATA456ba/serpent genes suggests they may have resulted from an initial fusion of the second and third exon, and the subsequent introduction of a new intron. We have identified four examples of insect and *Daphnia* GATA456ba/serpent homologs in which the first 13 AA from the third exon (3'CD) are now encoded within the middle exon (2ZF). The high degree of sequence conservation implies a transfer between the original coding exons, as opposed to a loss of the beginning of the third exon and gain of surrounding genomic sequence at the end of the middle exon. This presumably rare sequence of events likely occurred only once, and

was then preserved within GATA456ba/serpent orthologs. The one exception, *AmelGATA456ba*, likely represents an additional intron loss.

To summarize, the clades of arthropod GATA456 homologs defined by molecular phylogeny also exhibit clade-specific intron/exon structures. This correspondence provides a third line of evidence in support of our proposed orthologous relationships for these genes, as suggested by both molecular phylogenetic analysis and conserved syntenic gene order. The intron/exon structure also allows us to generate deeper inferences regarding interclade relationships (see Discussion).

Extensive gene duplication and sequence divergence within the nematode GATA family

We also analyzed the GATA genes of two nematode species, *Caenorhabditis briggsae* and *Caenorhabditis elegans*. The GATA gene family has undergone extensive duplication in nematodes, with eleven GATA factors identified in *C. elegans,* and thirteen in *C. briggsae*. These sequences display significant sequence divergence, and only the *elt-1*/GATA123 orthologs contain complete dual-zinc finger domains. The other predicted nematode GATA factors all lack the first zinc finger, similar to some of the insect GATAs.

Although the nematode GATA factors are highly derived in sequence, they resemble the arthropod GATA complement in displaying a biased expansion of the GATA456 paralogs. In a previous analysis, we assigned the *C. elegans* GATA factors to one of the two classes based upon their reported germ layer-specific function or expression. Four orthologs have roles in ectoderm (epidermis and nervous system) specification, while seven *(C. elegans)* or nine *(C. briggsae)* have roles in endomesoderm (intestine and muscle) specification (see Additional File 1). However, our phylogenetic

tree (see Additional File 2) suggests that *elt-1* is the sole GATA123 ortholog in both nematode species, and that all the remaining nematode GATAs group within the GATA456 class. The long branches of these additional GATAs, and the short regions of conserved sequence, make these inferences highly speculative. Nevertheless, our data suggests that, like the arthropods we have analyzed, both nematodes have undergone a greater expansion of GATA456 paralogs, relative to GATA123 paralogs.

To evaluate the relationships between the GATA factors in the two nematode genomes, we have conducted additional phylogenetic analyses using the complete gene sequences from these two *Caenorhabditis* genomes (see Figure III.4). This analysis provides clear support for nine common clades of *C. elegans* and *C. briggsae* GATAs,

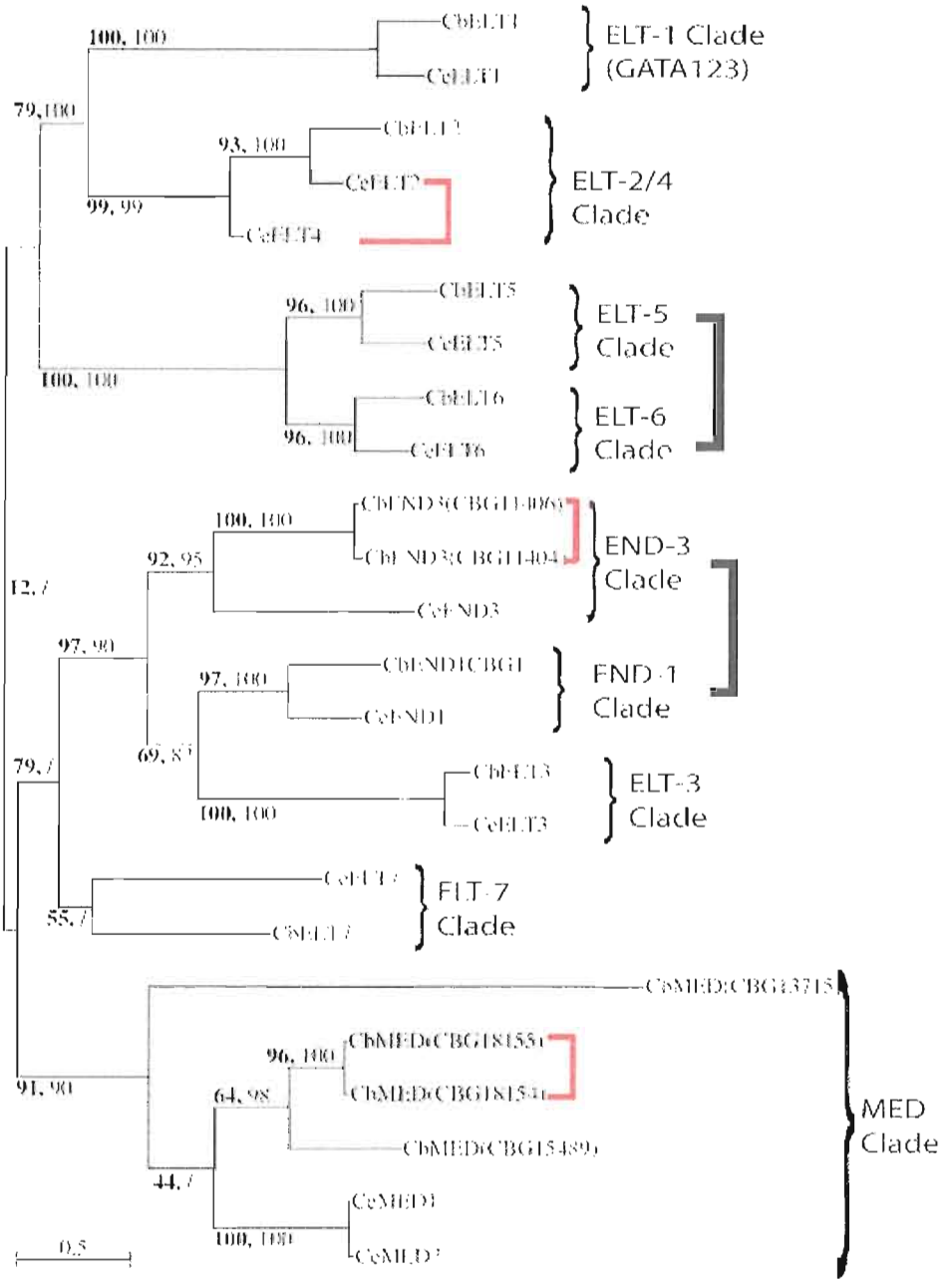

Figure III.4 Molecular phylogeny of nematode GATA factors.
Maximum Likelihood tree of *C. briggsae* (Cb) and *C. elegans* (Ce) GATA factors, showing both the PhyML-aLRT Chi2-based parametric statistic (bold) and Neighbor Joining Bootstrap percentiles (regular). Brackets to the right indicate the inferred ancestral clades. Black box brackets indicate genomic linkage found for both nematode orthologs in two clades, while red box brackets indicate linkage found only within one nematode species.

based upon the ability to define orthology between individual *C. elegans* and *C. briggsae* GATAs. This suggests that the last common ancestor of these two nematodes possessed at least nine distinct GATA genes. Furthermore, several nematode GATA factors appear to have resulted from more recent duplications within these clades, suggesting their duplication after the divergence of *C. elegans* and *C. briggsae* lineages some 80-110 million years ago (Coroian et al. 2006). These include *C. elegans elt-4*, *C. briggsae end-3*, and the *med* genes in both species.

We also observed chromosomal linkage of some nematode GATA genes. Most of the linked genes are the same ones identified as more recent duplicates within single clades that are specific to only one nematode species: the *C. elegans elt-4* and *elt-2* genes, as well as the *C. briggsae end-3* and two of the *med* genes. Some orthologs are linked in both nematode species, including the *elt-5/6* and *end-1/3* orthologs, indicating a linkage retained from an ancestral *Rhabditis* species. However, these linked genes are from closely related clades, and possess long internal branch lengths suggestive of an evolutionary origin within nematodes. We conclude that the linked nematode GATAs originated from more recent nematode and *Rhabditis* specific duplication events, and do not reflect any retention of deeper ancestral gene duplicates.

Similar GATA family gene number and linkage, but different intron/exon structures, in lophotrochozoans and arthropods.

The lophotrochozoan GATA complement is similar in copy number to the arthropod GATA complement. For example, all of the analyzed lophotrochozoans possessed three or four GATA456 homologs, and a single GATA123 homolog (see

Figure III.1). In the flatworm, *Schmidtea mediterranea,* and the limpet, *Lottia gigantea,* we have found three GATA456 paralogs in a reciprocal best hit BLAST analysis. The annelid polychaete, *Capitella capitata,* appears to possess an additional GATA456 homolog. The annelid leech *Helobdella robusta,* has 13 predicted GATA homologs; however, the number and extreme divergence of these leech GATAs relative to other lophotrochozoan genomes appear to make these uninformative in reconstructing the ancestral annelid condition.

As in arthropods, we also identified syntenic relationships of GATA456 genes in two lophotrochozoan genomes (see Figure III.2). In *Lottia*, all three of the GATA456 orthologs are contained within a 45-kilobase region, although the third gene appears to be inverted in orientation compared to the arthropod genes. In *Capitella*, two of the four GATA456 genes appear in an 11-kilobase region, though linkage to a third gene has not been found. None of the GATA genes appear linked in *Schmidtea*, consistent with the relatively degenerate nature of the flatworm GATA genes (data not shown), or perhaps a consequence of the short lengths of the currently available assembled genomic regions.

Although similar in number and linkage, it is unclear to what degree the lophotrochozoan and arthropod GATA456 duplicates can be considered orthologous. No relationships between individual arthropod and lophotrochozoan GATA456 duplicates are apparent from our molecular phylogenetic analyses.

It is also unclear to what degree these lophotrochozoan GATAs are related to one another. The *GATA456Nc* genes appear to be orthologous across *Capitella* and *Lottia* in both ML and MB analysis, and the *CcapGATA456Nd* is supported as a more recent duplicate of the *CcapGATA456Nc* gene in the ML analysis. However, the two well-

conserved GATA456 duplicates from both *Capitella* and *Lottia (Na, Nb)* group more closely within each species then across species, suggesting these could be recent lineage-specific duplicates. The sole *GATA456* identified in another polychaete, the *Platynereis dumerilii PdGATA456*, does not branch near the other lophotrochozoan GATAs, instead branching as the closest outgroup to the arthropod GATA456 clade. Finally, when we examined the intron/exon structure of all of the analyzed annelid (*Capitella capitata*), mollusk (*Lottia gigantea*), and flatworm (*Schmidtea mediterranea)* GATA genes, we found little evidence for the extensive modifications seen for the arthropod GATA456 homologs (see Additional File 1). Thus, while the orthologous relationships between arthropod GATA456 factors appear well supported, additional information will be needed to resolve the evolutionary relationships of lophotrochozoan GATA456 factors.

Discussion

We have identified the complete complement of GATA factors from nine additional protostome genomes, and we have reconstructed the evolution of protostome GATA factors using multiple approaches. Our initial estimates of orthology, from reciprocal best hit BLAST analysis, revealed an expansion of the GATA456 paralogs in protostomes, while the GATA123 genes appear to be retained as a single copy. The inclusion of additional arthropod genomes has allowed us to confidently assign the more divergent *Drosophila* GATA factors as GATA456 paralogs. Furthermore, we have demonstrated widespread linkage of many GATA456 duplicates, suggesting a mechanism of gene duplication via tandem duplication events and providing further evidence in support of duplicate-orthology. Finally, we infer from changes in intron/exon

structure the sequence of gene duplications that produced the GATA456 paralogs in arthropods.

Overview of the evolution of the protostome GATAs

We suggest two alternative scenarios for the evolution of the GATA transcription factors in protostomes (Figure 5). In the first scenario, the GATA456 family expanded very early during protostome evolution. In support of this scenario, there are similar numbers of GATA456 paralogs in lophotrochozoans (three, with a fourth in *Capitella*) and arthropods (three, with a fourth paralog in insects), even though they each appear to possess single GATA123 orthologs. Furthermore, the chromosomal linkage of GATA456 paralogs, not only in arthropods but also in the mollusk, *Lottia,* and partially in the annelid, *Capitella,* is suggestive of a deep origin.

However, our analyses more strongly support a second scenario, in which there have been independent duplications of a single GATA456 ortholog in both arthropods and lophotrochozoans. This second scenario is suggested by the lack of affinities between individual lophotrochozoan and arthropod GATA456 paralogs in molecular phylogenetic analyses. One possibility is that this cluster arose very close to the lophotrochozoan-ecdysozoan split, allowing little time for sequence divergence and retention of phylogenetic information between the GATA456 paralogs. Nevertheless, the conserved modifications of intron/exon boundaries between orthologous arthropod GATA genes suggest that intron loss and gain occurred before the duplication of certain

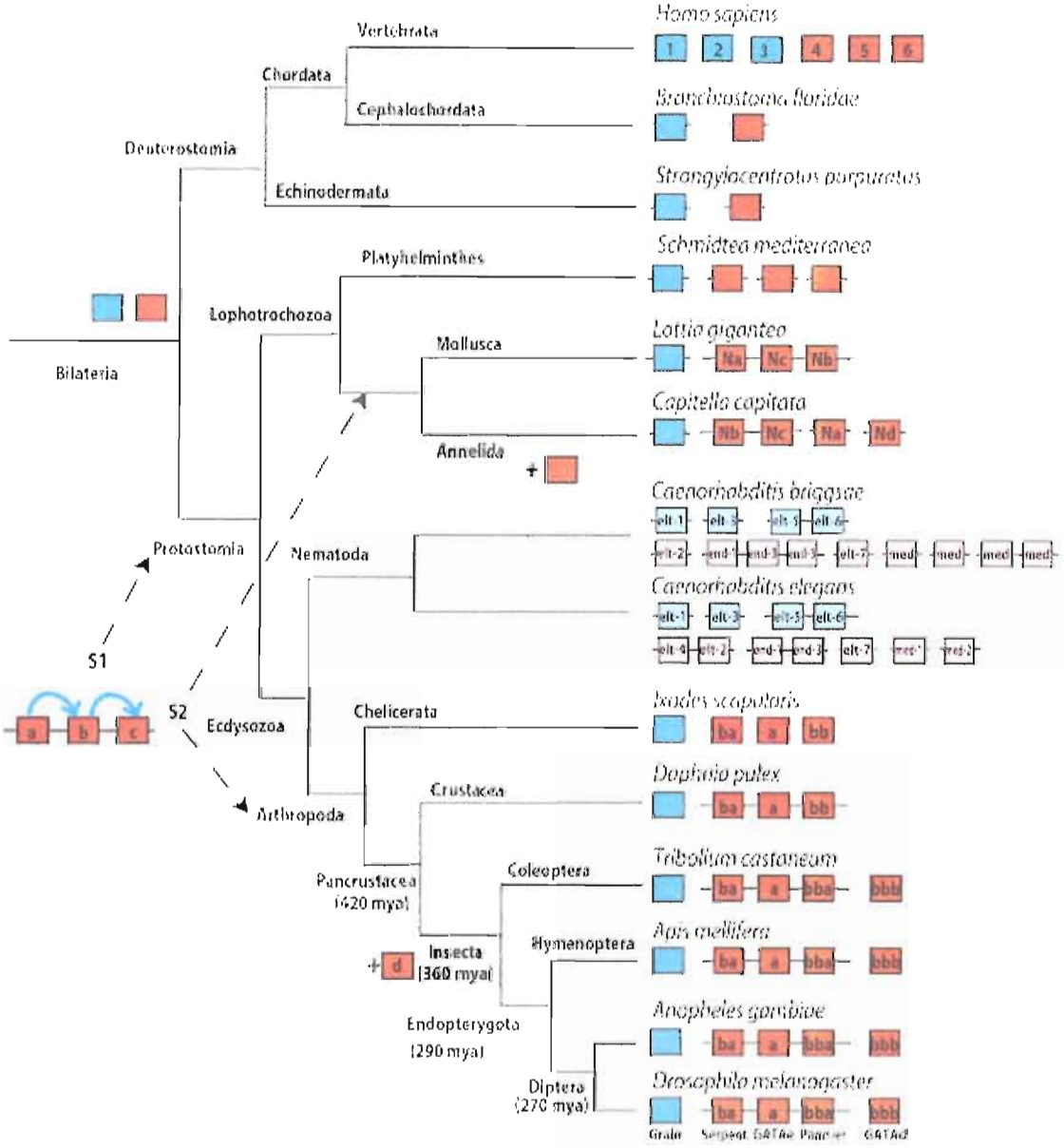

Figure III.5 Evolution of protostome GATA factors.
Alternate scenarios for the early/orthologous (S1) or late/convergent (S2) evolution of the arthropod and lophotrochozoan GATA456 gene clusters. Names at a node represent the name of the clade of organisms, and the time since the last common ancestor is given for some nodes in million years (mya) (tree topology, dates, and nomenclature from (Sequencing Consortium THG 2006, Phillipe et al. 2005, Gaunt and Miles 2002). Boxes represent individual GATA genes, with GATA123 orthologs in blue, and GATA456 orthologs in red. Light shaded nematode GATAs represent predictions based mainly upon functional conservation. Linked genes are represented with a connected line, and identified via their terminal letters (see Materials and Methods).

arthropod GATA456 paralogs (see below). Because all but one of the analyzed lophotrochozoan GATA456 genes have retained the ancestral intron/exon structure, we conclude that either the GATA456 paralog clusters in lophotrochozoans and arthropods and convergent intron losses. have independent origins, or that GATA456 genes in arthropods have undergone repeated

Orthology and evolutionary birth-order of the arthropod GATA gene complement.

We have used multiple lines of evidence to determine the origin and relationships of the more degenerate *Drosophila* GATA factors. A reciprocal best hit BLAST analysis suggested that only one (*Grain*) of the five *Drosophila* GATAs belongs to the GATA123 class, while the remaining four are in the GATA456 class. Additionally, in multiple phylogenetic analyses all five *Drosophila* GATAs formed single clades with other arthropod homologs, suggesting orthologous relationships between the GATA genes within each clade. Orthologous genes for each of three fly GATA456 genes--*Serpent, GATAe,* and *Pannier*--are present throughout arthropods. However, the fourth *GATA456* paralog, *GATAd*, appears to be an insect-specific duplicate found only in the beetle, bee, mosquito, and fruitfly genomes.

Additional evidence for the orthology of the four arthropod GATA456 genes comes from the observed conservation of gene order within a GATA456 paralog cluster among arthropods. The *Drosophila* GATA genes *Serpent, GATAe,* and *Pannier* are present within a tightly linked cluster, as are their best-hit orthologs in all the arthropod genomes we analyzed. Moreover, the relative orientations of these three genes are the same in every analyzed arthropod genome. In contrast, the insect specific GATA456 gene, orthologous to *GATAd*, is not linked in any of the four analyzed insect genomes.

Thus, the gene order is consistent with the predicted orthology defined by our phylogenetic analysis.

A third independent line of evidence for the orthologous relationships among arthropod GATAs emerged from our comparative analysis of the genomic intron/exon structures. The ancestral condition for the genomic structure of the conserved dual zinc finger domain of GATA transcription factors is three exons with conserved intron/exon boundaries, as found in all vertebrate, lophotrochozoan, and cnidarian GATAs analyzed. In contrast, arthropod GATA456 genes exhibit extensive modifications from this ancestral genomic organization. However, the suspected orthologs among the arthropod GATA456 homologs are united by unique clade specific intron/exon structures.

The observed synteny, as well as the pattern of intron losses and gains of the arthropod GATA factors, also provide an explicit mechanism for gene expansion via tandem-duplication and suggest an evolutionary birth order for the GATA genes during the expansion from one ancestral to four GATA456 homologs in arthropods. As illustrated in Figure III.6, we can infer the following sequence of duplication events. GATA456a/GATAe and the GATA456b precursor would have first arisen from an initial tandem duplication of a GATA456 gene that possessed the ancestral intron/exon structure. GATA456b then lost the second intron, resulting in a secondary state. The subsequent duplication of GATA456b formed GATA456ba/Serpent and GATA456bb. Gain of a novel second intron by GATA456ba/Serpent produced a third state. Following the divergence of insects from a pancrustacean (insect and crustacean) ancestor, a duplication of GATA456bb generated both the GATA456bba/Pannier

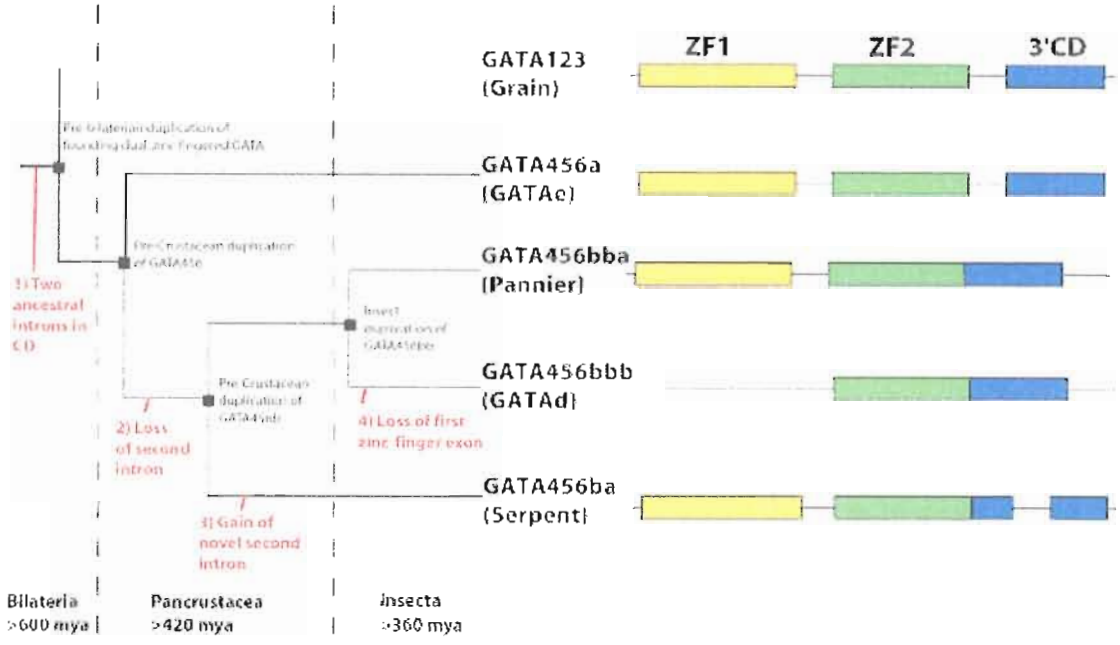

Figure III.6 Intron/exon structures define relationships and evolutionary birth order of arthropod GATA paralogs.
The intron/exon structures for the arthropod GATA456 genes suggest a birth order (as described in the discussion). The dotted line represents emergence of the last common bilaterian, pancrustacean, and insect ancestors. The red text represents points of intron/exon modifications, while the grey box describes the timing of gene duplication events. Intron/exon structures represent the inferred state for each paralog in the last common insect ancestor.

GATA456bbb/DmGATAd genes. GATA456bba genes appear to retain the second state, but the GATA456bbb genes appear to have lost the first exon (ZF1) early.

Our analysis shows the utility of such rare genomic changes as additional characters for resolving the relationship between deep branches in gene phylogenies. In this case, combining synteny and gene structure analysis with molecular phylogeny helped to resolve not only the obscure phylogenetic relationships of the highly derived *Drosophila* GATA456 genes, but also suggests the sequence of gene duplications that produced this gene family in arthropods. This evolutionary scenario makes predictions

about the sequences, intron/exon structure, and syntenies of GATA456 genes within arthropod phylogeny that can now be further tested by obtaining genome sequences from additional arthropods. This scenario also predicts similarities in expression and function of the orthologous GATA genes in arthropod development.

The identification of clear arthropod orthologs to *Drosophila* GATA factors also allows us to infer the origin of metazoan single-zinc fingered GATAs. Non-metazoan GATA factors, such as the fungal AreA proteins, generally possess single zinc-fingers, but most metazoan GATA factors possess dual zinc fingers. However, some invertebrate GATAs (e.g. *Drosophila GATAd* or *Serpent* isoforms) possess only single zinc fingers that might indicate their independent origin from single-zinc fingered ancestors. However, as described above, other arthropods possess orthologs of single finger *Drosophila* genes with two zinc fingers, indicating these genes all arose from dual-zinc fingered ancestral sequences. This conclusion is further supported by molecular phylogenetic analyses suggesting that all metazoan GATA factors are equally related to fungal outgroups (W.J.G, unpublished results). Finally, we also have been able to identify highly conserved individual amino acids that are diagnostic for individual arthropod orthologs and may be useful for identifying orthologous GATA factors in partially sequenced arthropods genomes (see Additional File 3).

Unresolved GATA456 orthology in other protostome phyla.

While we can use both gene sequence and synteny to infer GATA456 factor phylogeny among arthropods, it remains unclear whether the GATA456 expansion in arthropods is related to, or independent of those observed in other protostomes. For example, we can define nine *Caenorhabditis* clades, but we have little understanding of

how these relate to the four predicted ancestral arthropod GATA genes. Several of the nematode GATA genes are tightly linked, but most of these appear to involve very recent duplications. While sequenced genomes are now becoming available for additional nematode species, currently the data from the well supported *C. elegans* and *C. briggsae* GATA sequences provide the best platform to launch future inquiries into GATA gene family evolution in nematodes.

Likewise, we cannot fully resolve how lophotrochozoan GATA456 paralogs are related to each other, or to *Drosophila* GATA456 paralogs, despite similar copy-number and linkage. In each of three lophotrochozoan genomes, we identified three or more GATA456 paralogs. Two and three of these GATA456 paralogs are linked in an annelid (*Capitella)* and molluskan (*Lottia)* genome, respectively. In our molecular phylogenetic analysis, the *CcapGATANc* and *LgigGATANc* form a clade, and are both part of the GATA456 clusters in either organisms, suggesting these may represent orthologs across molluscs and annelids. However, our molecular phylogenetic analyses fail to define additional relationships for the GATA456 paralogs in lophotrochozoans, or between lophotrochozoans and arthropods.

Additionally, the sequence and expression of four of the five *Capitella* GATA factors was recently characterized (Boyle and Seaver 2008). They name these factors *CapI-gataA, CapI-gataB1, CapI-gataB2*, and *CapI-gataB3*, which appear to correspond to *CcapGATA123*, *CcapGATA456Na*, *CcapGATA456Nc*, and *CcapGATA456Nb*, respectively. Interestingly, the mRNA expression of the GATA123 ortholog was reported to be restricted to ectodermal derivatives, while the expression of the three GATA456 paralogs was described within nested mesendodermal territories. These

results provide additional evidence for a class-specific germ layer expression and expansion of GATA456 versus GATA123 transcription factors (Gillis et al. 2007).

One path to resolving gene-family relationships in other protostome phyla is to survey additional taxa, with a focus on slow evolving genomes and appropriate phylogenetic position. A better understanding of the *Drosophila* GATA factor evolution was possible only after additional arthropod genomes from suitable phylogenetic branches were included, many of which possessed less derived GATA sequences. However, the arthropods currently are an exceptionally well-sampled protostome phylum, and our findings suggest that the current sampling of lophotrochozoan genomes is not sufficient to resolve their gene family evolution. In order to extend our findings in arthropods to other phyla, it will be important to survey additional protostome genomes. Several ecdysozoan genomes have been targeted for whole genome sequencing, including those from basal taxa such as tardigrades and priapulids. Genome sequences from additional nematode species may help resolve the current ambiguity about the relationships of the *Caenorhabditis* GATAs.

Conclusions

In this study, we identified and examined the complete complement of GATA factors from nine newly sequenced protostome genomes. We have reconstructed the evolutionary relationships of these protostome GATAs using complementary forms of phylogenetic inference, including molecular phylogenetic analysis, genomic linkage and an examination of intron/exon boundaries. Our analysis indicates that protostome genomes have a single GATA123 ortholog, but multiple GATA456 paralogs. Furthermore, by including many arthropod genomes, we have been able to define

orthology for the more degenerate *Drosophila* GATA factors, including assigning the *GATAd, GATAe,* and *serpent* genes conclusively as GATA456 co-orthologs. Our examination of intron/exon structure modifications suggests a birth order of GATA456 paralogs, which could not be resolved in molecular phylogenetic analyses. This analysis has also identified similar tightly linked clusters of three GATA456 orthologs in both arthropods and lophotrochozoans, but additional taxa sampling will be required to define gene family relationships among diverse protostome phyla.

Methods

Identification of the conserved domains of the GATA transcription factor complement.

To identify putative GATA conserved domains, whole genome traces were downloaded to a local database and searched using two previously described *Platynereis* GATA factors and TblastN with each individual genome. Genome sequence from *T. castaneum* (Tcas_2.0) and *A. mellifera* (Amel4.0) was obtained from the Baylor College of Medicine Human Genome Sequencing Center . *I. scapularis* (iscapularis.TRACE-WIKEL.june07) and *A. gambiae (*AgamP3) genome sequence was obtained from the VectorBase (Lawson et al. 2007). *D. pulex, C. capitata,* and *L. gigantea* sequence data (v.1.0) was obtained from the US Department of Energy Joint Genome Institute . *S. mediterranea (v.3.1)* sequence data was produced by the Genome Sequencing Center at Washington University School of Medicine in St. Louis.

The TblastN hits from the genomic trace archives were validated and grouped using subsequent blast analyses. First, TblastN hits were validated by blastx against the Genbank NR genome, with a positive hit showing highest similarity to GATA sequences

in other organisms. Validated hits were then clustered, using blastn to search for like hits in the organism's trace archive, using these to group all positive traces and remove duplicates from the list of positive TblastN hits. The best deuterostome TblastN hit from each of the blastx analyses was recorded, and used for reciprocal best hit BLAST analysis to assign the initial orthology to known deuterostome classes. This process was repeated until no additional exons could be identified.

To assemble the individual exons, we used two distinct methods. In cases where a genome assembly was publicly available, contigs containing these exons were identified by blastn and compared to define the assembled exon structure for individual genes. In cases where no genome assembly was available, we attempted to first connect these exons by searching for traces with overlap between two exons. In the case where no single trace could be identified to connect two exons, we performed chromosome walks on the individual exon using the Tracembler program (Dong et al. 2007). These larger contigs, which was based upon overlapping sequence and also mate-pair relationships, were then used to determine linkage between genes.

Nomenclature for additional protostome GATA factors

We have named these additional protostome GATA factors using a nomenclature that reflects our inferred evolutionary relationships. In some cases, such as for the arthropod and insect GATAs, we can infer not only orthologous relationships, but also the sequence of duplications that lead to additional paralogs. In these cases, we use a binary naming system to describe each gene speciation event, adding an 'a' to one duplicate, and 'b' to the other. In cases where only uncertain orthology is inferred, we describe the orthology following our convention, and describe multiple, uncertain

orthologs using a capital 'N', followed by letters ranked by their degree of sequence conservation (from most conserved to least conserved).

Molecular phylogenetic analysis

A sequence file was made for each of the GATA genes using the conserved dual-zinc finger domain, consisting of the two zinc finger exons and the N-terminal portion of the following exon. These sequences were aligned using Clustalw (Thomsen et al. 1994), and then manual improvements were made in MacVector (see Additional File 4). Maximum likelihood analysis was conducted using PhyML-aLRT (Guindon and Gascuel 2003, Anisimova and Gascuel 2006) using a JTT model of evolution, and branch support given by the aLRT CHI2-based parametric statistic. Bayesian Inference was conducted using the MrBayes v3.1 (Huelsenbeck and Ronquist 2001), using the JTT model of evolution. The results are a consensus of two-converged runs of 3,000,000 generations, and branch supports given as posterior probabilities. Neighbor joining distance-based analyses was conducted using the MacVector program (v7.2.3)(Rastogi 2000), and the support given by bootstrap percentiles of 10000 replicates. For the nematode GATA factors, the complete sequence for each factor was aligned using Clustalx, a tree was generated using PhyML-aLRT, and includes support from both the PhyML-alrt CHI2-based parametric statistic and Bootstrap percentiles from a Neighbor Joining analysis in MacVector.

CHAPTER IV

THE ORIGINS OF VERTEBRATE GATA FACTORS: INSIGHTS FROM INVERTEBRATE DEUTEROSTOMES

Introduction

Within vertebrates, GATA transcription factors are required for the proper specification of cardiac and blood cell lineages, for the induction and differentiation of endoderm and mesendoderm, and in cell movement during gastrulation and neural projections (Patient 2002). Many of these roles appear to be widely conserved across diverse animals; for instance, in *Xenopus laevis,* overexpression of GATA4, 5, or 6 can induce endoderm formation (Afouda et al. 2005). In *Caenorhabditis elegans,* the *end-1* (a GATA4/5/6 ortholog) gene is necessary and sufficient to generate E or endodermal cell fate in *C. elegans*, but can also induce endoderm when overexpressed in *Xenopus* (Schoichet et al. 2001).

Relative to many gene families, the GATA transcription factors appear to be a relatively small and evolutionary tractable gene family, with only six members present in most vertebrates, five in insects, and eleven in nematodes. This gene family has undergone significant expansion in bilaterians compared to lower metazoans; for instance, only a single GATA gene can be found in two cnidarian genomes currently sequenced. Previous studies have shown that the six vertebrate GATA factors appear to belong within two classes of historically related genes, of GATAs –1, -2, and -3, versus

GATAs -4, -5, and –6 (Lowry and Atchley 2001). These two classes of GATA factors appear to be common throughout bilaterian animals, suggesting the last common ancestor of protostome and deuterostome animals possessed two GATA genes, having both a GATA123 and a GATA456 ortholog. Our recent survey of GATAs from the whole-genome sequence of multiple protostome genomes has shown that at least four GATA genes are present in every currently available protostome genome, including many additional gene duplicates within the GATA456 class (Gillis et al. 2008).

In contrast, two basal deuterostomes, the echinoderm *S. purpuratus*, and also the tunicate *C. intestinalis,* have been shown to possess just two GATA transcription factors, similar in number to the predicted ancestral bilaterian state (Lowry and Atchley 2001, Gillis et al. 2007). However, both of these genes appear to be relatively quickly evolved, showing only little supporting evidence – in conserved protein sequences, motifs, and genomic features – to belong within two different GATA classes. On the other extreme, a recent phylogenetic study of this gene family (He et al. 2007) concluded that the presence of only two GATAs in *S. purpuratus* and *C. intestinalis* represents secondary and independent losses in these lineages, relative to the eleven nematode and six vertebrate GATAs. In addition, both echinoderms and tunicates appear to have undergone radical shifts in their developmental modes relative to other phyla, making it difficult to ascertain the number, structural features, and role of the ancestral deuterostome GATA complement.

To resolve this ambiguous situation we analyzed the GATA complement within the whole genome sequence of members of two additional basal deuterostome phyla, the hemichordate *Saccoglossus kowalevskii* and the cephalochordate *Branchiostoma floridae,*

and analyzed these to help inform the overall pattern of deuterostome GATA evolution. Intriguingly, we found one well-conserved GATA gene for each of the two ancestral bilaterian GATA classes within each genome. Each of these GATA genes exhibits highly conserved sequence, structural, and genomic features compared to the predicted ancestral bilaterian gene set. Hence, our analyses show that both the hemichordate and cephalochordate retained very slowly evolved GATA genes within their genome. Thus, our study provides the strongest evidence so far for two distinct GATAs, one GATA123 like gene and one GATA456- like gene, in a deuterostome ancestor.

Materials and Methods

Identification of *Branchiostoma floridae* GATA sequences

Initial identification of two GATA gene fragments was conducted using tblastn analysis of the *B. floridae* trace archive. These fragments were used to search for the chromosomal regions containing these sequences in the draft genome (1.0) of *B. floridae*. BfGATA123 was found on JGI_Scaffold27 (2842113-2842359), and *BfGATA456* was identified on JGI_Scaffold160 (117014-51465).

Additional sequence on the 5' and 3' ends were identified via Blast 2 Sequences (bl2seq) (Tatusova 1999) comparisons against GATA123 and GATA456 orthologs from *Platynereis*, *S.purpuratus*, and vertebrate GATA sequences. We also identified 19 expressed sequence tags (ESTs) for the single BfGATA123 gene, which allowed a precise definition of the full-length coding sequence of this gene (1419 NT, 478AA). No published ESTs were available for the predicted GATA456 sequence.

We amplified by PCR these two GATA fragments from a *B. floridae* Gastrula-Neurula stage cDNA library, generously provided by James Langeland. For BfGAT456, we isolated two different sized clones (859 nt and 874) using the following primers: F1BfG456 (BfGATA456-'MYQ') 5'- ATGTACCAGAATCACTCCGTCGCG -3'; R1BfG456 (BfGATA456- '3'aln') 5'-ATTACTGGTGCTAGTTGGAGGCTTGC –3, designed to conserved regions from the 5' and 3' regions of our in silico predictions. Based upon comparisons to the chromosomal regions, these fragments appear to be alternate splice forms, with the smaller clone (BfGATA456-isoform b) possessing an alternative second exon, resulting in the loss of the first zinc finger in this isoform.

For BfGATA123 PCR amplified a 772 base pair fragment (corresponding to nt129-1070 of the published BfGATA123 cDNA) using nested gene specific primers F1BfG123 (BfGATA123 - F129) 5'- AGACATCGACGTGTTCTTCCACCA -3'; F2BfG123 (BfGATA123 - F300) 5'- CATGCAGTGGATCGAGAGTACCAA -3'; R1BfG123 (BfGATA123-R1128) 5'- TGTCTGGATGCCGTCCTTCTTCAT -3' R2BfG123 (BfGATA123-R1070) 5'- TAAAGTCCACAGGCGTTGCACACA -3').

Identification of *Saccoglossus kowalevskii* GATA sequences

A bioinformatics pipeline, Gene Family Finder (GFF) was developed to facilitate the identification of gene-family members within genomic trace archives, and used to search for GATA genes from the hemichordate *Saccoglossus kowalevskii*. This tool takes a user inputted protein sequence, and compares this to a local genomic trace DNA blast database using protein-translated nucleotide (tblastn) comparison. The program then groups all the initial hits into unique groups of redundant traces by taking into

account the greater divergence between nucleic acid sequences relative to amino acid sequence. Our program iterates through the initial hit list, performing a blastn search on the hit identifying redundant traces based upon a high e-value cutoff. These redundant traces are then collected into their own unique hit file, and used to remove these traces from the initial results list. These unique-hit redundant traces are assembled into a larger DNA contig using the CAP3 program (Huang 1999). These contigs are compared to the input sequence using the bl2seq program, and these results are parsed to display both the aligned region, and translation(s) of the contig based upon significant bl2seq identified frames. To test the utility of this program, we conducted searches within previously analyzed (see chapter II) and better annotated genomes, and were able to identify conserved open reading frames (ORFs) for the complete GATA gene complement. Searching the current trace archive of *S.kowalevskii* from the NCBI trace archive (8,246,246 sequences; last updated April 9, 2008) we identified a total of 8 ORFs that belong to a sole *SkGATA123*, and a sole *SkGATA456* gene.

Phylogenetic analysis

We manually aligned the newly identified *S. kowalevskii* and *B. floridae* GATAs to a previous alignment of the conserved dual-zinc finger domain (Gillis et al. 07). This data was analyzed using the phyml-alrt on a mac-powerpc processor; the JTT model of evolution was used, and Chi2-based branch supports generated using the approximate likelihood ratio test (Anismova and Gascuel 2006)(Guindon and Gascuel 2003).

Motif and splice site analysis

GATA1/2/3 and GATA4/5/6 motifs outside of the conserved dual-zinc finger domain were identified as described previously (Gillis et al. 2007), and were manually aligned to the *S. kowalevskii* and *B. floridae* orthologs. A motif was identified if it shared at least a 20% pairwise identity with another example of that motif. Splice boundaries were identified by using the Spalign program (Kapustin et al. 2008).

Results

Identification of hemichordate and cephalochordate GATA sequences

To investigate the origin of vertebrate GATA transcription factors and reliably reconstruct key ancestors leading to the vertebrate clade, we identified GATA sequences from the available genomes of two basal deuterostome invertebrates, the cephalochordate *Branchiostoma floridae* and the hemichordate *Saccoglossus kowalevskii*.

We have identified the presence of only two GATA factors within the cephalochordate genome, which has currently been sequenced to 8.1 x coverage. Initial identification were made using tblastn analysis of the *B. floridae* trace archives, using the local teleost blast servers (Catchen and Postlethwait, http://teleost.cs.uoregon.edu/blast/blast.html), in which we isolated two fragments of ~136 AA. These fragments both crossed three introns that contained the highly conserved dual zinc finger domain (Figure IV.1). We identified two genomic scaffolds from the pre-release genomic assembly JGI-assembled genome that contain these fragments (http://genome.jgi-psf.org/Brafl1/Brafl1.home.html). By conducting bl2seq sequence comparisons on larger regions of these scaffolds, we were able to identify the

less conserved 5' and 3' ends of each gene encoding the N-terminal and C-terminal regions of each protein.

In a blastn of these predicted *BfGATA* genes against sequenced EST libraries, we identified 19 ESTs for the *BfGATA123*, which allowed a precise definition of the full length mRNA of this gene (1419 NT, 478 AA). In addition, we confirmed the transcription of the *BfGAT123* gene by PCR amplification of a predicted 772 nt fragment with gene specific primers from a cephalochordate cDNA library.

For the second ortholog, *BfGATA456*, we were unable to identify any EST from the prerelease database. We therefore defined a contig for the 5' domain through the conserved dual zinc finger domain based on blt2seq sequence comparison with human and *Platynereis* GATA sequences. Gene specific primer were designed to a predicted 5' start codon, and the conserved dual zinc finger domain. We isolated two differentially sized clones via PCR from a cDNA library, which represented 859 and 874 nucleotide fragments. Using the Spalign program (Kaupustin et al. 2008), these fragments both aligned to the same region of JGI:scaffold 160, and are presumably alternative splice forms. These splice forms are identical with the exception of alternative second exons, with the smaller splice form incorporating a novel and seemingly unconserved exon which eliminates the first zinc finger domain.

We were able to computationally identify two GATA factors (*SkGATA123, SkGATA456)* from the *S. kowalevskii* genomic trace archive, which represent single GATA123 and GATA456 orthologs, respectively. Using the above-identified *BfGATAs*, we could identify four conserved open reading frames (ORFs) from each protein. Within the *SkGATA123*, we found two exons in the conserved domain, representing the first and

second zinc fingers, as well as two exons 5' to the conserved domain. We were unable to identify additional 3' sequences, including the 3' conserved domain exon described for other GATAs; however, it is possible this sequence is divergent or not represented in the current trace archive. For *SkGATA456*, we found three ORFs, which represent the complete conserved domain, as well as an additional single large 5' ORF.

Molecular Phylogenetic analysis of Deuterostome GATAs

To help define their orthology, we aligned the newly identified *S. kowalevskii* and *B. floridae* GATA factors to a previously defined alignment of animal GATA factors (Gillis et al. 2008) for use in molecular phylogenetic analysis. In addition, we included a GATA identified from another echinoderm, the starfish *Asterina minerata* GATA456-orthlog (Hinman and Davidson 2003) GATAe, as well as a GATA2/3 ortholog identified in the hagfish *Eptatretus burgeri.*

Using the alignment of the conserved dual-zinc finger domain from these species, we conducted a molecular phylogenetic analysis using a maximum likelihood technique. This analysis, shown in Figure IV.1, resolved the majority of GATA homologs into either the GATA123 or GATA456 subclass. However, the *BfGATA123* ortholog is the key exception, branching just before the division of these two subclasses. We believe that this is a result of *BfGATA123* being one of the most conserved GATAs identified, and also a high affinity of the *NvGATA* to the GATA123 class. In general, GATA1/2/3 orthologs appear to possess much shorter branch lengths relative to GATA4/5/6 counterparts, even in cases with an equal number of gene duplicates, as can be compared by the relatively short branch lengths of the vertebrate GATA1/2/3 versus the vertebrate

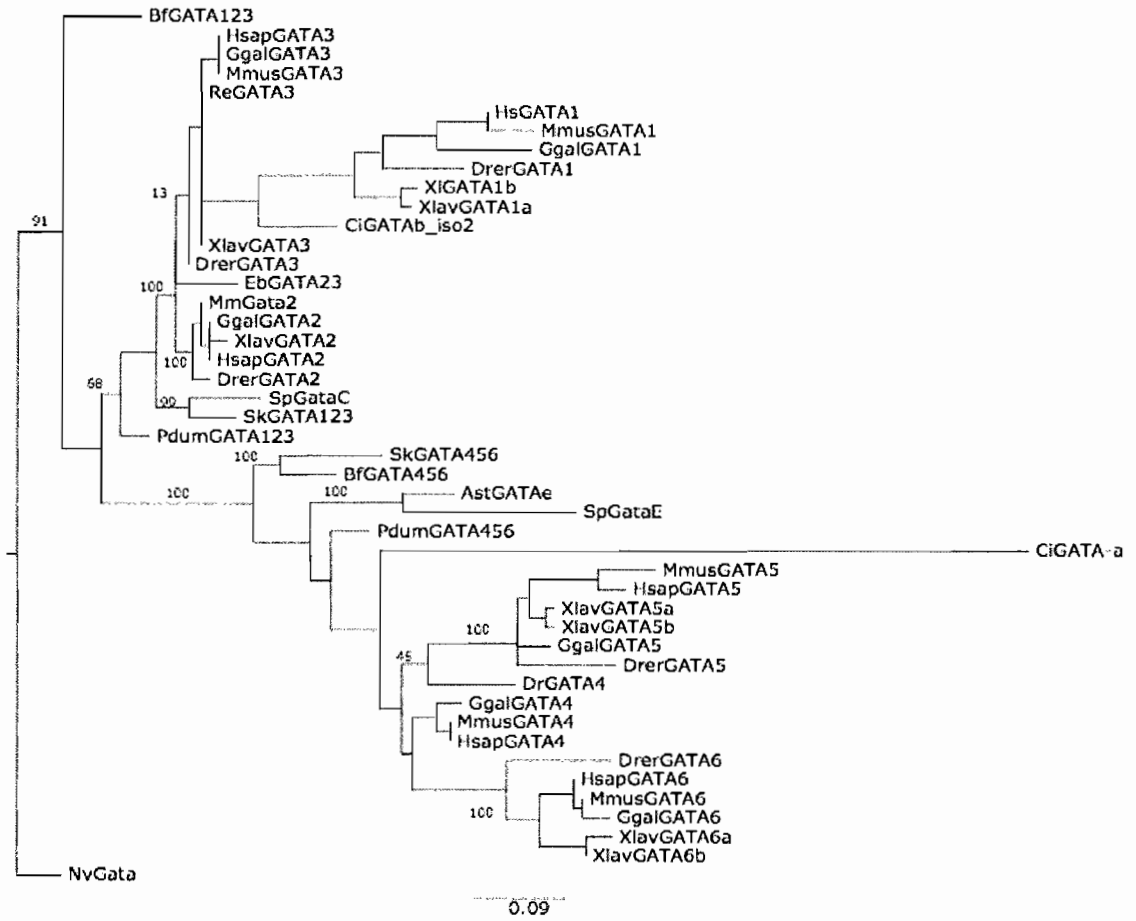

Figure IV.1 Phylogeny of deuterostome GATAs.
Chi-2 branch supports were generated by the approximate likelihood ratio test, and are represented on key nodes. Species names are as follows; As *Asterina miniata* (starfish); Bf, *Branchiostoma floridae* (cephalochordate), Ce, *Caenorhabditis elegans* (nematode); Ci, *Ciona intestinalis*(urochordate); Dm, *Drosophila melanogaster* (fly);Dr, *Danio rerio* (fish); Ep, *Eptatretus burgeri* (hagfish); Gg, *Gallus gallus* (chicken); Hs, *Homo sapiens* (human); Nv, *Nematostella vectensis* (cnidarian); Sk, *Saccoglossus kowalevskii*; Sp, *Strongylocentrus purpuratus* (echinoderm); Pd, *Platynereis dumerilii* (annelid); Re, *Raja erinacea* (skate); Xe, *Xenopus laevis* (frog).

GATA4/5/6 genes. Indeed, many of the shorter-branch genes appeared to diverge sooner then would be predicted based upon a priori knowledge of the species tree, presumably for a similar reason.

Identification of class specific motifs

In line with our phylogenetic analyses, the cephalochordate and hemichordate GATAs possess class-specific motifs outside of the zinc fingers identified in a previous analysis (Gillis 07). *BfGATA123* exhibits one of the most complete and well-conserved set of ortholog-specific motifs from our data set (see Table 1), containing all 7 previously

Table IV.1 Conservation of Invertebrate Deuterostome GATAs.

	123_N1		123_N2		123_N3		123_N4		123_N5	Dual-ZF Domain			123_C1		123_C2	
Gene	Pd	Nv	Pd	Nv	Pd	Nv	Pd	Nv	Pd	Pd123	Pd456	Nv	Pd	Nv	Pd	Nv
BfGATA123	10	30	27	47	53	61	58	52	29	94	82	90	11	17	2	7
SkGATA123	0	14	31	50	56	61	64	64	38	77	72	74	-	-	-	-
SpGATAc	-	-	7	14	56	56	70	58	21	88	81	84	17	17	25	11
CiGATAb	-	-	10	27	7	7	44	38	21	90	81	84	11	17	33	11
HsGATA1	-	-	-	-	-	-	-	-	-	82	74	77	-	-	-	-
HsGATA2	11	18	29	50	48	33	58	64	39	92	82	86	29	23	30	12
HsGATA3	6	17	38	31	52	34	65	58	36	92	82	87	41	5	33	25

	456_N1		456_N2		456_N3		456_N4	Dual-ZF Domain		
Gene	Pd	Nv	Pd	Nv	Pd	Nv	Pd	Pd123	Pd456	Nv
BfGATA456	20	16	60	20	41	37	8	87	90	82
SkGATA456	13	15	56	23	41	32	0	82	88	81
SpGATAe	18	15	46	10	17	25	17	80	84	80
CiGATAa			21	10	13	14	0	57	56	58
HsGATA4	-	-	43	18	40	21	13	83	90	79
HsGATA5	-	-	35	16	32	12	13	77	84	76
HsGATA6	-	-	50	12	34	22	17	76	84	74

The percent identity shared between motifs and conserved domains from cephalochordate (Bf), hemichordate (Sk), and echinoderm (Sp) GATAs compared to polychaete (Pd) and sea anemone (Nv) GATAs. Scores based upon pairwise alignment percent identity scores in individual alignments.

identified motifs, all of which possess a percent identity of 38-76% with at least one other example of that motif. *BfGATA456* contains all 4 motifs identified for within human GATA4/5/6 orthologs, and an additional N-terminal motif previously only identified in

the *Platynereis PdGATA456,* the sea urchin *SpGATAe*, and the sole anemone GATA *NvGATA*.

Similar to *BfGATA456*, *SkGATA456* possesses 3 N-terminal motifs identified within human GATA4/5/6 orthologs, and an additional N-terminal motif previously identified only in the *Platynereis PdGATA456,* the sea urchin *SpGATAe*, and the sole anemone GATA *NvGATA*. We have not yet found any *SkGATA123* conserved motif positioned towards the C-terminus from the conserved dual-zinc finger domain, but we have found all five previously identified N-terminal motifs.

Conserved splice site boundaries within the two deuterostome GATA classes.

A comparison of the genome assemblies and the translated amino acid sequence allowed for the mapping of splice sites. In a comparison (Figure IV.2) of the splice sites of the two cephalochordate and hemichordate GATA factors with vertebrate and anemone orthologs, we find conservation of two internal exons among all organisms. These exons encode the first and second zinc finger domain and are 47-55 AA and 41-42 AA in length, respectively. This appears to correlate with the high conservation of the dual zinc-finger domain in almost all animal GATA factors (Gillis 2007).

Although we see some variation in 5' and 3' exons, a class-specific pattern emerges when compared to the position of the conserved motifs. The cephalochordate *BfGata456*, the hemichordate *SkGATA456,* and the echinoderm GATA456 ortholog *SpGATAe*, as well as the human GATA4, 5, and 6 genes, possess a single exon 5' to the conserved domain, which possesses all of the identified motifs. However, *BfGATA123, SkGATA123,* and the human GATA -2 and -3 genes all possess two 5' exons, with the N1 and N2 motifs in the first exons, and N3-5 motifs in the second.

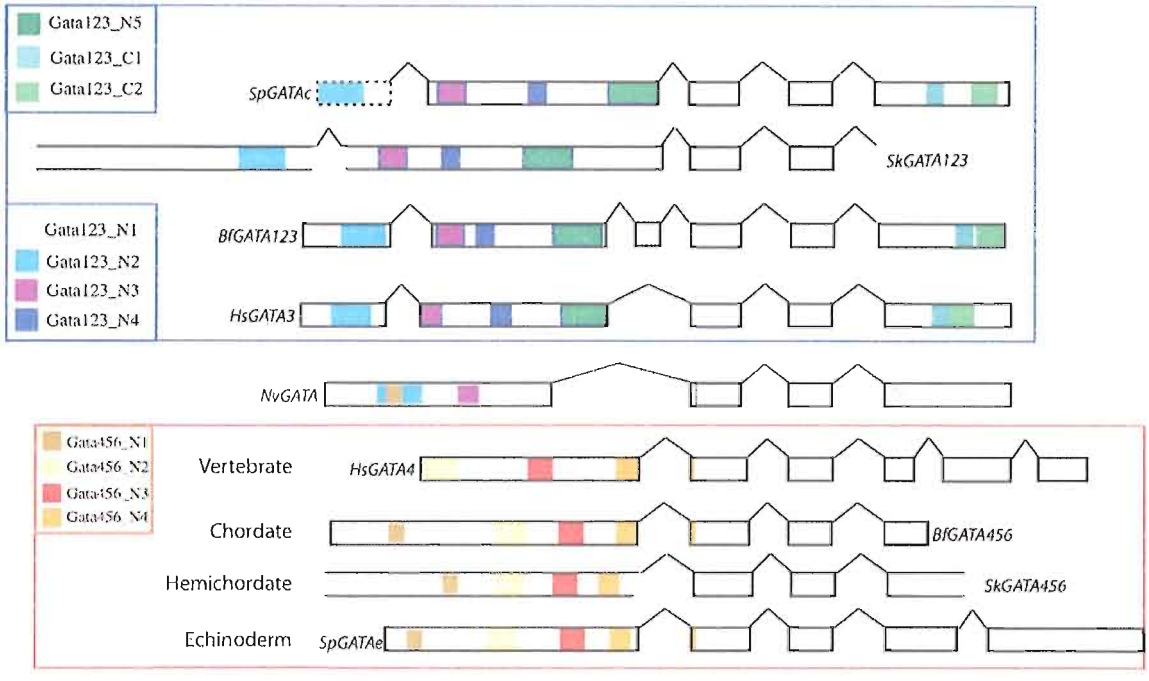

Figure IV.2 Exon Intron structure and conserved motif of Deuterostome GATAs. The identified exons are represented by solid blocks or a dotted line if boundaries are not confirmed by EST.. GATA123 orthologs for human (*HsGATA3*), hemichordate (*SkGATA123*), echinoderm (*SpGATAc*), and cephalochordate (*BfGATA123*) are found within blue block (top), and GATA456 orthologs are found in the red block (bottom). The sole cnidarian GATA from *Nematostella* (*NvGATA*) is shown centrally. Motifs are represented within the exons as colored blocks specified in the legends. The dotted line for the first *SpGATAc* exon represents its possible pseudo-exon status, and the open bars for SkGATAs are due to uncertainty regarding the exact ends of the exons.

Additionally, the 3' exons in the echinoderm, cephalochordate, and vertebrate GATA123 orthologs possess only a single large exon, which contains the tail end of the conserved domain, as well as two identified C-terminal motifs. We have not yet found additional exons 3' to the conserved dual-zinc finger domain in the hemichordate *SkGATA123*.

The 3' ends of the human GATA 4, 5, and 6 genes are split into 3 short exons, the first of which contains the tail part of the conserved domain and is conserved with the cephalochordate *BfGATA456*. However, the GATA456 ortholog in *S. purpuratus* has

only two large C-terminal exons, and only a short exon containing the 3' conserved domain is found in the hemichordate or cephalochordate GATA456 orthologs.

Prediction of an additional 5' exon in the *S. purpuratus* GATA1/2/3 gene (GATAc)

Based upon the conservation of two 5' exons in chordates and hemichordates (which represent a sister group to echinoderms), we found it surprising that only one exon was found in the previously characterized SpGATA1/2/3 ortholog, SpGATAc. Additionally, motifs found on the 5' most exon in other deuterostome GATA1/2/3 could be also identified in protostome GATA123 orthologs (Gillis et al. 2007). Together, this suggested that the *S. purpuratus* GATAc gene may have lost its first exon, or that the current gene is misannotated.

We conducted tblastn searches using other GATA1/2/3 factors to search for signs of an additional 5' exon. Although no similar sequences were identified using vertebrate or hemichordate sequence, using the 5' exon from *BfGATA123* we were able to identify a ~207 nt region approximately 18 kb upstream of the *SpGATAc* gene that shows a significant (.002) similarity. Indeed, a hypothetical translation of this region showed a 69 amino acid open reading frame. The region appeared to contain a GATA123_N2 motif, which was 64% identical to the BfGATA123_N2 motif on the amino acid level, (27% to Pd. 50% to NV, 56% to Vert (MmGATA2)). However, this ORF appears to begin abruptly within this motif, and does not contain a 5' start codon. However, it is possible we are missing an additional short 5' exon, the genomic reference is off, or that this exon has been lost, or we have identified a pseudo-exon.

Discussion

Invertebrate deuterostomes possess sole GATA123 and GATA456 orthologs

To examine GATA transcription factor evolution in deuterostomes, including vertebrates, we have identified the presence of single-copy GATA123 and GATA456 orthologs in two basal deuterostomes, the cephalochordate *Branchiostoma floridae* and the hemichordate *Saccoglossus kowalevskii*. These results extend on previous identification of single-copy GATA123 and GATA456 orthologs identified both in echinoderms (*S. purpuratus)* as well as urochordates (*C. intestinalis)*; however, the newly identified genes appear to show an even greater level of sequence conservation. This conservation has allowed us to confirm the presence of near complete sets of class-specific sequence motifs in these orthologs, which appear to correlate with conserved intron/exon boundaries in these genes. Thus, this level of sequence conservation has allowed us to use multiple lines of phylogenetic inference to independently confirm that the ancestral deuterostome and chordate - like the bilaterian ancestor – possessed only two GATA transcription factors.

Our analysis allows us to reconstruct several aspects of the ancestral deuterostome (Ud) GATA orthologs. One clear aspect is the presence of two N-terminal exons in the UdGATA123, versus only a single exon for UdGATA456-like genes. Furthermore, the two UdGATA123 exons appear to have maintained their boundaries relative to the presence of conserved-sequence motifs, with the UdGATA123 N1-N2 appearing in the first exon and N3-N5 the second exon. The GATA456 orthologs possess only a single N-terminal exon containing all of the conserved sequence motifs. Interestingly, these patterns appear to break down in the sole cnidarian GATA; although the NvGATA

possess conserved GATA123-like N1–N4 motifs, these are all contained in a single exon, suggesting that there has been an introduction of a new splice-site within the sole pre-existing exon leading to the UdGATA123 gene. We can also say that the human GATAs-4, -5, and -6 appear to have lost the initial N-terminal motif, which appears to be present within the deuterostome invertebrate GATA456s, as well as within the protostome *Platyneries dumerilii.* A blast search of this region against the NR protein database fails to find this motif in any vertebrate GATA, suggesting that this may have been lost early during vertebrate evolution. Additionally, the GATA456 N2-N4 motifs in the human GATAs –4, -5, and –6 appear to be spaced further apart then in their deuterostome invertebrate counterparts, suggesting that this exon has grown in length in the vertebrate lineage. In general, the hemichordate and cephalochordate GATAs appear to retain more ancestral sequence features, such as shared percent identity to protostome and cnidarian orthologs, the presence of a more-complete ancestral exon/intron state, and of class-specific motifs, then either urochordate or echinoderm sequence. Furthermore, although the vertebrate GATA proteins (with the exception of the GATA-1 like sequences) appear to moderately well conserve sequence, they have undergone multiple duplications relative to the deuterostome invertebrates.

Independent expansions of vertebrate and protostome GATA456 factors

In a previous analysis, we had identified frequent duplication of the protostome GATA456 orthologs; however, in deuterostome species we have identified additional GATA duplicates only in vertebrates so far (Figure IV.3). As we presented strong evidence that the basal chordate only had two distinct GATA factors, we now have additional strong support that the vertebrate GATA family has expanded by retention of

GATA genes that originated by two rounds of whole genome duplications (Dehal and Boore 2005).

We can further test this prediction by analyzing additional genomes closer to these whole genome duplication events; for instance, lampreys and hagfish are thought to have only undergone one round of whole genome duplication, therefore we would predict to find two GATA123 and GATA456 orthologs in this lineage. Additionally, as cartilaginous fish are thought to be closest to vertebrates that experienced the second round of whole-genome duplication, it may be possible that they retained the additional GATA123 or GATA456 duplicate thought to be lost in either lineage.

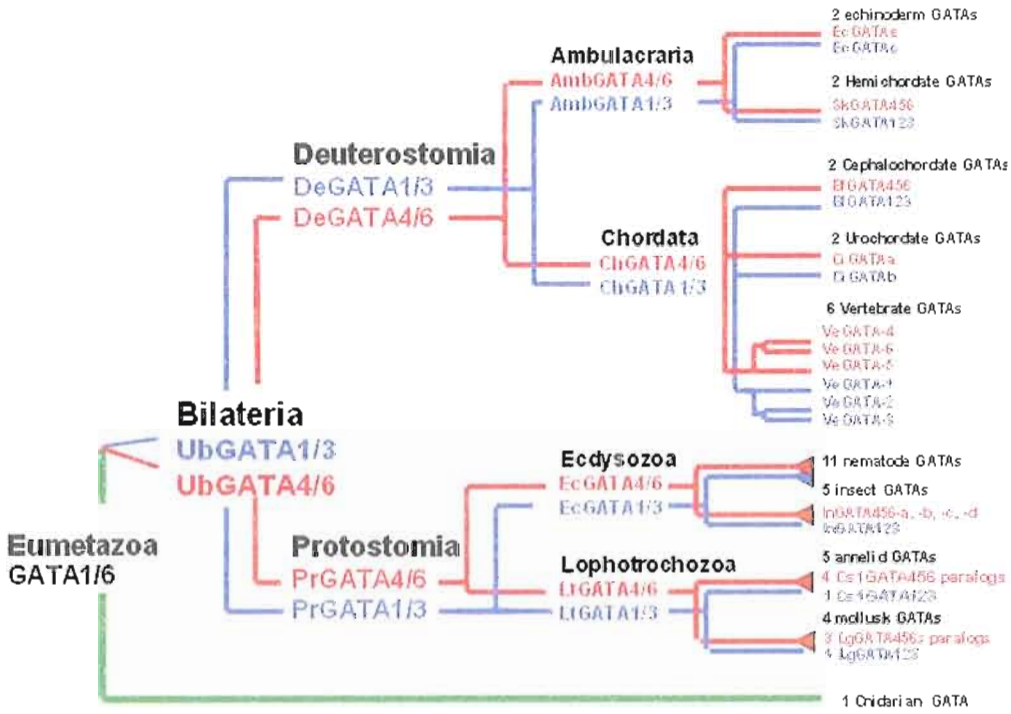

Figure IV.3 Overview of Animal GATA factor evolution
A schematic showing our current depiction of GATA evolution in animals. An early GATA duplication occured at the base of the bilaterian tree, resulting in GATA1/2/3 (blue) and GATA4/5/6 (red) orthologs. Expansions of the GATA4/5/6 orthologs have occurred in protostomes via tandem duplication, either in independent lineages (as shown above), or perhaps earlier during protostome evolution (not shown).

In no case have we identified additional deuterostome GATA duplicates via tandem duplication, an event that appears to be common in protostome evolution. The only vertebrate we have identified as possessing more then six GATA genes is *Xenopus laevis*, which possesses duplicate GATA1, GATA5, and GATA6 genes. However, *Xenopus* has undergone a relatively recent tetraploid event, meaning these additional duplicates are also likely duplicated via whole-genome duplication. Interestingly, only six GATA genes appear to have been maintained in zebrafish, which also has undergone a third round of genome duplication.

Greater Sequence Conservation of GATA123 orthologs

Our comparisons of the deuterostome GATA123 genes to the sole cnidarian GATA (NvGATA) suggest that they are more slowly evolving then their GATA456 counterparts. This can be seen both in the higher percent identity shared between the deuterostome GATA123 and NvGATA conserved domains (Table IV.1), the high affinity of the BfGATA123 with the NvGATA, and the total number of common motifs we can identify.

One possibility is that GATA123 genes are constrained more based upon retention of a deep ancestral function (at least one common between cnidarians and bilaterians), while the GATA456 class is more diverged, perhaps due to the selection or incorporation of bilaterian specific roles. However, the conservation of GATA123 orthologs appears to be in contrast with expression patterns described for deuterostome and cnidarian GATA factors. Whereas *NvGATA* mRNA is largely restricted to the endoderm in *Nematostella*, with only a small ectodermal expression domain, the

Vertebrate GATAs-1, -2, and –3 are mostly within ectodermal tissues and blood and not in the endoderm.

Prediction of a novel exon via intron exon structure

Additionally, we show how a detailed comparison of smaller conserved motifs and splice sites can be used to more correctly predict gene structure. We found a potential 5' exon to the previously characterized *S. purpuratus* GATAc gene based upon the absence of a N-terminal motif that is generally present both in deuterostomes and protostomes. However, the finding of just two conserved genes in the basal deuterostomes opens up the possibility of exploring the overall function of the GATA123 or GATA456 subclasses. It has been shown that redundancy between GATA orthologs may mask their overall phenotypes. For instance, single or even double gene perturbation of GATA4, -5, and –6 orthologs is not sufficient to completely block cardiac mesoderm induction in zebrafish or *Xenopus* (Peterkin et al. 2007, Afouda et al. 2005), and triple perturbation of these orthologs reveals a much potent endodermal defect in *Xenopus*. The overlapping expression domains in the CNS for the GATA2 and 3 (Nardelli et al. 1999) and in hematopoietic lineages for GATA1/2/3 orthologs (Patient 2002) likewise suggest that this may the case for these orthologs as well. The single copy-state of these genes in both a hemichordates and a cephalochordate, in addition to possession of morphological features which are thought to be more conservative to the ancestral deuterostome and chordate respectively, make these ideal models to examine the ancestral function of the GATA4/5/6 and GATA1/2/3 vertebrate classes.

CHAPTER V

DEVELOPMENT AND CHARACTERIZATION OF *PLATYNEREIS* GATA ANTIBODIES

Introduction

Embryologists have long been fascinated with the homology of embryonic germ layers, or distinct layers of cells within the developing embryo, across a wide variety of animals. For instance, bilaterian animals are all thought to be triploblastic, possessing three germ layers; an outer ectoderm which gives rise to the outer epidermis and nervous tissue, an inner endodermal layer which forms the midgut, and the intermediate mesodermal layer which gives rise to muscle.

There are many genes that appear to be playing similar roles in the formation of these germ layers, however, very few analyses have put these genes in a specific cell-lineage context. *Caenorhabditis elegans* being one of the few systems where this can be understood in such a fashion, and many of the important germ-layer specification genes have first been identified in this system. However, the extreme divergence and duplication of many *C. elegans* relative to other organisms often makes comparisons to other organisms intractable.

The polychaete *Platynereis dumerilii* may serve as a good model system to study the comparative roles of early developmental genes for several reasons. First, the *Platynereis* embryo develops in a stereotypic spiral cleavage, with distinct endodermal, mesodermal, and ectodermal cells thought to arise very early in development. Thus, we have fewer cells at the key stages of germ layer formation the many other model systems

(e.g. Flies, Urchins, Vertebrates), making reproducible identification possible, similar to the situation for *C. elegans.* However, unlike *C.elegans, Platynereis* genes appear to be relatively well conserved, allowing accurate phylogenetic reconstruction of gene orthology across distantly related animals.

We have previously identified 2 *Platynereis* genes, *PdGATA123* and *PdGATA456*, in the GATA family of transcription factors, which are orthologs to vertebrate GATA1/2/3 and GATA4/5/6 genes, respectively. Similar to their vertebrate orthologs, *PdGATA123* mRNA is expressed in ectodermal lineages, whereas *PdGATA456* appeared to be expressed in endomesodermal tissues. However, to gain more insight into the cellular deployment of these genes, we have raised polyclonal antibodies to these two *Platynereis* GATA orthologs. Herein we describe the initial characterization of these two PdGATA antibodies via western blotting and immunohistochemistry. Additionally, we describe an immunhistochemical analysis of the *PdGATA456* antibody over a time course of endomesoderm formation and specification (10-24 Hours), and identified initial cells in which *PdGATA456* has undergone nuclear localization, as a marker for the activate GATA456 protein.

Materials and Methods

Antibody Development

The Monoclonal Antibody Facility at the University of Oregon generated the rabbit and rat polyclonal antibodies used in this study. For antibody production, 6 fusion proteins containing 100-200AA fragments of the *PdGATA* proteins were constructed using a his-tagged pET-28a vector (Novagen). These polypetides were designed to minimize potential overlap of antigenic sites between paralogs. The polypeptides inserted into the

fusion proteins were as follows: *PdGATA123*(ABK32791.1) aa181-288, aa157-288, and aa158-362, and *PdGATA456* (ABK32792.1) aa13-117, aa13-167, aa13-217. The fusion proteins were expressed in Escherichia coli BL21 (DE3) Escherichia coli BL21 (DE3). *PdGATA123*(aa 158-362) recombinant protein was purified using Qiagen Ni-NTA agarose, and was used to immunize rabbits. *PdGATA456(*aa13-217) recombinant protein separated by SDS-PAGE, and electroeluted from the gel, and used to immunize rats. To test antibody specificity, six fusion proteins containing the same regions from were constructed by using the pGEX4-T-1 vector (Amersham Biosciences).

Immunoblotting

Immunoblotting protocol adapted from Schneider and Bowerman (2007). To extract proteins from larval stages (24-48 hours past fertilization),~2000–4000 specimens were collected, washed to remove their jelly coats, and centrifuged to remove sea water (3000 rpm for 30 s at 4°C). 750 µl of ice-cold RIPA buffer (0.1% SDS, 0.5% DOC, 1% NP-40, 150 mM NaCl, 1 mM CaCl2, 50 mM Tris/HCl [pH 7.4), and protease inhibitors) was added, and the samples were homogenized (3 intervals of 1 minute beating, 30 seconds icing) using the Mini-Beadbeater (Biospec Products). Extracts were separated by SDS-PAGE gel electrophoresis, transferred to nitrocellulose (Amersham), and then probed with *PdGATA123* (rabbit; 1:10,000) or *PdGATA456* (rat:1:10,000). HRP-conjugated secondary antibodies (Jackson) and enhanced chemiluminescence (Amersham) were used for detection.

Immunohistochemistry:

Immunostaining protocol adapted from Schneider and Bowerman (2007). Synchronized batches of 10- to 24 hr-stage embryos collected, washed with acidified sea

water to remove their jelly coats, and then treated in 25 ml TCMFSW (50 mM Tris, 495 mM NaCl, 9.6 mM KCl, 27.6 mM Na2SO4, 2.3 mM NaHCO3, 6.4 mM EDTA [pH 8.0]) twice for 3 min prior to fixation to permeabilize the eggshell. Embryos were transferred to fixative (4% PFA in phosphate-buffered saline [PBS] supplemented with 0.1% Tween [PBT]) and fixed overnight at 4°C on a nutator. After five washes with PBT, embryos were either used directly, stored for up to 5 days in PBS supplemented with 0.01% Azide at 4°C, or dehydrated in methanol and preserved at -20°C. Fixed embryos were blocked for 1 hr at room temperature in 1xPBS, 0.1% DMSO, 0.1% Tween, 2 mg/ml BSA, 4% normal goat serum (Block 1) and were incubated with primary antibodies in Block 1 overnight at 4°C. After six washes in PBT, specimens were incubated with secondary antibodies in Block 1 for 2 hr at room temperature. After six washes in PBT, specimens were mounted in PBS or Slowfade (Molecular Probes),

Primary antibodies and reagents used for this study include affinity-purified rabbit β-catenin (1:100; Schneider et al., 1996), rat-tubulin (1:100; Serotec), mouse histone (1:200; Abcam), mouse and rabbit gamma-tubulin (1:2000; Sigma-Aldrich), rat GATA456 (1:1000), or rabbit GATA123 (1:1000). Fluorescent-conjugated secondary antibodies (Jackson) were used.

Affinity purification of PdGATA123 antiserum:

Affinity purification of the GATA123 antiserum was adapted from Koelle and Horvitz (1996). Approximately 1 mg of Protein for GST-Tagged GATA123 (aa157-288) fusion protein was separated using SDS-PAGE gel electrophoresis, and transferred to nitrocellulose. Transferred proteins were visualized using Ponceau S staining, and the band containing the GATA123 fusion protein was cut out. This band was dried, then

poorly bound protein was removed by treatment with 100 mM glycine/HCl ph2.5 for 5 min. After 2x2 min. washes in TBS [12.1g Tris; 9g NaCl; 1 l dH2O; pH 7.6], the membrane was blocked for 1 hour in TBS + 3% BSA on a rocker. After 2x2 min. washes in TBS, 1 ml of serum diluted in 4 ml of PBS was added, and then incubated on a rocker at 4*C. The supernatant was reserved, and then washed 2x 5min in TBS and 2x5min in PBS. The bound sera was eluted 2x by adding 500 µl glycine, incubating 10 min. with occasional vortexing and then adding the elute to a tube containing 100 µl 1M tris pH8.0 to return final ph to 7.0. Supernatant and elutes were compared on western to determine efficacy of affinity purification, and the aliquots of the elutes stored at –80° C.

Results

Characterization of *Platynereis* anti-GATA polyclonal antibodies

In order to study the role of the *Platynereis* GATA proteins and characterize their role in germ-layer development, we generated polyclonal antibodies against recombinant GATA proteins. We designed smaller antigenic polypeptide from 5' regions of the proteins, as seen in the alignment in Figure V.1. The specificity of these antibodies was tested by western blot, first against isolated recombinant protein, and then against the Platyneries Larval extracts.

We generated a rabbit-anti*PdGATA123* sera 06-08, which showed reactivity to recombinant GST-tagged protein. This antibody detected a strong band from *Platynereis* larval protein lysates at approximately 54 kD , but a longer exposure seen in Figure V.2 detected several fainter bands at higher and lower molecular. When tested in

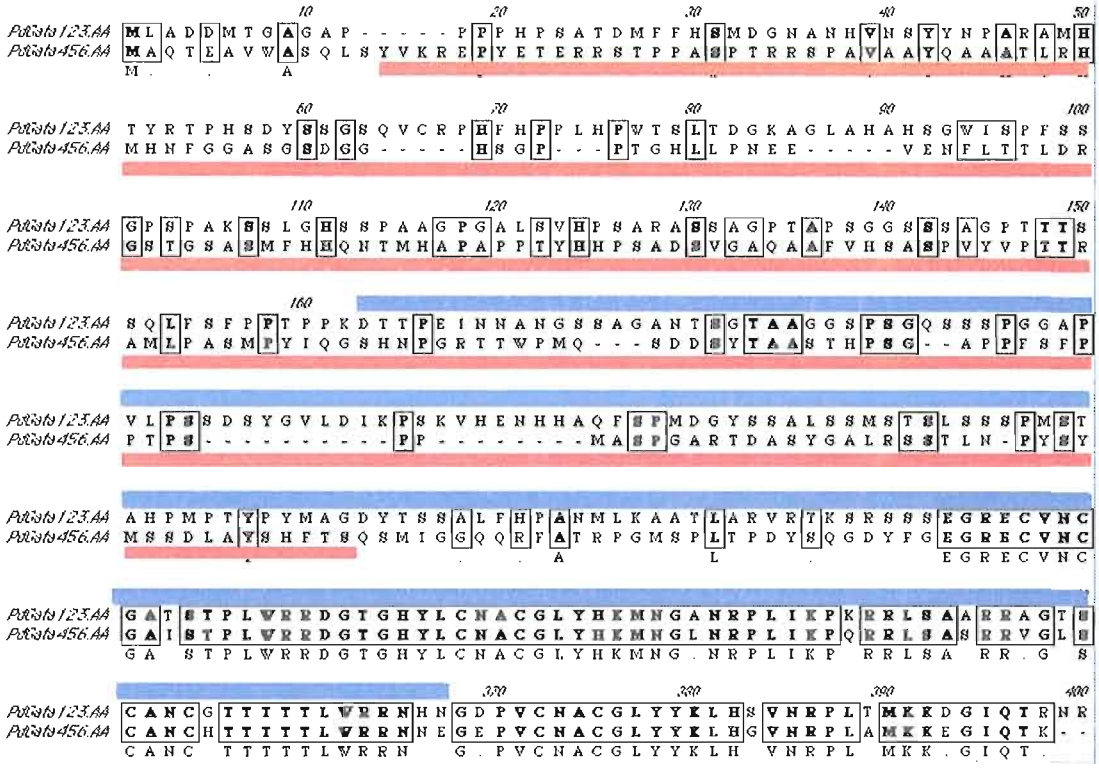

Figure V.1 GATA antigen alignment
Pairwise alignment of the two Platynereis GATA proteins, showing the region of the proteins used to generate the anti-PdGATA antibodies; GATA123 (blue) is located above the alignment, whereas the GATA456 (red) antigen is located below the alignment

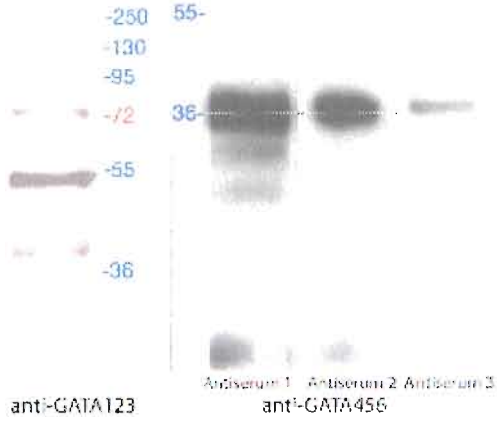

Figure V.2 GATA_Ab_Westerns
Westerns Blots of protein extract from 1 day Platyneries Larva to test anti-PdGATA antibodies. Rabbit-antiPdGATA123 serum 06-08 (left,1:2000) recognizes primarily a 54 kD protein. The rat-antiPdGATA456 antibodies from antisera-1(left, 1:1000) , -2(middle, 1:1000), and -3(right, 1:1000), all recognize a 36 kD protein.

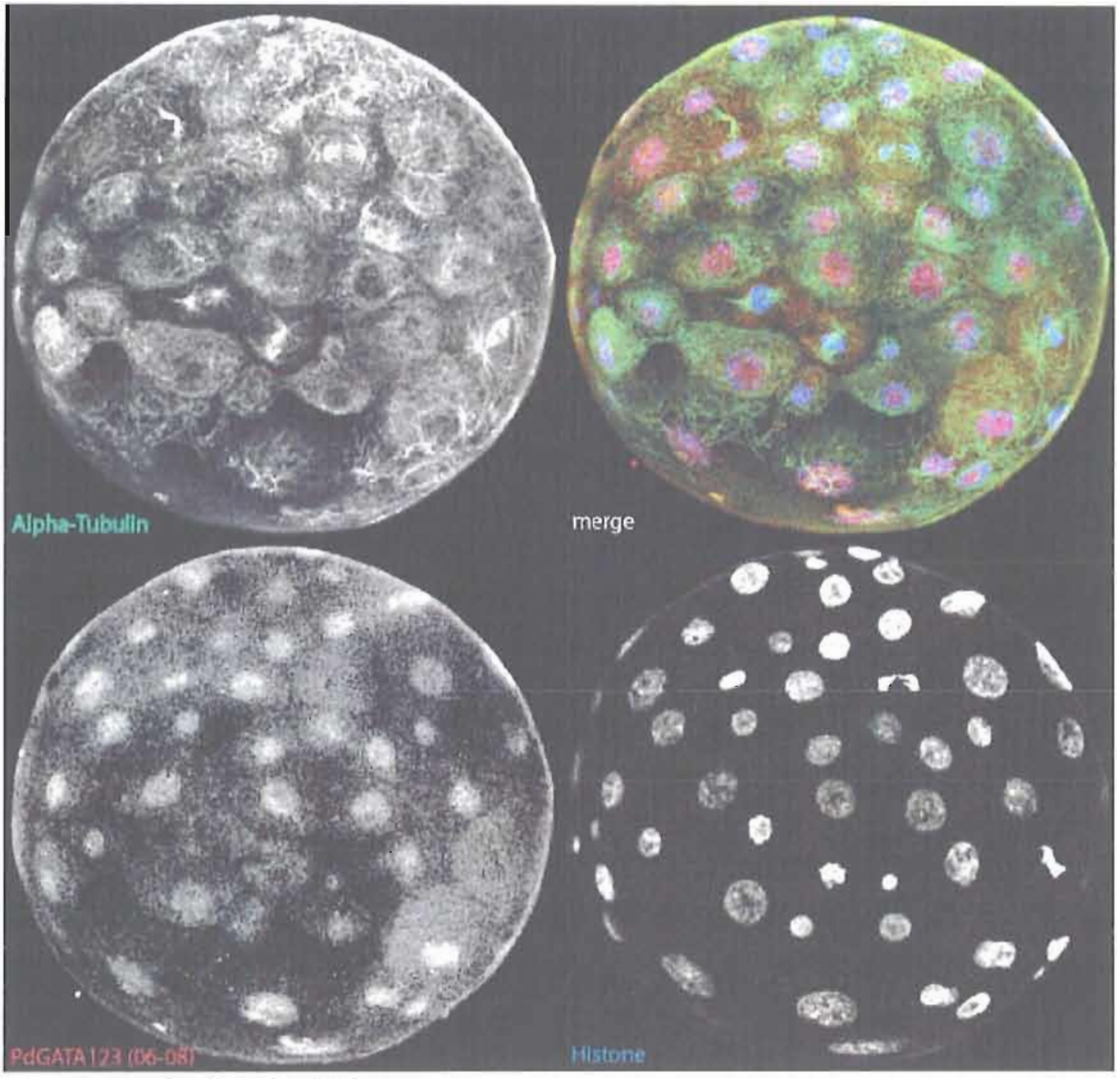

Figure V.3 GATA123 IHC
Whole mount immunohistochemistry (WMIHC) of an early (10 Hour past fertilization) *Platynereis* embryo. The anti-PdGATA123 antibody localizes ubiquitously to nuclei (Bottom Left, Red Channel) in non-mitotic nuclei, as detected by counterstaining with a nuclear stain (Histone, Bottom Right, blue), and a cytoplasmic gamma-tubulin stain (Top Left, Green).

We isolated sera from three rats that showed reactivity to recombinant *PdGATA456* protein, 876-1, 876-2, and 876-3 in western blots. All three of these detected a strong band around 36 kD separated 1-day *Platynereis* larval protein lysates, with 876-2 and 876-3 showing relatively clean bands, while 876-1 showed several fainter bands at lower molecular weights (Figure V.2). Only 876-2 appeared to show reactivity

in immunohistochemistry experiments, which appeared to show a differential nuclear localization pattern in the developing *Platynereis* embryo.

Domains of nuclear GATA456 protein in the early Platynereis embryo

In order to understand the pattern of nuclear GATA456 protein, we performed multiple time courses of fixations and a subsequent detailed immunohistochemical analysis. We identified several instances in which nuclear localization of GATA456 appears to occur in non-clonally related cells, suggesting several independent developmental events that trigger nuclear *PdGATA456* localizations (Initiation phase) In order to determine individual cells, we made use of the following histological makers; Gamma-Tubulin – a well known protein localized to centrosomes during mitosis – exhibits also differential cytoplasmic staining during embryonic cell cycles and histone as a nuclear stain.

The earliest nuclear localization of *PdGATA456* is seen in the four vegetal macromeres (Figure V.4) referred to as entomeres as they are destined to give rise to the endoderm. At this point (~96 cells), the two primary mesentoblast (4d1 and 4d2) have each undergone one asymmetric cell division, giving rise to two smaller vegetal daughter cells, the secondary mesoblasts 4d12 and 4d22, two larger animal daughter cells, the primary mesoblasts 4d11 and 4d21. Sometimes differences in the intensity of *PdGATA456* entomeric nuclear staining are seen, especially in 4C or 4D, apparently a technical issue, as it correlates with the similar variance of intensity seen with other

nuclear stains.

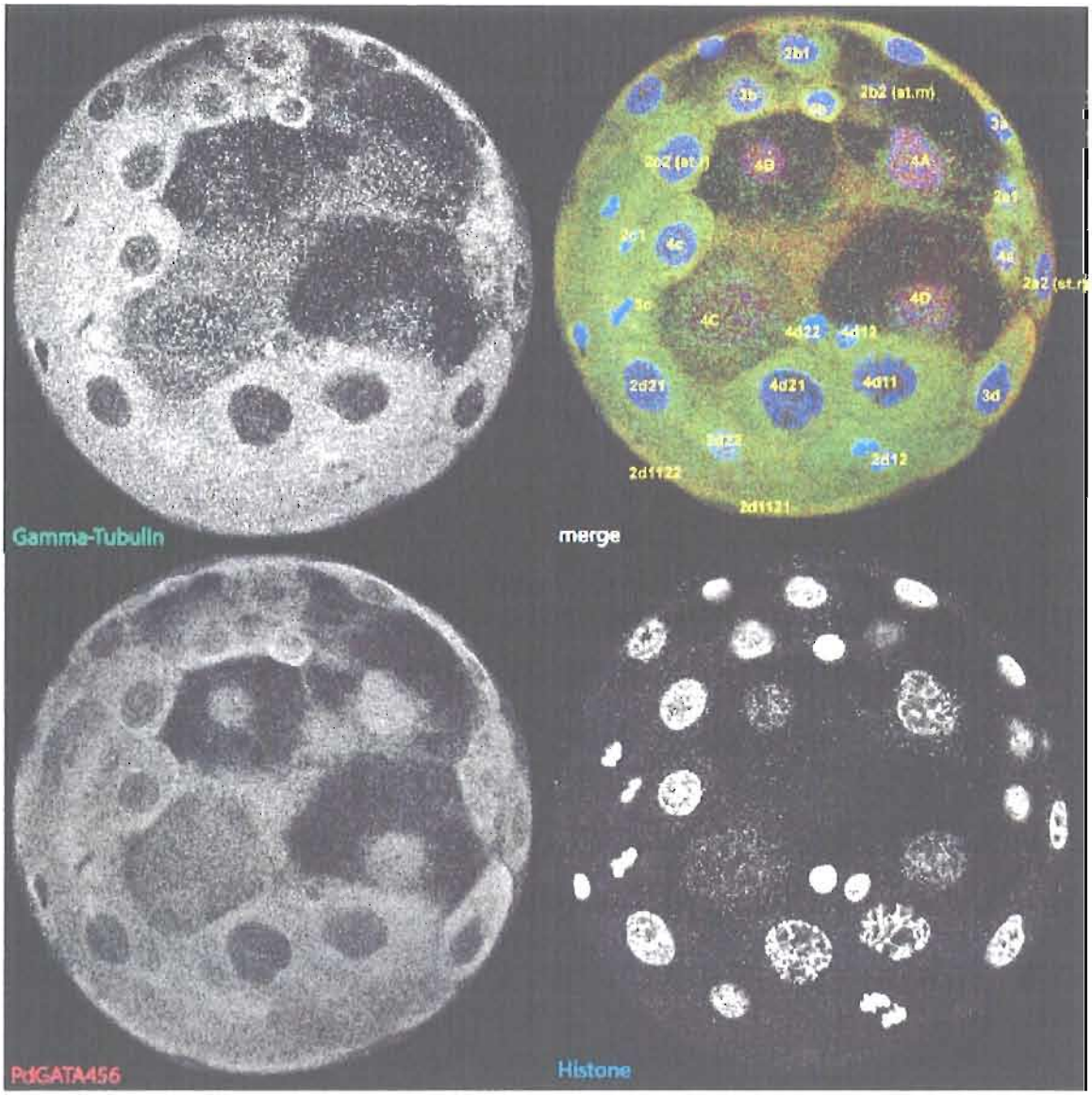

Figure V.4 GATA456_IHC_11H15
Whole mount IHC of a *Platynereis* embryo at ~11Hours 15 Min. past fertilization. This Posterior View of the early blastopore showing PdGATA456 antibody weakly localizing to entoplastic nuclei (4A, 4B, 4C, 4D, Bottom Left, Red Channel), similar to the faint nuclear staining seen for these macromeres as opposed to surrounding micromeres (Histone, Bottom Right, blue), and in contrast to the cytoplasmic staining seen with a Gamma-Tubulin antibody (Top Left, Green). Labels for identified cells are overlaid on the merged image.

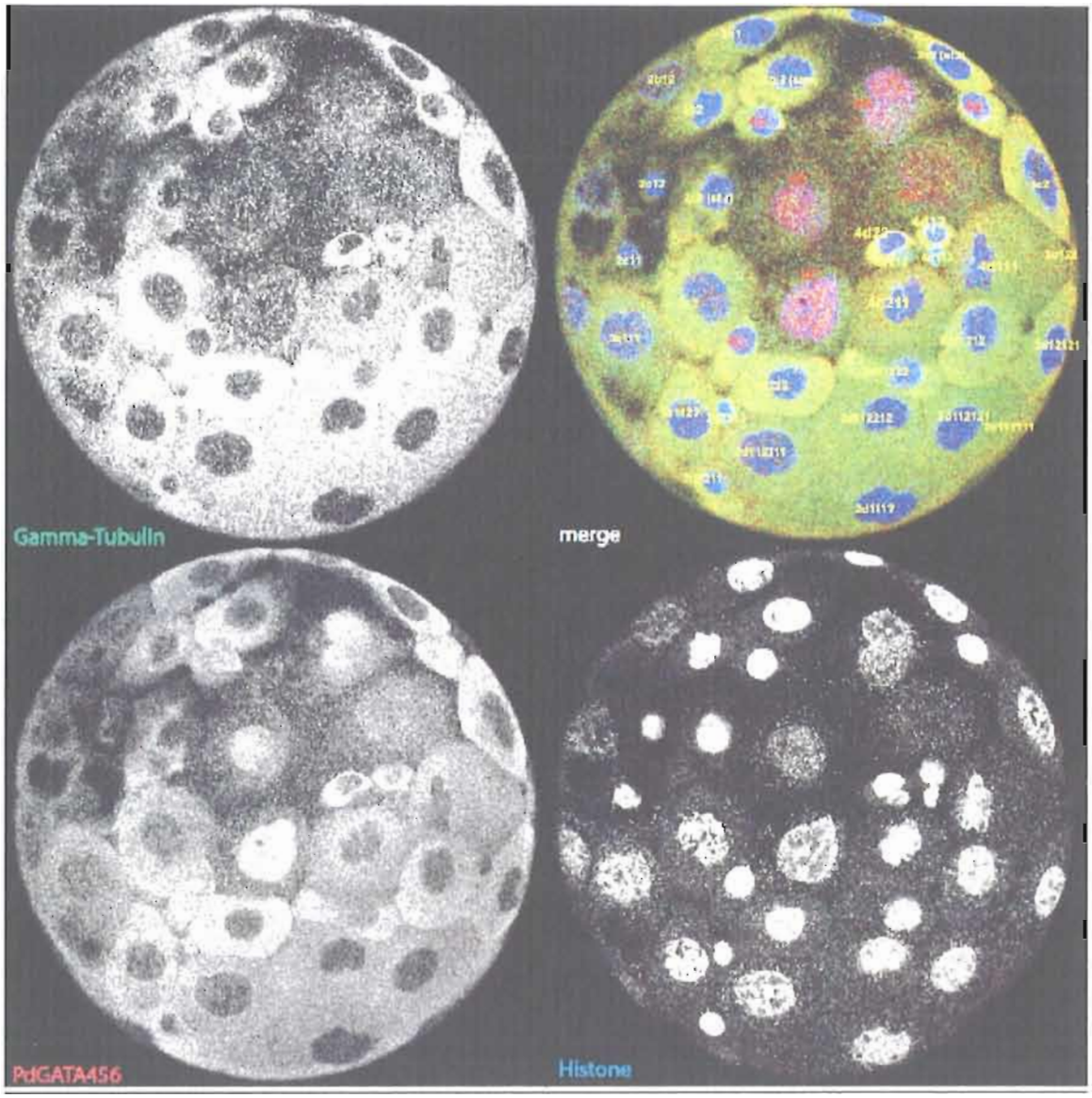

Figure V.5 GATA456_IHC_11H45
Whole mount IHC of a *Platynereis* embryo at ~11Hours 45 Min. past fertilization, approximately 1 cell cycle later. PdGATA456 antibody weakly localizing to entoplastic nuclei (4A, 4B, 4C, 4D, Bottom Left, Red Channel), as well as the tertiary micromeres 4a, 4b, 4c. Compare to nuclear (Histone, Bottom Right, blue), and cytoplasmic staining (Gamma-Tubulin, Top Left, Green). Labels for identified cells are overlaid on the merged image.

Around the 110-cell (pregastrula) stage, faint yet distinct nuclear localization occurs within three small cells, which we believe to be the three tertiary micromeres, 4a,

4b, 4c. (Figure 5). At this point, 4d11 and 4d21 have undergone a second asymmetric division, giving rise to a two more smaller secondary mesoblasts (4d112 & 4212) and two larger primary mesoblasts (4d111 and 4d211). Although it is difficult to follow these definitively due to their weak staining, I believe these are may be precursors of a small population of GATA456+ cells associated with the stomodaeum in later stages (see below).

Based upon our previous expression analysis (see discussion), we had expected to see nuclear localization in 4d lineages. However, as shown above, *PdGATA456* nuclear localization is not seen at the birth of 4d, after the bilateral division of 4d1 and 4d2, or two subsequent descendents from 4d1 and 4d2. We analyzed embryos at ~157-cell stage, which supports a third round of asymmetric cell division of the primary mesoblasts which generates two more smaller cells (4d1112 & 4d2112), none of which with nuclear GATA456 (data not shown). We begin to see staining in what appear to be 4d progenitors at the 181-cell stage, at which point a strong GATA456+ nuclear localization can be detected in medium-sized cells on the lateral surface of the primary mesoblasts. At this point, the primary mesoblasts each appear to have undergone 6-7cell divisions, though we have been unable to follow every division at this point. However, from the relative location and large-cell boundary shared between these GATA456+ cells and the primary mesoblasts we infer that these are indeed 4d descendants. On each side, these cells are soon joined another, slightly smaller GATA456+ cells laterally, and then another, smaller cells between these two (Figure V.6).

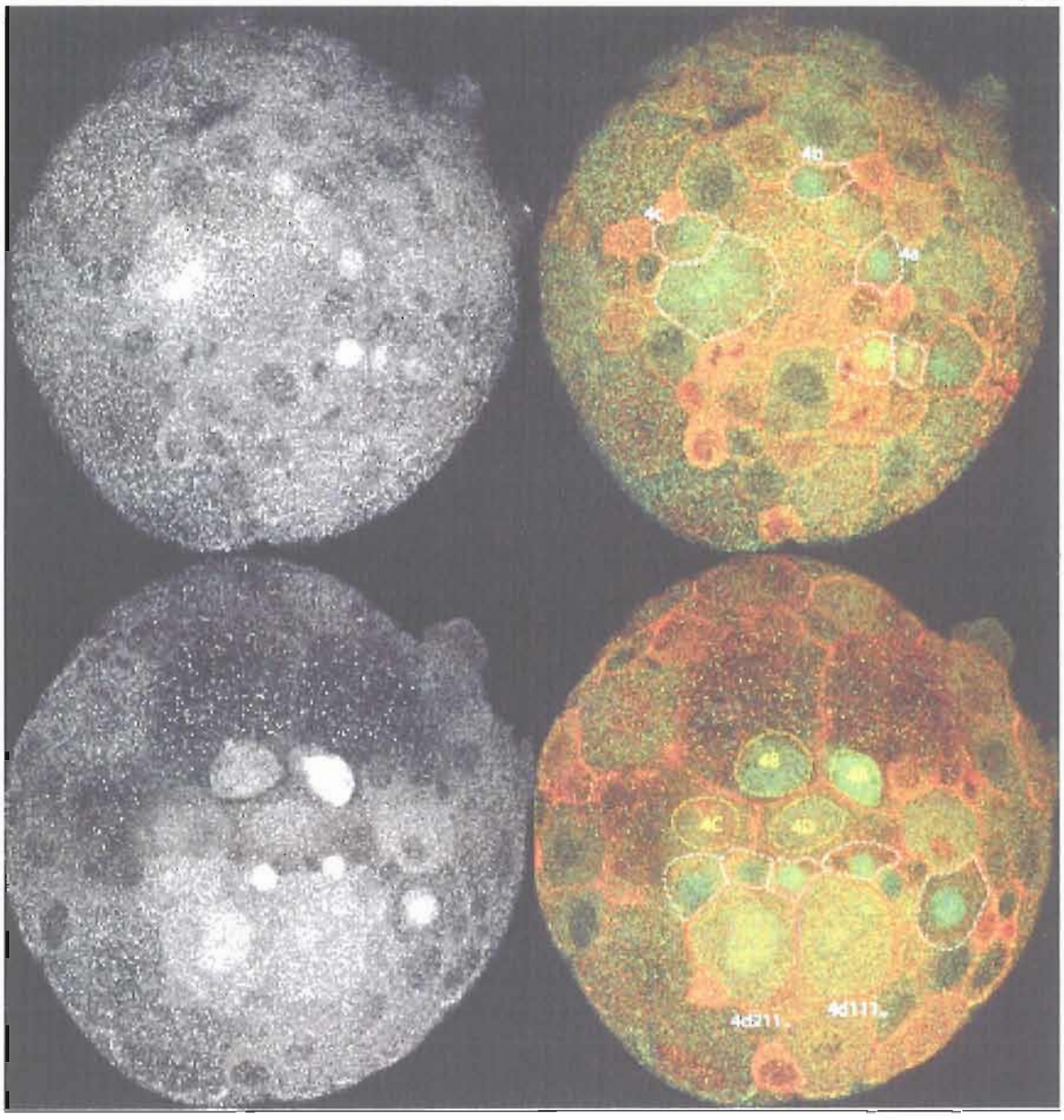

Figure V.6 GATA456_1IHC_14H35M

Whole mount IHC of a Platynereis embryo at 14Hours 45 Min. past fertilization. Two slices from a z-stack image file taken from the posterior pole (ventral to top), with the top image being a shallower image. Embryos have been stained with PdGATA456 (green, left) with Beta-catenin (red, right) as a marker for cell boundaries . GATA456+ cells are outlined in white, while the larger mesentoblasts are outlined in a dotted red line, and four entomoric nuclei are named and circled in yellow. The proposed identity for the three ventral-most GATA456+ cells have been suggested (4a, 4b, 4c), as well as the for some of the 4d mesentoblasts

Around this time, another micromere between the two mesoblast becomes GATA456+, ventral to the mesentoblast junction and anterior to the first secondary mesoblasts (data not shown). Slightly later, an additional medial GATA456+ cell appears , and these cells appear to take on a bilateral symmetry. At this point, in some especially crisp stainings, such as Figure V.6, the large mesoblasts themselves show faint GATA456+ nuclear accumulation.

Nuclear PdGATA456+ domains during cell proliferation and morphogenesis

We conducted a survey of GATA456 localization during from early initiation phase to a 24 hours post fertilization (HPF) trochophore larva, a very poorly understood period of *Platynereis* development (Figure V.7.,V.8, Table V.I) From around 14 hours post fertilization to the 24 hour trochophore, there is an expansion from ~5 GATA456+ cells to ~70. It appears that most of the GATA456+ cells appear to be within the paired mesoblastic bands, which previous lineage analyses have shown solely derived from the 4d lineage (Ackerman et al. 2005). By 18 HPF, the mesoblastic cells form paired kidney-shaped masses of ~20 cells/side, with more then half of these cells being GATA456 positive. These paired bands are separated from each by the posterior widening of the slit-like blastopore (see below). At 19 HPF, these paired mesoblasts stretch laterally, ventrally, and anteriorly, to form two horse-shoe shaped bands. These bands now possess two clear lateral bulges, each with a single large GATA456+ cell, and what looks like the beginning of a third bulge. At 22 Hours, each bands possesses over 40 cells, with ~20 of these being clearly GATA456+, and a second large laterally protruding GATA456+ forms a second .spoke ~2/3 of the way of the band.

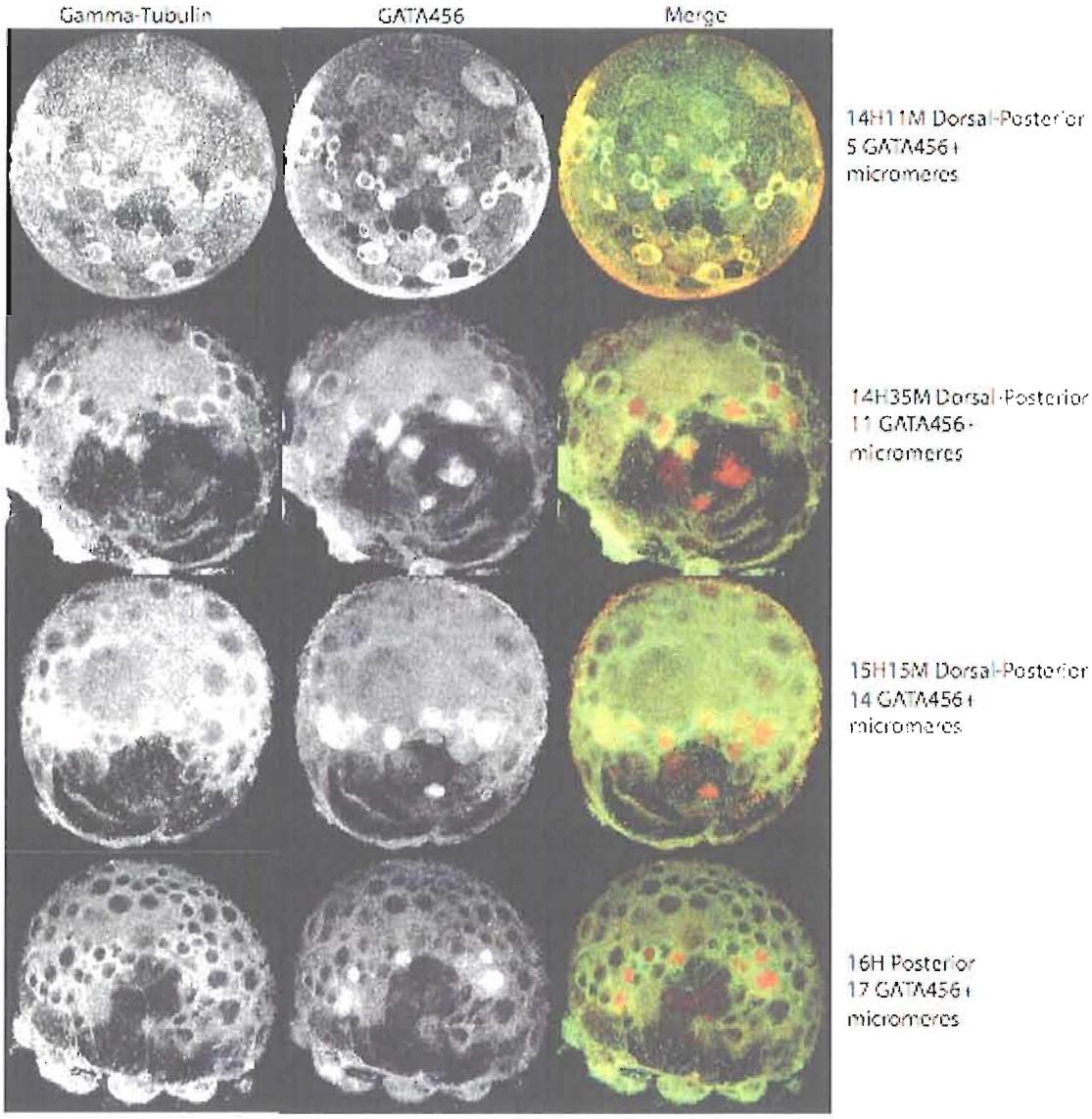

Figure V.7 GATA456 14-16H
Whole mount IHC of a *Platynereis* embryo at 14 – 16 Hour past fertilization (H- hours, M-minutes). The *PdGATA456* antibody (red) is separated in the middle column, and counts of GATA456+ cells are based upon a clear increase in nuclear-cytoplasmic staining, relative to the cytoplasmic stain of Gamma-tubulin (Left column, Green) Images represent individual slices from a z-stack image series, and may not display all GATA456+ cells. Posterior Views, Dorsal to Top.

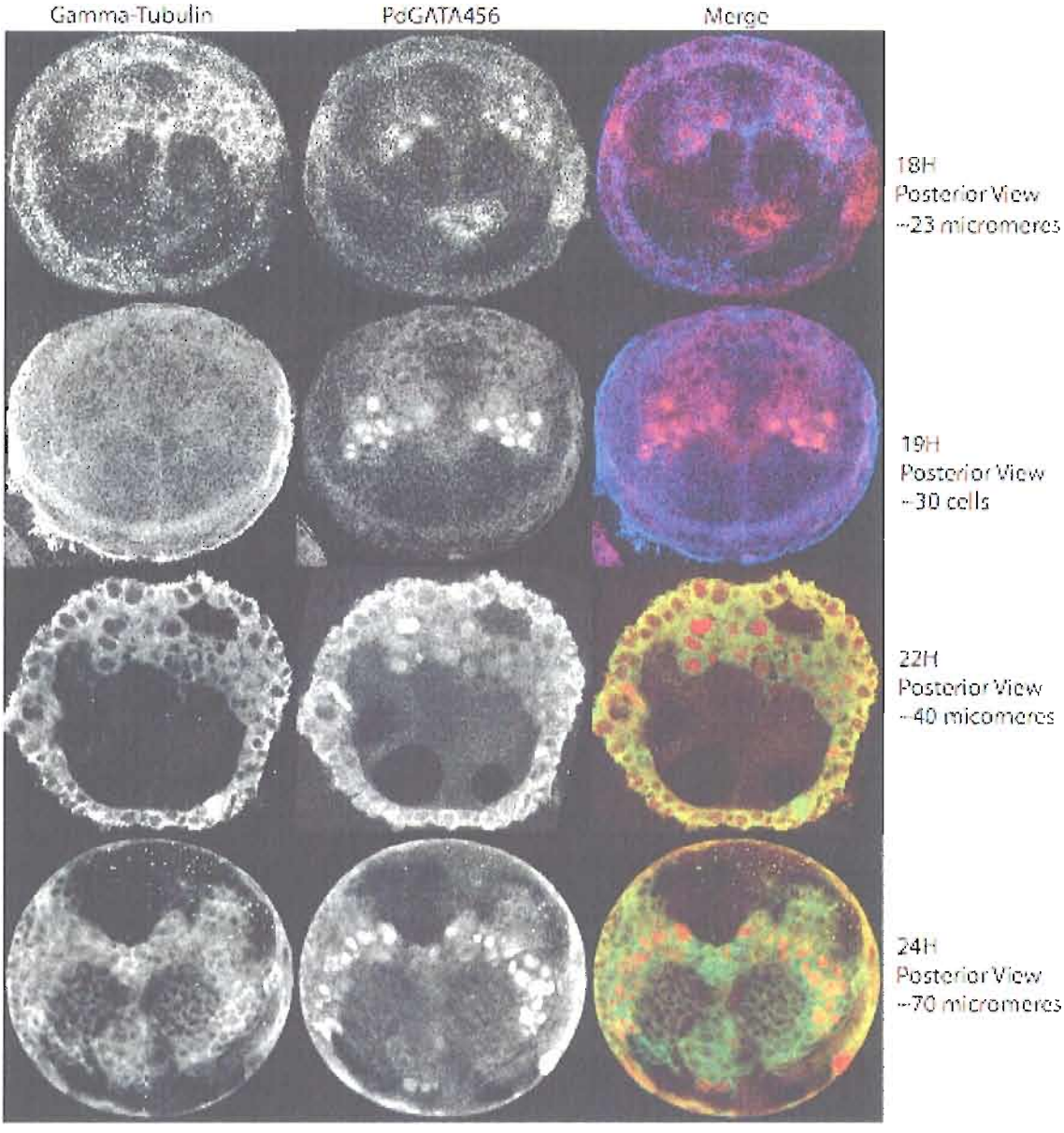

Figure V.8 GATA456 18-24H
Whole mount IHC of a *Platynereis* embryo at 18 – 24 Hour past fertilization (H- hours, M-minutes). The *PdGATA456* antibody (red) is separated in the middle column, and counts of GATA456+ cells are based upon a clear increase in nuclear-cytoplasmic staining, relative to the cytoplasmic stain of Gamma-tubulin (Left column, Green) Images represent individual slices from a z-stack image series, and may not display all GATA456+ cells. Posterior Views, Dorsal to Top.

Table V.1 GATA456 Cell Counts

Stage	N	*left micromeres*	*right micromeres*	*center micromeres*	total micromeres	entomeres
11H 15M	3	-	-	-	-	3.8
11H 45M	4	*1.0*	*1.0*	*1.0*	3.0	3.8
12H 15M	4	*1.0*	*1.0*	*1.0*	3.0	4.0
12H 45M	4	*1.0*	*1.0*	*1.0*	2.5	3.0
14H 11M	4	*1.8*	*1.8*	*1.3*	4.8	4.0
14H 35M	4	*4.3*	*4.3*	*1.5*	10.0	3.8
14H 55M	3	*3.3*	*3.0*	*3.3*	8.7	4.0
15H 15M	9	*5.9*	*6.8*	*1.2*	13.1	3.1
16H 02M	4	*7.8*	*8.0*	*1.0*	16.8	4.0
17H 10M	2	*11.0*	*9.0*	*1.0*	20.0	4.0
14H 35M	9	*3.7*	*3.8*	*2.8*	10.0	4.0
12H	2	*1.0*	*1.0*		2.0	-
16H	7	*4.0*	*3.8*	*1.8*	10.0	3.0
17H	1	*6.0*	*6.0*	-	12.0	-
18H	1	*12.0*	*11.0*	-	23.0	-
19H	4	*15.0*	*14.5*	-	29.0	-
22H	5	*16.8*	*18.8*	*3.8*	38.5	4.0
23H	5	*26.0*	*28.0*	*4.6*	58.8	4.0
24H	10	*31.0*	*33.2*	*5.3*	68.5	4.0

Average GATA456+ cell counts for different stages used in this analysis. These averages are based upon N-number of individual image stacks per stage, with counts of GATA456+ cells based upon a clear increase in nuclear-cytoplasmic staining, relative to the cytoplasmic stain of Gamma-tubulin (Left column, Green). These GATA456+ cells were separated into left, right, and central micromeres, as well as total micromere versus entomere staining.

At 24 hours, we can begin to place the domains of GATA456 expression in the context of several key morphologically distinct features. According to a previous report

looking at *brachyury* gene expression (Arendt et al. 2001), by 22 hours past fertilization, the

closed blastopore forms a slit below the ventral midline, is widened at either end like a dumbbell, with the anterior widening occurring at the base of the stomodeum or mouth rudiment, and the posterior widening forms the proctodaeum (anus). On the ventral/anterior sides of the proctodaeum, we see a distinctive pair of large GATA456+ cells, with a few smaller GATA456 cells directly anterior to each of these cells. These large cells directly abut a number of GATA456 negative cells, which appear to be the ectodermally-derived hindgut. The position and size of the large posterior GATA456+ cells are suggestive of these being the remnants of the primary mesoblasts, although we have cannot assert this definitively without a more detailed lineage analysis. The paired mesoblastic bands, which appear to be primarily composed of GATA456+ cells, abut the lateral edges of these putative primary mesoblasts, which now extend anteriorly and laterally and then medially along the anterior surface of the entomeres. Each paired mesoblastic wing appears to be a contiguous band, but possess three laterally-protruding thickened sections along their length. The anterior/medial margin of the mesoblastic bands appear to on either side of the anterior widening of the closed blastopore, which is thought to be the ectodermally-derived foregut. An additional small population of GATA456 cells (6-10), which form morphologically distinct pouch, can be seen underneath the anterior portion of the stomodeaum.

If these are clonally related, this means each of the GATA456+ cells begin to accumulate through development, primarily in cells on the ventral and lateral sides of the

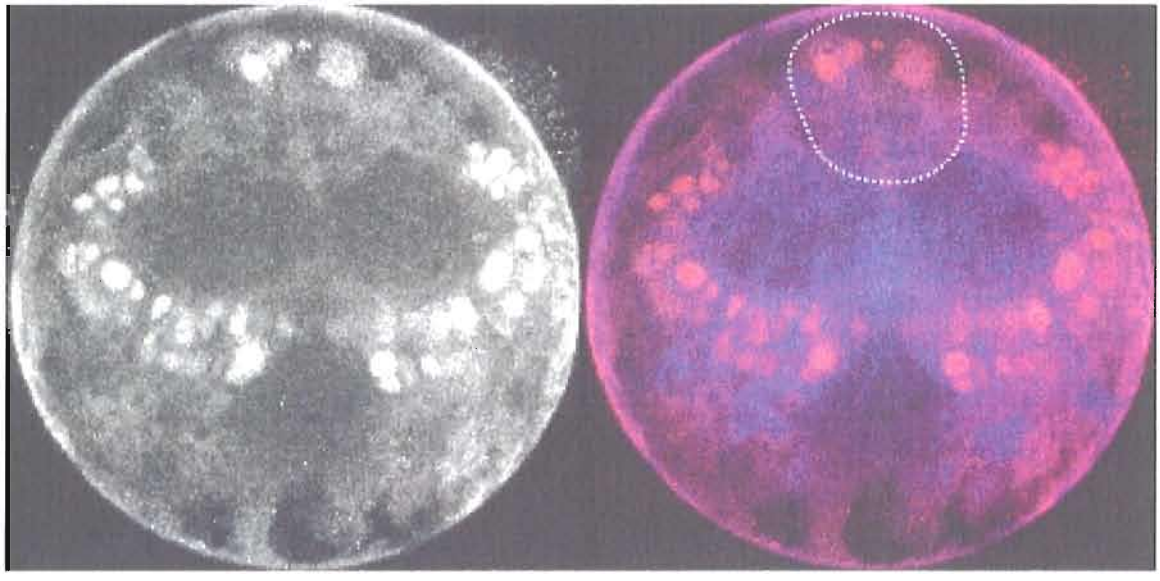

Figure V.9 GATA456 24H
A merged image (using Imaris Extended Field of View option) of several z-planed sections from a whole mount IHC movie of a 24 hour *Platynereis* embryo. PdGATA456 Staining is separated in the left column, and is red in the merged image. Gamma tubulin is in the blue channel. These are taken from the ventral side of the embryo, anterior to top, and the stomodeum is marked with the white dotted line.

mesoblasts, and eventually consist of more then half of the total cells within the two paired bands of developing mesoderm (Figure V.9). We were interested if the proliferation was occurring from the primary mesoblasts, or if there was a contribution from secondary mesoblasts. If these are all derived from the primary mesoblasts, this would mean at least 32 more divisions within a 10 hour period, or a division approximately every 18 minutes. Alternatively, if these are clonally related from the 3 initial GATA456+ cells, this would mean each of the 3 initial 4d derived cell populations has undergone ~ 4 rounds of cell. many of these cells are later derivatives of 4d, as the large primary mesoblasts become less obvious by 18 hours. We can find many examples of dividing secondary mesoblast cells which display high level of cytoplasmic GATA456 protein has accumulated, as well as catching frequent mitosis of the primary

mesoblasts, suggesting that both secondary GATA456+ mesoblasts and the primary mesoblasts are dividing to populate the mesoblastic bands.

Discussion

We have generated polyclonal antibodies against two *Platynereis* GATA transcription factors, *PdGATA123* and *PdGATA456,* and characterized these using immunoblotting and immunohistochemistry.

The first antibody, against *PdGATA123*, detects all non-mitotic nuclei in every stage analyzed. As this appeared drastically different from our previous description of the mRNA expression (Gillis et. al 2007), it is unclear whether this is specifically detecting *PdGATA123*, or is cross reacting with several other antigens. One possibility is that this antibody may be acting as a pan-GATA antibody, as it was created using a portion of the first GATA zinc finger, which is highly conserved on the amino acid across most animal GATAs. However, it is curious that this ubiquitous nuclear staining persists, even after affinity purifying the sera against a smaller GATA123 fragment containing none of the conserved GATA domain.

The staining displayed by the *PdGATA456* antibody appeared to more closely follow our previously described expression for the *PdGATA456* mRNA, but allowed us to conduct a more detailed analysis on the cellular level. We appeared to detect at least six early defined initializations of GATA456+ cells, in both entomeres and micromeres. Based upon our attempts to follow these populations throughout development, and comparing these to previous cell lineage analyses (Wilson 1892, Wilson 1898, Schneider et al. 2002, Schneider and Bowerman 2002, Ackerman et al. 2005), we believe the ectomesodermal population and the body-wall muscle arise from distinct lineages.

We found the earliest GATA456+ nuclear localization appeared to be in the fourth quartet micromeres 4a-4c, whose fate has not yet been determined for *Platynereis,* but which we believe may be the precursors to later ectomesodermal GATA456+ cells, which appear below the anterior rim of the stomodeaum. These cells appear to be in the only region of muscle that does not overlap with 4d staining (Ackerman et al. 2005), suggesting these are not connected to the mesoblast bands. We cannot yet definitively track the fate of the three 4a-c micromeres, but we have seen three GATA456+ cells on the left, right, ventral sides of the blastopore until around ~17 hours. At this point, the closed blastopore is forming it's dumbbell shape, with a ventral thickened portion shifting anteriorly and ventrally, where it will end up just posterior to the equatorial ciliary band referred to as the prototroch. During later stages, a small number of GATA456+ cells appears around the anterior rim of the stomodeaum, which appear to be the ectomesodermal, or non-4d derived muscle. This *Platynereis* ectomesoderm has previously been proposed to originate from the 3a-3c micromeres (Ackerman et al. 2005) based upon cell position. Indeed, in many other spiralians, ectomesoderm has been shown to arise from 3a-b, as well in 2a, 2b, and/or 2c, micromeres in many mollusks (Heijnl et al. 2007). However, this has not been thoroughly examined in annelids, with neither 3a-c or 4a-c micromere fates examined using intracellular fate injections, it is tempting to think that the initial GATA456+ 4a-4c micromeres may actually be the precursors of this later ectomesodermal staining, moved to a more anterior position during epibolic gastrulation movements.

However, 4a-4c have appears to give rise to midgut in other mollusks (Heijnl et al. 2007), and it is possible that they are playing a similar role in *Platynereis* (Schaub

2007) . GATA456 homologs in other most bilaterians play important role in endoderm and mesoderm specification (Patient 2002). mRNA Expression for three GATA456 paralogs have been characterized in another polychaete species, *Capitella Sp I.*, polychaete species, with one localized to the mesoderm, one to the endoderm, and a third with mixed endodermal/mesodermal expression (Bolye and Seaver 2008). It's possible that GATA456+ is transitory, or otherwise difficult to follow, in endodermal 4a-c micromeres. This endodermal role for *PdGATA456* would also appear to correspond with the faint GATA456 nuclear staining we observe in the entomeric nuclei. These nuclei acquire an amorphic shape early in development, and stain poorly with nuclear counterstain, and in fact are most obvious due to their weak staining with *PdGATA456.* One possibility, consistent with lineage tracing showing gut-derived from entomeres, is that the entomeres completely break down and then are resorbed by surrounding micromeres, such as 4a-4c. However, a more detailed time-course analyses or other lineage tracing of third and fourth quartet micromeres will be required to definitively answer these question.

Our analysis also shows the GATA456+ cells occurring throughout the 4d-dervied trunk mesoderm. Our data suggests that at least three different populations of GATA456+ mesodermal precursors exist; a central portion that starts out with cells at the junction between the large 4d1/4d2 mesentoblasts, and cells that appear be lateral to the 4d1 and 4d2 cells. At this point, our knowledge of *Platynereis* cell-lineages only extend to approximately 8 hours past fertilization, and therefore we the exact identity of these cells is still undetermined. However, their relative positions of these cells make it is

unlikely that these are clonally related, and that independent nuclear localization of GATA456 has occurred in three independent cell populations.

The earliest staining in the 4d lineage appears in two medium-sized cells lateral to the primary mesoblasts, which appear to give rise (at least in part) to the paired mesodermal wings. However, we have seen cell proliferation from both from the primary mesoblasts and secondary 4d-derived mesoblasts, suggesting that GATA456 is initialized in additional 4d derived secondary mesoblasts, as well as the possibility GATA456 activation may be clonally inherited in proliferating secondary mesoblasts. However, we see strong GATA456 staining within most of the cells in the 4d derived mesoderm bands throughout development.

In addition, we have followed these mesoblastic bands throughout their development, which reveals a glimpse into the cellular basis for development of *Platynereis* mesoderm, as well as serving as basis for future analyses. One observation is the presence three lateral thickenings of the paired mesodermal bands, each of which appears to initiate with the protrusion of a large, GATA456 positive cell. These thickenings may represent mesodermal segments or somites that occur within the three larval segments (Prud'homme et al. 2003, Steinmartz 2007, Saudemont et al. 2008), and may be some of the earliest precursors of the formation of the three larval segments. Additionally, we hypothesize that the two large bilateral symmetric positioned GATA456+ cells on either side of the proctodaeum might be the remnants of the two primary mesoblasts. These cells appear to be excluded from these segmental blocks, and may be continue to act as mesodermal stem-cell precursors in the future posterior

growth zone generating the mesoderm for newly forming posterior segments added throughout development (de Rosa et al. 2005) .

CHAPTER VI

CONCLUSION

Before this work, it was unclear how the GATA transcription factors have evolved in independent animal lineages, and which functional roles might be conserved. Whereas individual GATA homologs have been suggested to play conserved roles in endomesoderm specification across bilaterian animals, as well in heart and blood formation, in vertebrates, flies, and worms, to date no study has established these conserved roles on the basis of gene orthology across these species. My identification and phylogenetic analyses of the GATA family across the metazoans has provided the evolutionary framework required for reconstructing the ancestral state of this gene family at key nodes of animal evolution. In addition to placing these genes in an evolutionary context, I have conducted the first developmental analyses of GATA orthologs identified in an key intermediate species, the marine annelid *Platynereis dumerilii*, in order to explore the conservation of germ-layer function of GATA factors across vertebrate and invertebrate animals.

In order to reconstruct the relationships of GATA factors across Bilateria, we have identified GATA factors from the polychaete *Platynereis dumerilii*, which represent the first lophotrochozoan GATAs identified. Using a degenerate primer PCR approach, we identified two highly conserved *Platynereis* GATA factors. Contrary to prior predictions of a single ancestral bilaterian GATA (Rehorn et al.1996), our phylogenetic analyses demonstrated that protostomes indeed possessed orthologs to the two

subfamilies previously described among vertebrate GATAs, the GATA –1, -2, -3 class and the GATA –4, -5, -6 class. We observed a distinct germ layer restricted expression of individual *PdGATA* orthologs; *PdGATA123* was expressed in neuroectoderm, while *PdGATA456* was expressed in the mesoderm. Additionally, this analysis enabled us to identify potential GATA123 and GATA456 orthologs between the multiple *Drosophila* and *C. elegans* GATAs. Indeed, comparison of the current literature for germ-layer specific roles for individual GATA homologs across several protostome and deuterostome animals and consistent with our findings in *Platynereis*, GATA456 orthologs appear to be restricted to endomesoderm, mesoderm, or endoderm, whereas the GATA123 orthologs appeared to be more important for neuroectoderm or ectodermal derivatives across Bilateria.

Despite finding putative GATA123 and GATA456 orthologs in several protostome species, many fruitfly and nematode GATAs remained unresolved using the previously annotated protostome GATAs. Furthermore, the *Drosophila Serpent* ortholog, had previously been suggested to be orthologous to both classes of vertebrate GATA, based upon apparent conservation of roles in endoderm specification (GATA4/5/6-like) and also in blood-cell specification (GATA1/2/3-like), suggesting that these arose from a single ancestral bilaterian GATA. To reconstruct the evolutionary origin of these highly derived GATA factors, we used in silico data mining of whole-genome sequence to identify the complete complement of GATA factors from nine recently sequenced protostome animals. Using a combination of molecular phylogenetic analyses, changes in exon/intron boundaries, and chromosomal synteny, I was able to identify the relationship of every *Drosophila* GATA factor. Furthermore, with the exception of a few

fast-evolving lineages (nematodes and leeches), GATA456 genes underwent frequent and apparently independent duplications in lophotrochozoan and ecdysozoan animals, while GATA123 type genes remained single-copy. Interestingly, all arthropods possessed an orthologous single-copy Serpent-like gene, which contained an intact C-terminal zinc finger missing in *Drosophila.* These well-conserved GATAs allowed us to demonstrate that the *Serpent* GATA was indeed a GATA456 like gene, and that perhaps the role of GATA factors in blood formation convergently arose (or was inversely lost) in protostome and deuterostome animals, yet the role in endoderm formation may be ancestral and class-specific.

Previous analyses suggested that the six vertebrate GATA factors arose from two rounds of whole genome duplication (Lowry and Atchley 2000), which suggested the ancestral deuterostome - like the ancestral bilaterian – possessed only two GATA factors. However, this framework was only weakly supported by previous analyses of invertebrate deuterostomes. Although two GATAs had been identified from both an echinoderm (*Strongylocentrotus purpuratus)* (Pancer et al.1999, Hinman and Davidson 2003) and urochordate *(Ciona intestinalis)* (D'Ambrosio et al. 2003, Rothbacher et al. 2007) in earlier studies (citations), these provided only limited support for being members of the GATA123 and GATA456 classes (Gillis et al. 07). In order to support our prediction of two GATA factors in the deuterostome invertebrates, we identified GATA factors from the genomes of two deuterostome invertebrates, the cephalochordate *Branchiostoma floridae* and the hemichordate *Saccoglossus kowalevskii.* In both of these genomes, we identified sole well-conserved orthologs to the vertebrate GATA1/2/3 and GATA4/5/6 genes. We identified nearly exact sets of previously characterized class-

specific motifs in these genes. By characterizing the intron/exon boundaries with the retention of class-specific motifs, we identified a gain of a new splice site within the first-exon of deuterostome GATA1/2/3 genes, splitting this into two smaller exons. Furthermore, we suspected a loss of the first exon missing in the previously described *S. purpuratus* GATA1/2/3 ortholog (*SpGATAe)*, and found conserved sequence of this 'lost' first exon just upstream to the *SpGATAe* locus.

Finally, I returned to investigate the role of the *Platynereis* GATA orthologs via the generation of *Platynereis*-specific antibodies to the individual *PdGATA* orthologs. Although it is unclear if we generated specific anti-*PdGATA123* serum, I was able to obtain a specific anti-*PdGATA456* serum which identified a single protein in western blot of *Platynereis* protein, and appeared to show endomesodermal restricted nuclear protein localization in whole-mount immunohistochemistry. We have examined *PdGATA456* protein localization before and during gastrulation, and examined several populations of cells in which this protein localization first occurs. Similar to the *PdGATA456* mRNA expression, most *PdGATA456* nuclear localization occurs in progeny of the primary mesentoblast 4d. We were only able to detect *PdGATA456* localization in 4d progeny, not until several cell divisions after the initial birth of 4d, in what appear to be three separate 4d descendant populations. In addition to the expected trunk mesoderm staining, we also found *PdGATA456* in the nuclei of the entomeric macromeres, suggest *PdGATA456* plays a previously unidentified role in endoderm. It is possible that this endodermal localization could be a maternally derived protein, which would explain why we failed to see endodermal expression of *PdGATA456* mRNA via in situ hybridization experiments. Additionally, we found *PdGATA456* in the fourth-quartet micromeres (4a-

4c), whose fate has previously remained unknown. Although we have not been able to follow these cells throughout embryonic development, it is possible this cells contributing to non-4d derived ectomesoderm, and indeed *PdGATA456* protein can be seen in nuclei in the anterior of the embryo, in mesoderm surrounding the developing stomodeaum. However, it is apparent that the GATA456 protein localization is much more similar to our current expectation for GATA456 GATAs, with roles throughout the endoderm and mesoderm.

So what can we say about the overall role of GATAs in the evolution of animal germ layers? My data supports that the ancestral bilaterian, the Urbilaterian, possessed at least two GATA factors, while the last-common ancestor of cnidarians and bilaterians, the so-called Urmetazoan, possessed only a singe GATA factor. The expression of the sole GATA in the sea anemone *Nematostella vectensis* is restricted to the inner layer, which some workers refer to as an example of a bifunctional mesendoderm layer, but during later development is expressed in an ectodermal component. A duplication prior to the bilaterian ancestor allowed one GATA to be completely restricted to the formation of the endomesoderm, and the other to the ectoderm, which suggests the retention of two gene-duplicates within a prebilaterian genome through subfunctionalization as hypothesized in the duplication-degeneration-complementation model (Figure VI.1) (Force et al.1999). It appears that GATAs are indeed expressed in the endomesodermal precursors throughout Bilateria, and this early role is even required for the formation of mesoderm tissue in some lineages. Within protostomes, the endomesodermal GATA456 genes have duplicated, and some of these duplicates still display mesoderm-specific roles; for instance, the three GATA456 orthologs in the polychaete *Capitella SpI.* are

expressed only in the endoderm (GATAB-1), mesodermal (GATAB-3), or both (GATAB-2) (Boyle and Seaver 2008). However, there does not appear to be a specific orthologous GATA456 restricted to mesoderm specification throughout bilaterians. In fact, it appears that in some cases in vertebrates (e.g. *Xenopus*), the vertebrate GATA456 genes lack early mesodermal expression, suggesting that these are not playing a direct and initiating role in forming the mesoderm, but exhibit later roles in heart and blood specification. However, one highly speculative possibility is that an ancestral mesoderm is a default-state due to a lack of GATA expression, e.g., trunk cells that lack both GATA456 and GATA123, may be channeled into the mesodermal fate. This can be seen to a certain extent in some of the nematode GATA mutant, where loss of *end-1* and *end-3* cause a transformation of E (endodermal) cell fate to C (ectomesoderm) cell fate (Maduro 2006).

The strongest statement our comparative analysis can make at this point is that the urbilaterian GATA456 gene was playing a role in endomesoderm specification, distinct from an ectodermal GATA123 gene. A key next step will be to examine in detail the roles of GATA factors, and especially GATA456 orthologs, in the endomesodermal gene-regulatory network in a wide number of species. However, we appear to be nearing the answer to many fundamental questions: Is the mesodermal germ layer homologous across Bilateria, or to cnidarian or ctenophores tissues? How conserved are the 'conserved tool-kit' genes and can these be used to determine the homology of germ layers and other developmental characters?

BIBLIOGRAPHY

Aboobaker, A. A., and Blaxter, M. L. 2003. Hox gene loss during dynamic evolution of the nematode cluster. *Curr. Biol.* 13:37-40.

Ackermann, C., Dorresteijn, A., and Fischer, A. 2005. Clonal domains in postlarval *Platynereis dumerilii* (Annelida: Polychaeta). *J. Morphol.* 266:258-280.

Afouda, B., Ciau-Uitz, A., and Patient, R. 2005. GATA4, 5 and 6 mediate TGFbeta maintenance of endodermal gene expression in *Xenopus* embryos. *Development* 132:763-774.

Aguinaldo, A. M., Turbeville, J. M., Linford, L. S., Rivera, M. C., Garey, J. R., Raff, R. A., and Lake, J. A. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* 387:489-493.

Anisimova, M., and Gascuel, O. 2006. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Syst. Biol.*. 55:539 - 552.

Arendt, D., Denes, A. S., Jekely, G., and Tessmar-Raible, K. 2008. The evolution of nervous system centralization. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*. 363:1523-1528.

Arendt, D., Technau, U., and Wittbrodt, J. 2001. Evolution of the bilaterian larval foregut. *Nature*. 409:81-85.

Arendt, D., Tessmar, K., de Campos-Baptista, M. I., Dorresteijn, A., and Wittbrodt, J. 2002. Development of pigment-cup eyes in the polychaete *Platynereis dumerilii* and evolutionary conservation of larval eyes in Bilateria. *Development*. 129:1143-1154.

Arendt, D., Tessmar-Raible, K., Snyman, H., Dorresteijn, A. W., and Wittbrodt, J. 2004. Ciliary photoreceptors with a vertebrate-type opsin in an invertebrate brain. *Science* 306:869-871.

Baguna, J., Martinez, P., Paps, J., and Riutort, M. 2008. Back in time: a new systematic proposal for the Bilateria. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 363:1481-1491.

Bakken, T., and Wilson, R. S. 2005. Phylogeny of nereidids (Polychaeta, Nereididae) with paragnaths. *Zool. Scr.* 34:507-547.

Bartolomaeus, T., and Purschke, G. 2005. Morphology, molecules, evolution and phylogeny in polychaeta and related taxa. in *Dev. Hydrobiol. 179* Springer, Dordrecht, Great Britain, pp 1-356.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C., and Eddy, S. R. 2004. The Pfam protein families database. *Nucl. Acids Res.* 32: D138-141.

Bergter, A., Brubacher, J. L., and Paululat, A. 2008. Muscle formation during embryogenesis of the polychaete *Ophryotrocha diadema* (Dorvilleidae) - new insights into annelid muscle patterns. *Front. Zool.* 5:1.

Boore, J. L. 2001. Complete mitochondrial genome sequence of the polychaete annelid *Platynereis dumerilii. Mol. Biol. Evol.* 18:1413-1416.

Boore, J. L., and Brown, W. M. 2000. Mitochondrial genomes of *Galathealinum*, *Helobdella*, and *Platynereis*: sequence and gene arrangement comparisons indicate that Pogonophora is not a phylum and Annelida and Arthropoda are not sister taxa. *Mol. Biol. Evol.* 17, 87-106.

Boyer, B. C., Henry, J. J., and Martindale, M. Q. 1998. The cell lineage of a polyclad turbellarian embryo reveals close similarity to coelomate spiralians. *Dev. Biol.* 204:111-123.

Boyer, B. C., and Henry, J. Q. 1998. Evolutionary modifications of the Spiralian developmental program. *Am. Zool.* 38:621-633.

Boyer, B. C., Henry, J. Q., and Martindale, M. Q. 1996. Modified spiral cleavage: The duet cleavage pattern and early blastomere fates in the acoel turbellarian Neochildia fusca. *Biol. Bull.* 191:285-286.

Boyle, M. J., and Seaver, E. C. 2008. Developmental expression of foxA and gata genes during gut formation in the polychaete annelid, Capitella sp. I. *Evolution & Development* 10:89-105.

Broitman-Maduro, G., Maduro, M. F., and Rothman, J. H. 2005. The noncanonical binding site of the MED-1 GATA factor defines differentially regulated target genes in the *C. elegans* mesendoderm. *Dev. Cell* 8:427-433.

Brokelmann, J., and Fischer, A. 1966. On the cuticle of *Platynereis dumerilii* (Polychaeta). *Z. Zellforsch Mikrosk. Anat.* 70:131-135.

Burton, P. M. 2008. Insights from diploblasts; the evolution of mesoderm and muscle. *J. Exp. Zoolog. B Mol. Dev. Evol.* 310:5-14.

Casanova, G. 1954. Budding rate of segments during growth and regeneration in the annelid *Platynereis massilienis* (Moquin-Tandon). *C. R. Seances Soc. Biol. Fil.*148, 1446-1448.

Casanova, G., and Coulon-Roso, J. 1967. On the alimentary behavior of *Platynereis massiliensis*. in Moquin-Tandon *C. R. Acad. Sci Hebd. Seances Acad. Sci. D* 264:2152-2153.

Catchen, J., Postlethwait, J., and Conery, J. Teleost Local BLAST Search.

Coroian, C., Broitman-Maduro, G., and Maduro, M. F. 2006. Med-type GATA factors and the evolution of mesendoderm specification in nematodes. *Dev. Biol.* 289:444-55.

D'Ambrosio, P., Fanelli, A., Pischetola, M., and Spagnuolo, A. 2003. Ci-GATAa, a GATA-class gene from the ascidian *Ciona intestinalis*: isolation and developmental expression. *Dev. Dyn.* 226:145-148.

de Rosa, R., Prud'homme, B., and Balavoine, G. 2005. Caudal and even-skipped in the annelid *Platynereis dumerilii* and the ancestry of posterior growth. *Evol. Dev.* 7:574-587.

Dehal, P., and Boore, J. L. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* 3:e314.

Dehal, P. S., and Boore, J. L. 2006. A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database. *BMC Bioinformatics* 7:201.

Denes, A. S., Jekely, G., Steinmetz, P. R., Raible, F., Snyman, H., Prud'homme, B., Ferrier, D. E., Balavoine, G., and Arendt, D. 2007. Molecular architecture of annelid nerve cord supports common origin of nervous system centralization in bilateria. *Cell* 129:277-88.

Dong, Q., Wilkerson, M., and Brendel, V. 2007. Tracembler - software for in-silico chromosome walking in unassembled genomes. *BMC Bioinformatics* 8:151.

Dorresteijn, A. 2005. Cell lineage and gene expression in the development of polychaetes. *Hydrobiologia* 535:1-22.

Dorresteijn, A., and Westheide, W. 1999. The reproductive strategies and developmental patterns in annelids - Preface. *Hydrobiologia* 402:VII-VII.

Dunn, C. W., Hejnol, A., Matus, D. Q., Pang, K., Browne, W. E., Smith, S. A., Seaver, E., Rouse, G. W., Obst, M., Edgecombe, G. D., Sorensen, M. V., Haddock, S. H., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R. M., Wheeler, W. C., Martindale, M. Q., and Giribet, G. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452:745-749.

Fedorov, A., and Fedorova, L. 2006. Where is the difference between the genomes of humans and annelids? *Genome Biol.* 7:203.

Fischer, A. 1977. Autonomy for a specific gene product in oocytes: experimental evidence in the polychaetous annelid, *Platynereis dumerilii*. *Dev. Biol.* 55:46-58.

Fischer, A., and Dorresteijn, A. 2004. The polychaete *Platynereis dumerilii* (Annelida): a laboratory animal with spiralian cleavage, lifelong segment proliferation and a mixed benthic/pelagic life cycle. *Bioessays* 26:314-325.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., and Postlethwait, J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531-1545.

Fritzenwanker, J. H., Saina, M., and Technau, U. 2004. Analysis of forkhead and snail expression reveals epithelial-mesenchymal transitions during embryonic and larval development of *Nematostella vectensis*. *Dev. Biol.* 275:389-402.

Fukushige, T., Hawkins, M. G., and McGhee, J. D. 1998. The GATA-factor elt-2 is essential for formation of the *Caenorhabditis elegans* intestine. *Dev. Biol.* 198:286-302.

Gaunt, M. W., and Miles, M. A. 2002. An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks. *Mol. Biol. Evol.* 19:748-761.

Gilleard, J. S., and McGhee, J. D. 2001. Activation of hypodermal differentiation in the *Caenorhabditis elegans* embryo by GATA transcription factors ELT-1 and ELT-3. *Mol. Cell Biol.* 21:2533-2544.

Gilleard, J. S., Shafi, Y., Barry, J. D., and McGhee, J. D. 1999. ELT-3: A *Caenorhabditis elegans* GATA factor expressed in the embryonic epidermis during morphogenesis. *Dev. Biol.* 208:265-280.

Gillis, W., Bowerman, B., and Schneider, S. 2008. The evolution of protostome GATA factors: Molecular phylogenetics, synteny, and intron/exon structure reveal orthologous relationships. *BMC Evol. Biol* .8:112.

Gillis, W. J., Bowerman, B., and Schneider, S. Q. 2007. Ectoderm- and endomesoderm-specific GATA transcription factors in the marine annelid *Platynereis dumerilli*. *Evol. Dev.* 9:39-50.

Goszczynski, B., and McGhee, J. D. 2005. Reevaluation of the role of the med-1 and med-2 genes in specifying the *Caenorhabditis elegans* endoderm. *Genetics* 171:545-555.

Guindon, S. e. p., and Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52:696-704.

Hagger, J. A., Fisher, A. S., Hill, S. J., Depledge, M. H., and Jha, A. N. 2002. Genotoxic, cytotoxic and ontogenetic effects of tri-n-butyltin on the marine worm, *Platynereis dumerilii* (Polychaeta: Nereidae). *Aquat. Toxicol.* 57:243-255.

He, C., Cheng, H., and Zhou, R. 2007. GATA family of transcription factors of vertebrates: phylogenetics and chromosomal synteny. *J Biosci* 32:1273-1280.

Hejnol, A., and Martindale, M. Q. 2008. Acoel development supports a simple planula-like urbilaterian. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 363:1493-1501.

Hejnol, A., Martindale, M. Q., and Henry, J. Q. 2007. High-resolution fate map of the snail *Crepidula fornicata*: The origins of ciliary bands, nervous system, and muscular elements. *Dev. Biol.* 305:63-76.

Henry, J. J. 2002. Conserved mechanism of dorsoventral axis determination in equal-cleaving spiralians. *Dev. Biol.* 248:343-355.

Henry, J. J., and Martindale, M. Q. 1999. Conservation and innovation in spiralian development. *Hydrobiologia* 402:255-265.

Henry, J. J., and Perry, K. J. 2008. MAPK activation and the specification of the D quadrant in the gastropod mollusc, *Crepidula fornicata*. *Dev. Biol.* 313:181-195.

Henry, J. Q., Hejnol, A., Perry, K. J., and Martindale, M. Q. 2007. Homology of ciliary bands in spiralian trochophores. *Integ. Comp. Biol.* 47:865-871.

Henry, J. Q., Martindale, M. Q., and Boyer, B. C. 2000. The unique developmental program of the acoel flatworm, *Neochildia fusca*. *Dev. Biol.* 220:285-295.

Henry, J. Q., Okusu, A., and Martindale, M. Q. 2004. The cell lineage of the polyplacophoran, *Chaetopleura apiculata*: variation in the spiralian program and implications for molluscan evolution. *Dev. Biol.* 272:145-160.

Henry, J. Q., Perry, K. J., and Martindale, M. Q. 2006a. Cell specification and the role of the polar lobe in the gastropod mollusc *Crepidula fornicata*. *Dev. Biol.* 297:295-307.

Henry, J. Q., Perry, K. J., and Martindale, M. Q. 2006b. Molecular controls of axis specification and cell determination in marine invertebrate embryos and larvae. *Integ. Comp. Biol.* 46:E59-E59.

Hinman, V. F., and Davidson, E. H. 2003. Expression of a gene encoding a Gata transcription factor during embryogenesis of the starfish *Asterina miniata*. *Gene Expr. Patterns* 3:419-422.

Hinman, V. F., Nguyen, A., and Davidson, E. H. 2007. Caught in the evolutionary act: precise cis-regulatory basis of difference in the organization of gene networks of sea stars and sea urchins. *Dev. Biol.* 312:584-595.

Huang, X., and Madan, A. 1999. CAP3: A DNA sequence assembly program. *Genome Res.* 9:868-877.

Huelsenbeck, J. P., and Ronquist, F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754-755.

Hutchinson, T. H., Jha, A. N., and Dixon, D. R. 1995. The polychaete *Platynereis dumerilii* (Audouin and Milne-Edwards): a new species for assessing the hazardous potential of chemicals in the marine environment. *Ecotoxicol. Environ. Sa.* 31:271-281.

Hutchinson, T. H., Jha, A. N., Mackay, J. M., Elliott, B. M., and Dixon, D. R. 1998. Assessment of developmental effects, cytotoxicity and genotoxicity in the marine polychaete (*Platynereis dumerilii*) exposed to disinfected municipal sewage effluent. *Mutat. Res.* 399:97-108.

Ip, Y. T., Maggert, K., and Levine, M. 1994. Uncoupling gastrulation and mesoderm differentiation in the *Drosophila* embryo. *Embo J.* 13:5826-5834.

Ip, Y. T., Park, R. E., Kosman, D., Bier, E., and Levine, M. 1992a. The dorsal gradient morphogen regulates stripes of rhomboid expression in the presumptive neuroectoderm of the *Drosophila* embryo. *Genes Dev.* 6:1728-1739.

Ip, Y. T., Park, R. E., Kosman, D., Yazdanbakhsh, K., and Levine, M. 1992b. dorsal-twist interactions establish snail expression in the presumptive mesoderm of the *Drosophila* embryo. *Genes Dev.* 6:1518-1530.

Jekely, G., and Arendt, D. 2007. Cellular resolution expression profiling using confocal detection of NBT/BCIP precipitate by reflection microscopy. *Biotechniques* 42:751-755.

Jha, A. N., Dominquez, I., Balajee, A. S., Hutchinson, T. H., Dixon, D. R., and Natarajan, A. T. 1995. Localization of a vertebrate telomeric sequence in the chromosomes of two marine worms (phylum Annelida: class polychaeta). *Chromosome Res.* 3:507-508.

Jha, A. N., Hagger, J. A., Hill, S. J., and Depledge, M. H. 2000. Genotoxic, cytotoxic and developmental effects of tributyltin oxide (TBTO): an integrated approach to the evaluation of the relative sensitivities of two marine species. *Mar. Environ. Res.* 50:565-573.

Jha, A. N., Hutchinson, T. H., Mackay, J. M., Elliott, B. M., and Dixon, D. R. 1996. Development of an in vivo genotoxicity assay using the marine worm *Platynereis dumerilii* (Polychaeta: Nereidae). *Mutat. Res.* 359:141-150.

Jha, A. N., Hutchinson, T. H., Mackay, J. M., Elliott, B. M., and Dixons, D. R. 1997. Evaluation of the genotoxicity of municipal sewage effluent using the marine worm *Platynereis dumerilii* (Polychaeta: Nereidae). *Mutat. Res.* 391:179-188.

Kapustin, Y., Souvorov, A., Tatusova, T., and Lipman, D. 2008. Splign: algorithms for computing spliced alignments with identification of paralogs. *Biol. Direct* 3:20.

Karis, A., Pata, I., van Doorninck, J. H., Grosveld, F., de Zeeuw, C. I., de Caprona, D., and Fritzsch, B. 2001. Transcription factor GATA-3 alters pathway selection of olivocochlear neurons and affects morphogenesis of the ear. *J. Comp. Neurol.* 429:615-630.

Kerner, P., Zelada Gonzalez, F., Le Gouar, M., Ledent, V., Arendt, D., and Vervoort, M. 2006. The expression of a hunchback ortholog in the polychaete annelid *Platynereis dumerilii* suggests an ancestral role in mesoderm development and neurogenesis. *Dev. Genes Evol.* 216:821-828.

Kluge, B., Lehmann-Greif, M., and Fischer, A. 1995. Long-lasting exocytosis and massive structural reorganisation in the egg periphery during cortical reaction in *Platynereis dumerilii* (Annelida, Polychaeta). *Zygote* 3:141-156.

Kmiecik, D., Sellos, D., Belaiche, D., and Sautiere, P. 1985. Primary structure of the two variants of a sperm-specific histone H1 from the annelid *Platynereis dumerilii. Eur. J. Biochem.* 150:359-370.

Ko, L. J., and Engel, J. D. 1993. DNA-binding specificities of the GATA transcription factor family. *Mol. Cell Biol.* 13:4011-4022.

Koelle, M. R., and Horvitz, H. R. 1996. EGL-10 regulates G protein signaling in the *C. elegans* nervous system and shares a conserved domain with many mammalian proteins. *Cell* 84:115-125.

Koh, K., and Rothman, J. H. 2001. ELT-5 and ELT-6 are required continuously to regulate epidermal seam cell differentiation and cell fusion in *C. elegans*. *Development* 128:2867-2880.

Kornhauser, J. M., Leonard, M. W., Yamamoto, M., LaVail, J. H., Mayo, K. E., and Engel, J. D. 1994. Temporal and spatial changes in GATA transcription factor expression are coincident with development of the chicken optic tectum. *Brain Res. Mol. Brain Res.* 23:100-110.

Kraus, Y., and Technau, U. 2006. Gastrulation in the sea anemone *Nematostella vectensis* occurs by invagination and immigration: an ultrastructural study. *Dev. Genes Evol.* 216:119-132.

Kulakova, M., Bakalenko, N., Novikova, E., Cook, C. E., Eliseeva, E., Steinmetz, P. R., Kostyuchenko, R. P., Dondua, A., Arendt, D., Akam, M., and Andreeva, T. 2007. Hox gene expression in larval development of the polychaetes Nereis virens and *Platynereis dumerilii* (Annelida, Lophotrochozoa). *Dev. Genes Evol.* 217:39-54.

Lambert, J. D. 2008. Mesoderm in spiralians: the organizer and the 4d cell. *J. Exp. Zoolog. B Mol. Dev. Evol.* 310:15-23.

Lambert, J. D., and Nagy, L. M. 2001. MAPK signaling by the D quadrant embryonic organizer of the mollusc *Ilyanassa obsoleta*. *Development* 128:45-56.

Lambert, J. D., and Nagy, L. M. 2003. The MAPK cascade in equally cleaving spiralian embryos. *Dev. Biol.* 263:231-241.

Lawson, D., Arensburger, P., Atkinson, P., Besansky, N. J., Bruggner, R. V., Butler, R., Campbell, K. S., Christophides, G. K., Christley, S., Dialynas, E., Emmert, D., Hammond, M., Hill, C. A., Kennedy, R. C., Lobo, N. F., MacCallum, M. R., Madey, G., Megy, K., Redmond, S., Russo, S., Severson, D. W., Stinson, E. O., Topalis, P., Zdobnov, E. M., Birney, E., Gelbart, W. M., Kafatos, F. C., Louis, C., and Collins, F. H. 2007. VectorBase: a home for invertebrate vectors of human pathogens. *Nucl. Acids Res.* 35:D503-505.

Lawson, M. A., Whyte, D. B., and Mellon, P. L. 1996. GATA factors are essential for activity of the neuron-specific enhancer of the gonadotropin-releasing hormone gene. *Mol. Cell Biol.* 16:3596-3605.

Leptin, M. 2005. Gastrulation movements: the logic and the nuts and bolts. *Dev. Cell* 8:305-320.

Lowry, J. A., and Atchley, W. R. 2000. Molecular evolution of the GATA family of transcription factors: Conservation within the DNA-binding domain. *J. Mol. Evol.* 50:103-115.

Maduro, M. F. 2006. Endomesoderm specification in *Caenorhabditis elegans* and other nematodes. *Bioessays* 28:1010-1022.

Maduro, M. F., Broitman-Maduro, G., Mengarelli, I., and Rothman, J. H. 2007. Maternal deployment of the embryonic SKN-1-->MED-1,2 cell specification pathway in *C. elegans*. *Dev. Biol.* 301:590-601.

Maduro, M. F., Meneghini, M. D., Bowerman, B., Broitman-Maduro, G., and Rothman, J. H. 2001. Restriction of mesendoderm to a single blastomere by the combined action of SKN-1 and a GSK-3beta homolog is mediated by MED-1 and -2 in C. elegans. *Mol. Cell* 7:475-485.

Maduro, M. F., and Rothman, J. H. 2002. Making worm guts: the gene regulatory network of the *Caenorhabditis elegans* endoderm. *Dev. Biol.* 246:68-85.

Marcellini, S., Technau, U., Smith, J. C., and Lemaire, P. 2003. Evolution of Brachyury proteins: identification of a novel regulatory domain conserved within Bilateria. *Dev. Biol.* 260:352-361.

Martindale, M. Q. 2005. The evolution of metazoan axial properties. *Nat. Rev. Genet.* 6:917-927.

Martindale, M. Q., Finnerty, J. R., and Henry, J. Q. 2002. The Radiata and the evolutionary origins of the bilaterian body plan. *Mol. Phylogenet. Evol.* 24:358-365.

Martindale, M. Q., and Henry, J. J. Q. 1999. The origins of mesoderm. A cell lineage analysis in basal metazoans. *Dev. Biol.* 210:207-207.

Martindale, M. Q., and Henry, J. Q. 1995. Modifications of cell fate specification in equal-cleaving nemertean embryos - alternate patterns of spiralian development. *Development* 121:3175-3185.

McGhee, J. D., Sleumer, M. C., Bilenky, M., Wong, K., McKay, S. J., Goszczynski, B., Tian, H., Krich, N. D., Khattra, J., Holt, R. A., Baillie, D. L., Kohara, Y., Marra, M. A., Jones, S. J., Moerman, D. G., and Robertson, A. G. 2007. The ELT-2 GATA-factor and the global regulation of transcription in the *C. elegans* intestine. *Dev. Biol.* 302:627-645.

Meng, A., Tang, H., Ong, B. A., Farrell, M. J., and Lin, S. 1997. Promoter analysis in living zebrafish embryos identifies a cis-acting motif required for neuronal expression of GATA-2. *Proc. Natl. Acad. Sci.* 94:6267-6272.

Molkentin, J. D. 2000. The Zinc Finger-containing transcription factors GATA-4, -5, and -6. Ubiquitiously expressed regulators of tissue-specific gene expression. *J. Biol. Chem.* 275:38949-38952.

Muller, W. A. 1973. Autoradiographic studies on the synthetic activity of neurosecretory cells in the brain of *Platynereis dumerilii* during sexual development and regeneration. *Z. Zellforsch Mikrosk. Anat.* 139:487-510.

Murakami, R., Okumura, T., and Uchiyama, H. 2005. GATA factors as key regulatory molecules in the development of *Drosophila* endoderm. *Dev. Growth Differ.* 47:581-589.

Nardelli, J., Thiesson, D., Fujiwara, Y., Tsai, F. Y., and Orkin, S. H. 1999. Expression and genetic interaction of transcription factors GATA-2 and GATA-3 during development of the mouse central nervous system. *Dev. Biol.* 210:305-321.

Neave, B., Rodaway, A., Wilson, S. W., Patient, R., and Holder, N. 1995. Expression of zebrafish GATA 3 (gta3) during gastrulation and neurulation suggests a role in the specification of cell fate. *Mech. Dev.* 51:169-182.

Nieto, M. A. 2002. The snail superfamily of zinc-finger transcription factors. *Nat. Rev Mol. Cell Biol.* 3:155-166.

Page, B. D., Zhang, W., Steward, K., Blumenthal, T., and Priess, J. R. 1997. ELT-1, a GATA-like transcription factor, is required for epidermal cell fates in *Caenorhabditis elegans* embryos. *Genes Dev.* 11:1651-1661.

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

Patient, R. M., JD. 2002. The GATA family vertebrates and invertebrates. *Curr. Opin. Genet. Dev.* 12:416-422.

Pattyn, A., Simplicio, N., van Doorninck, J. H., Goridis, C., Guillemot, F., and Brunet, J. F. 2004. Ascl1/Mash1 is required for the development of central serotonergic neurons. *Nat. Neurosci.* 7:589-595.

Peterkin, T., Gibson, A., and Patient, R. 2007. Redundancy and evolution of GATA factor requirements in development of the myocardium. *Dev. Biol.* 311:623-635.

Philippe, H., Brinkmann, H., Martinez, P., Riutort, M., and Baguna, J. 2007. Acoel flatworms are not platyhelminthes: evidence from phylogenomics. *PLoS ONE* 2:e717.

Philippe, H., Lartillot, N., and Brinkmann, H. 2005. Multigene Analyses of Bilaterian Animals Corroborate the Monophyly of *Ecdysozoa, Lophotrochozoa, and Protostomia. Mol. Biol. Evol.* 22:1246-1253.

Podbilewicz, B. 2006. Cell fusion. *WormBook*, 1-32.

Price, A. L., and Patel, N. H. 2008. Investigating divergent mechanisms of mesoderm development in arthropods: the expression of Ph-twist and Ph-mef2 in *Parhyale hawaiensis. J. Exp. Zoolog. B Mol. Dev. Evol.* 310:24-40.

Prud'homme, B., de Rosa, R., Arendt, D., Julien, J. F., Pajaziti, R., Dorresteijn, A. W., Adoutte, A., Wittbrodt, J., and Balavoine, G. 2003. Arthropod-like expression patterns of engrailed and wingless in the annelid *Platynereis dumerilii* suggest a role in segment formation. *Curr. Biol.* 13:1876-1881.

Prud'homme, B., Lartillot, N., Balavoine, G., Adoutte, A., and Vervoort, M. 2002. Phylogenetic analysis of the Wnt gene family. Insights from lophotrochozoan members. *Curr. Biol.* 12:1395.

Raible, F., Tessmar-Raible, K., Osoegawa, K., Wincker, P., Jubin, C., Balavoine, G., Ferrier, D., Benes, V., de Jong, P., Weissenbach, J., Bork, P., and Arendt, D. 2005. Vertebrate-type intron-rich genes in the marine annelid *Platynereis dumerilii. Science* 310:1325-1326.

Rastogi, P. 2000. MacVector. Integrated sequence analysis for the Macintosh. *Methods Mol. Biol.* 132:47-69.

Rebscher, N., Zelada-Gonzalez, F., Banisch, T. U., Raible, F., and Arendt, D. 2007. Vasa unveils a common origin of germ cells and of somatic stem cells from the posterior growth zone in the polychaete *Platynereis dumerilii. Dev. Biol.* 306:599-611.

Rehorn, K. P., Thelen, H., Michelson, A. M., and Reuter, R. 1996. A molecular aspect of hematopoiesis and endoderm development common to vertebrates and *Drosophila*. *Development* 122:4023-4031.

Reuter, R. 1994. The gene serpent has homeotic properties and specifies endoderm versus ectoderm within the *Drosophila* gut. *Development* 120:1123-1135.

Rodaway, A., and Patient, R. 2001. Mesendoderm. an ancient germ layer? *Cell* 105:169-172.

Rodaway, A., Takeda, H., Koshida, S., Broadbent, J., Price, B., Smith, J. C., Patient, R., and Holder, N. 1999. Induction of the mesendoderm in the zebrafish germ ring by yolk cell-derived TGF-beta family signals and discrimination of mesoderm and endoderm by FGF. *Development* 126:3067-3078.

Rothbacher, U., Bertrand, V., Lamy, C., and Lemaire, P. 2007. A combinatorial code of maternal GATA, Ets and beta-catenin-TCF transcription factors specifies and patterns the early ascidian ectoderm. *Development* 134:4023-4032.

Rouse, G. W., and Pleijel, F. 2001. Polychaetes. Oxford University Press, New York.

Ruppert, E. E., Fox, R. S., and Barnes, R. D. 2004. Invertebrate zoology : a functional evolutionary approach. Thomson-Brooks/Cole, Belmont, CA.

Saudemont, A., Dray, N., Hudry, B., Le Gouar, M., Vervoort, M., and Balavoine, G. 2008. Complementary striped expression patterns of NK homeobox genes during segment formation in the annelid *Platynereis*. *Dev. Biol.* 317:430-443.

Schaub, C. 2007. Molekulare Mechanismen der Mesodermbildung in *Platynereis dumerilii* Annelida, Polychaeta. *In* am Fachbereich Biologie und Chemie. Justus-Liebig-Universitat Giessen, Giessen, Germany.

Schneider, S., Fischer, A., and Dorresteijn, A. W. C. 1992. A morphometric comparison of dissimilar early development in sibling species of *Platynereis* Annelida, Polychaeta. *Rouxs Archives of Developmental Biology* 201, 243-256.

Schneider, S. Q., and Bowerman, B. 2007. beta-Catenin asymmetries after all animal/vegetal- oriented cell divisions in *Platynereis dumerilii* embryos mediate binary cell-fate specification. *Dev. Cell* 13:73-86.

Seipel, K., and Schmid, V. 2005. Evolution of striated muscle: Jellyfish and the origin of triploblasty. *Dev. Biol.* 282:14-26.

Seipel, K., and Schmid, V. 2006. Mesodermal anatomies in cnidarian polyps and medusae. *Int. J. Dev. Biol.* 50:589-599.

Sellos, D., Krawetz, S. A., and Dixon, G. H. 1990. Organization and complete nucleotide sequence of the core-histone-gene cluster of the annelid *Platynereis dumerilii*. *Eur. J. Biochem.* 190:21-29.

Sempere, L. F., Martinez, P., Cole, C., Baguna, J., and Peterson, K. J. 2007. Phylogenetic distribution of microRNAs supports the basal position of acoel flatworms and the polyphyly of Platyhelminthes. *Evol. Dev.* 9:409-415.

Sequencing Consortium, T. H. G. 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* 443:931-949.

Shapira, M., Hamlin, B. J., Rong, J., Chen, K., Ronen, M., and Tan, M. W. 2006. A conserved role for a GATA transcription factor in regulating epithelial innate immune responses. *Proc. Natl. Acad. Sci.* 103:14086-14091.

Shim, Y. H. 1999. elt-1, a gene encoding a Caenorhabditis elegans GATA transcription factor, is highly expressed in the germ lines with msp genes as the potential targets. *Mol. Cells* 9:535-541.

Shoichet, S. A., Malik, T. H., Rothman, J. H., and Shivdasani, R. A. 2000. Action of the *Caenorhabditis elegans* GATA factor END-1 in *Xenopus* suggests that similar mechanisms initiate endoderm development in ecdysozoa and vertebrates. *Proc. Natl. Acad. Sci.* 97:4076-4081.

Simionato, E., Kerner, P., Dray, N., Le Gouar, M., Ledent, V., Arendt, D., and Vervoort, M. 2008. atonal- and achaete-scute-related genes in the annelid *Platynereis dumerilii*: insights into the evolution of neural basic-Helix-Loop-Helix genes. *BMC Evol. Biol.* 8:170.

Smith, J. A., McGarr, P., and Gilleard, J. S. 2005. The *Caenorhabditis elegans* GATA factor elt-1 is essential for differentiation and maintenance of hypodermal seam cells and for normal locomotion. *J. Cell Sci.* 118:5709-5719.

Spieth, J., Shim, Y. H., Lea, K., Conrad, R., and Blumenthal, T. 1991. elt-1, an embryonically expressed *Caenorhabditis elegans* gene homologous to the GATA transcription factor family. *Mol. Cell Biol.* 11:4651-4659.

Steinmetz, P. R., Zelada-Gonzales, F., Burgtorf, C., Wittbrodt, J., and Arendt, D. 2007. Polychaete trunk neuroectoderm converges and extends by mediolateral cell intercalation. *Proc. Natl. Acad. Sci.* 104:2727-2732.

Steinmetz, P. R. H. 2006. Comparative molecular and morphogenetic characterisation of larval body regions in the polychaete annelid *Platynereis dumerilii*. in dem Fachbereich der Biologie. Philipps-Universitat Marburg Marburg, Germany.

Steinmetz, P. R. H. 2007. Vasa unveils a common origin of germ cells and of somatic stem cells from the posterior growth zone in the polychaete *Platynereis dumerilii*. *Dev. Biol.* 306:599-611.

Suyama, M., Torrents, D., and Bork, P. 2004. BLAST2GENE: a comprehensive conversion of BLAST output into independent genes and gene fragments. *Bioinformatics* 20:1968-1970.

Sykes, T. G., Rodaway, A. R., Walmsley, M. E., and Patient, R. K. 1998. Suppression of GATA factor activity causes axis duplication in *Xenopus*. *Development* 125:4595-4605.

Tatusova, T. A., and Madden, T. L. 1999. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.* 174:247-250.

Technau, U. 2001. Brachyury, the blastopore and the evolution of the mesoderm. *Bioessays* 23:788-794.

Technau, U., and Scholz, C. B. 2003. Origin and evolution of endoderm and mesoderm. *Int. J. Dev. Biol.* 47:531-539.

Tessmar-Raible, K., and Arendt, D. 2003. Emerging systems: between vertebrates and arthropods, the Lophotrochozoa. *Curr. Opin. Genet. Dev.* 13:331-340.

Tessmar-Raible, K., Raible, F., Christodoulou, F., Guy, K., Rembold, M., Hausen, H., and Arendt, D. 2007. Conserved sensory-neurosecretory cell types in annelid and fish forebrain: insights into hypothalamus evolution. *Cell* 129:1389-1400.

Tessmar-Raible, K., Steinmetz, P. R., Snyman, H., Hassel, M., and Arendt, D. 2005. Fluorescent two-color whole mount in situ hybridization in *Platynereis dumerilii* Polychaeta, Annelida, an emerging marine molecular model for evolution and development. *Biotechniques* 39:460-464.

Tessmar-Raible, K. G. 2004. The evolution of neurosecretory cell types in biliaterian brains. *In* de, Fachbereich Biologie. der Phillips-Universitat Marburg, Marburg, Germany.

Thompson J.D., Higgins D. G., and Gibson, T. J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673-4680.

Tsarovina, K., Pattyn, A., Stubbusch, J., Muller, F., van der Wees, J., Schneider, C., Brunet, J.-F., and Rohrer, H. 2004. Essential role of Gata transcription factors in sympathetic neuron development. *Development* 131:4775-4786.

Tzetlin, A. B., and Filippova, A. V. 2005. Muscular system in polychaetes Annelida. *Hydrobiologia* 535:113-126.

van der Wees, J., van Looij, M. A., de Ruiter, M. M., Elias, H., van der Burg, H., Liem, S. S., Kurek, D., Engel, J. D., Karis, A., van Zanten, B. G., de Zeeuw, C. I., Grosveld, F. G., and van Doorninck, J. H. 2004. Hearing loss following Gata3 haploinsufficiency is caused by cochlear disorder. *Neurobiol. Dis.* 16:169-178.

van Doorninck, J. H., van Der Wees, J., Karis, A., Goedknegt, E., Engel, J. D., Coesmans, M., Rutteman, M., Grosveld, F., and De Zeeuw, C. I. 1999. GATA-3 is involved in the development of serotonergic neurons in the caudal raphe nuclei. *J. Neurosci.* 19:RC12.

Vandenbiggelaar, J. A. M., Kuhtreiber, W. M., Serras, F., Dorresteijn, A., Beekhuizen, H., and Schaap, D. 1986. Analysis of cell communication mechanisms involved in the induction of the stem-cell of the mesodermal bands in embryos of *Patella vulgata* Mollusca. *Acta Histochem.*, 32:29-33.

Velarde, R. A., Sauer, C. D., Walden, K. K., Fahrbach, S. E., and Robertson, H. M. 2005. Pteropsin: a vertebrate-like non-visual opsin expressed in the honey bee brain. *Insect Biochem. Mol. Biol.* 35:1367-1377.

Waltzer, L., Bataillé, L., Peyrefitte, S., and Haenlin, M. 2002. Two isoforms of Serpent containing either one or two GATA zinc fingers have different roles in *Drosophila* haematopoiesis. *EMBO J.* 21:5477-5486.

Weber, H., Symes, C. E., Walmsley, M. E., Rodaway, A. R., and Patient, R. K. 2000. A role for GATA5 in Xenopus endoderm specification. *Development* 127:4345-4360.

Wilson, E. B. 1890. The origin of the mesoblast-bands in annelids. *J. Morphol.* 4:205-219.

Wilson, E. B. 1892. The cell-lineage of *Nereis*. A contribution to the cytogeny of the annelid body. *J. Morphol.* 6:361-480.

Wright, J. M., Wiersma, P. A., and Dixon, G. H. 1987. Use of protein blotting to study the DNA-binding properties of histone H1 and H1 variants. *Eur. J. Biochem.* 168:281-285.

Yin, Z., and Frasch, M. 1998. Regulation and function of tinman during dorsal mesoderm induction and heart specification in *Drosophila*. *Dev. Genet.* 22:187-200.

Yin, Z., Xu, X. L., and Frasch, M. 1997. Regulation of the twist target gene tinman by modular cis-regulatory elements during early mesoderm development. *Development* 124:4971-4982.

Zeeck, E., Hardege, J. D., Willig, A., Ikekawa, N., and Fujimoto, Y. 1994. Sex pheromones in marine polychaetes: steroids from ripe *Nereis succinea*. *Steroids* 59:341-344.

Zhu, J., Hill, R. J., Heid, P. J., Fukuyama, M., Sugimoto, A., Priess, J. R., and Rothman, J. H. 1997. end-1 encodes an apparent GATA factor that specifies the endoderm precursor in *Caenorhabditis elegans* embryos. *Genes Dev.* 11:2883-2896.