UNIVERSITY OF CALIFORNIA, SAN DIEGO

On the evolution of the *sghC1q* gene family, with bioinformatic and transcriptional case studies in zebrafish

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Marine Biology

by

Tristan Matthew Carland

Committee in charge:

Lena G. Gerwick, Chair
Eric E. Allen
Phillip A. Hastings
Victor Nizet
Victor D. Vacquier

2011

The dissertation of Tristan Matthew Carland is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2011

DEDICATION

To my family and friends

To my parents

To my fianceé

You are the makings of this life

I can never thank you enough

EPIGRAPH

Computers are incredibly fast, accurate, and stupid.
Human beings are incredibly slow, inaccurate, and brilliant.
Together they are powerful beyond imagination.
— A. Einstein

All we have to decide is what to do with the time that is given to us.
— J.R.R. Tolkein

Enlightenment is man's release from his self-incurred tutelage. Tutelage is man's inability to make use of his understanding without direction from another. Self-incurred is this tutelage when its cause lies not in lack of reason but in lack of resolution and courage to use it without direction from another.
Sapere aude! *"Have courage to use your own reason!"*
— I. Kant

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

In giving thanks I cannot possibly do so without first thanking my advisor, Dr. Lena Gerwick. Few students can claim that their advisor taught them the basics of any topic in a one-on-one, hands-on manner. Her tutelage is what has enabled me to become the researcher that I am today, she has been an academic matriarch and open-minded scientific critic beyond the call of duty. She has always met the challenge of multi-disciplinary research (particularly in my case) with patience, understanding and enthusiasm.

My long academic career is greatly thanks to the constant support of my Mother and Father, Kathleen and Ronald Carland. They have always encouraged my interests in science, and shaped my mind with a home of open-minded logical debate. Thanks also to the Rivero family and particularly my fianceé Carmen for giving me a loving home away from home, full of cats and spicy food. Thanks to my grandfather Dr. Daniel Walsh for helping to inspire my earliest interests in science, my sister Miranda for visiting me in San Diego and letting me sleep in her basement, and my brother Benjamin for the lifetime of support that only a brother could share. Thanks also to the Gerwick laboratory, for so very many things.

My entry into science is thanks in great part to Dr. Jon Norenburg. Had he not checked his email and given the chance to a random intern, my history would be quite different. Working with him at the Smithsonian was an even mix of direction and wonder, being put to work and loosed upon "the Nation's Attic". As if that wasn't enough, he sent me to Dr. Rachel Collin who gave me yet another dream Summer. Dr. Collin was a friendly advisor who set me to work on a biodiversity database and let me stay at a tropical research station in Panama.

Thanks are owed to the Gridnexus team of the University of North Carolina Wilmington, my first university research experience. The leadership of Dr. Ronald Vetter and infective interest of Dr. Jeffrey Brown could only be trumped by the sharpness of Dr. Ann E. Stapleton. She (a botanist) took me into her lab and gave me invaluable guidance and research opportunity. The many fine teaching computer science professors at UNCW, especially Dr. Clayton Ferner, and Dr. David R. Berman are the foundation of my programming abilities. Thanks also

Dr. Thomas E. Lankford for the ichthyology experience I had always wanted.

My doctoral committee of Dr. Victor D. Vacquier, Dr. Eric E. Allen, Dr. Victor Nizet, and Dr. Phillip A. Hastings have been incredibly supportive of my time here; especially Dr. Nizet who shared a grant that funded my research for a year and Dr. Vacquier who has been very involved in helping me to find my path at SIO and beyond. I am very thankful to Dr. Vacquier for his career advice and that of: Dr. Jeffrey Graham who leveled with me when I needed it, Dr. Terry Gaasterland who got me started at SIO and showed me the gravity of bioinformatics, Dr. Shiela Podell the goddess of perl, Dr. Daniel Udwary the master of chicken, Dr. Francisco Villa the big-brother, Dr. William Gerwick who would always entertain my crazy ideas, and Eddie Kisfaludy who made my nautical dreams come true. Thanks also to Dr. Jules Jaffe for indulging my video gaming interests for a good cause, knowing how to handle me, and being a good influence.

Thanks to the Socrates team of Shelley Glenn, Johnnie Lyman, and Maarten Chrispeels for supporting my adventures as a scientist at a high school. Great thanks to Thomas McElfresh for helping me to survive the adventure; and thanks Carmen Velez, Zephen Specht, Jillian Blatti, and the rest of the Socrates 2010-2011 cohort for their comraderie. Special thanks to the wonderful crew of the CAL-ECHOES cruise for the best work-nights of my SIO career.

Last but certainly not least, no graduate student is an island; my friends/peers/cohorts/inlaws are what keep this ship floating. Thanks to old guard Tom Loescher, Jon Miller, Trey Canter, Bryson Osborne, Adam Duncan, Daniel Conley, Bill Shipman, and Kristen Pelick. Renowned office mates Dr. Trina Norden-Krichmar and Emily Trentacoste have been daily sources of support; as have the temporary office mates Dr. Irma Mercado and Amrit Sareen; and helpers Monica Brunneto and Shumpei Maruyama. Friends and cohorts Dr. Jeffrey B. Locke, Danny Richter, Cameron Coates, Niclas Engene, Ty Samo, Juan Ugalde, Samantha Mascuch, Emiley Eloe, Dr. Emily Monroe, Josh Wingerd, and Mike Wilson (among many others) have been indespensable in this journey. Final thanks to the personable SIO Graduate Department, the unbeatable Ludicrous Gibs, and the unsung heroes of the TG Committee.

Chapter 2 is a full reprint of the publication: Carland, T. M. and Gerwick, L. (2010). The C1q domain containing proteins: Where do they come from and what do they do? Developmental and Comparative Immunology, 34(8):785-790, with permission from all coathors.

Chapter 3 is a full reprint of the publication: Carland, T. M., Locke, J. B., Nizet, V., and Gerwick, L. (2011). Differential expression and intrachromosomal evolution of the sghC1q genes in zebrafish (Danio rerio). Developmental and Comparative Immunology, (In press), with permission from all coathors.

Chapter 4 is a modified version of a manuscript in preparation for submission under the title "EST Keeper: a Flash based web-tool for extracting complete and non-redundant ORFs from BLAST alignment sequence hits" with permission from coathor Dr. Lena Gerwick.

Appendix C is a high school biology activity that was developed as a product of the Socrates Fellows program at UC San Diego, a project of ScienceBridge, supported by funds from the National Science Foundation GK12 STEM Fellows in Education, awarded to Maarten Chrispeels, Division of Biological Sciences. An updated version with all materials is available at sciencebridge.ucsd.edu.

VITA

2011 - Ph.D.	Doctor of Philosophy, Marine Biology Scripps Institution of Oceanography, University of California, San Diego
2009 - M.S.	Master of Science, Marine Biology Scripps Institution of Oceanography, University of California, San Diego
2005 - B.S.	Bachelor of Science, Computer Science University of North Carolina, Wilmington
2005 - B.S.	Bachelor of Science, Marine Biology University of North Carolina, Wilmington

PUBLICATIONS

Tristan M. Carland, Jeffrey B. Locke, Victor Nizet, Lena Gerwick. Differential expression and intrachromosomal evolution of the (*sghC1q*) genes in zebrafish (*Danio rerio*). *Developmental & Comparative Immunology*, doi: 10.1016/j.dci.2011.05.013

Tristan M. Carland, Lena Gerwick, EST Keeper: a Flash based web-tool for extracting complete and non-redundant ORFs from BLAST alignment sequence hits. (In Preparation)

Tristan M. Carland, Lena Gerwick, The C1q domain containing proteins: Where do they come from and what do they do?, *Developmental & Comparative Immunology*, Volume 34, Issue 8, August 2010, Pages 785-790.

Jeffrey L. Brown, Clayton S. Ferner, Thomas C. Hudson, Ann E. Stapleton, Ronald J. Vetter, Tristan Carland, Andrew Martin, Jerry Martin, Allen Rawls, William J. Shipman, and Michael Wood. GridNexus: A Grid Services Scientific Workflow System. *International Journal of Computer Information Science (IJCIS)*, Volume 6, No 2, June 2005, Pages 72-82.

ABSTRACT OF THE DISSERTATION

On the evolution of the *sghC1q* gene family, with bioinformatic and transcriptional case studies in zebrafish

by

Tristan Matthew Carland

Doctor of Philosophy in Marine Biology

University of California, San Diego, 2011

Lena G. Gerwick, Chair

In this thesis the evolution of the sghC1q gene family is explored throughout the metazoan lineage and within the zebrafish (*Danio rerio*) genome. This involved novel bioinformatic analyses, extensive synthesis of the literature, development of a bioinformatic tool, and the transcriptional assessment of the full complement of *sghC1q* genes within *D. rerio* during infection and early development.

The secreted globular head C1q (*sghC1q*) genes can be characterized as a family of genetic loci each encoding a signal peptide followed by a complement component 1q globular (gC1q) motif. Members of this family have been referred to as precerebellin-like (cblnl), C1q-like or ovary specific C1q-like factors. Previous studies have found gene family members in multiple organisms with varying num-

bers of copies within a species. The genes are known to be transcribed in response to infection and/or during development.

The domain of the C1q globular head (gC1q or ghC1q) appears to be ancient; present even in prokaryotes. With increasing complexity of organisms, this domain can sometimes be found accompanying first a signal peptide motif (indicative of secretion), and later with a collagen region. A comprehensive naming scheme is suggested based on these evolutionary adaptations. Computational modeling shows the globular head to be structurally conserved throughout the metazoa.

The EST Keeper program was developed to facilitate these studies in identification of sets of non-redundant homologous genes from BLAST results that often contain redundant copies and gene fragments. It was built as a Flash based webservice and can be used to find gene families within genomes and EST datasets.

Twenty *sghC1q* genes were found in the zebrafish (*D. rerio*) genome (Zv9) and transcriptionally assessed. Two of the examined twenty genes showed significant up-regulation within 24 h of infection with the fish pathogen *Streptococcus iniae*, and eleven were expressed during early development. Due to the clustered nature of these genes on chromosomes two and seven, intrachromosomal duplication events are hypothesized and explored.

Chapter 1

Preface

1.1 Introduction

The immune system is the key to the relationship between pathogens and hosts. It has been defined as the system that prevents disease but is known to also take part in other processes such as neuronal development [Stevens et al., 2007] and clearance of damaged tissues [Murphy et al., 2008], fitting with the proposed "Danger Model" of immunological response stating that the immune system has evolved to handle threats to the organism [Matzinger, 2002]. The immune system is the means by which organisms identify and manage anything that is not "self". Identifying what is "self" and "non-self" is a complex problem given how quickly potential pathogens can evolve, deciding and managing the ensuing response is yet another complex problem, especially when some "non-self" organisms can be very helpful.

Unicellular microorganisms (microbes) are everywhere, covering most every last thing on this planet. Every organism, from plants, to insects, to other microbes, and especially vertebrates, are covered (inside and out) with microbes. We used to think of microbes as pathogens, a term that refers to any microbe that can cause disease. Microbes can be pathogenic, but most will never come into contact with a vertebrate, and many can be helpful. Every milliliter of seawater can contain millions of bacteria and an order of magnitude more viruses [Hobbie et al., 1977] [Wommack et al., 1999], from these communities our planet

gets half of its oxygen supply [Falkowski et al., 2000]. The human body contains more bacterial cells than human cells (estimated ten times more), particularly in the gut, most of which are useful symbionts [Berg, 1996]. Be they symbiotic friend or pathogenic foe, microbes are everywhere, evolving far more rapidly than we do, and our immune system has evolved to interact with them.

Approximately 450-500 million years ago, our ancestors, the fishes (fish*es* denotes multiple species of fish [Nelson, 2006]) underwent a massive radiation of lineages that would eventually lead to the tetrapod lineage that would someday forsake the ocean [Clack, 2002]. This period coincides with many events including: our ancestors' evolution of the jaw [Ellis, 2001], our ancestors' two genome duplication events [Holland and Garcia-Fernndez, 1996] [Postlethwait et al., 1998], a period of massive global cooling and extinction [Finnegan et al., 2011], and our ancestors' evolution of the antibodied adaptive immune system [Litman et al., 2010]. There are two primary systems of immunity among organisms (detailed below), one of which all organisms have (even bacteria and plants), and the other is possessed only among the descendants of that radiation - the jawed vertebrates [Flajnik and Kasahara, 2001, Pancer and Cooper, 2006]. Fishes present a particularly interesting area of study for immunology because they are still jawed vertebrates possessing the same fundamental immune components as mammals yet also represent the greatest diversity of the vertebrates, allowing us to learn a great deal about the history, evolution, and alternative possibilities of the immune system.

1.1.1 Fish(es) as model organisms

Proper experimental design dictates that in an experiment, only one variable should be changed at a time and all else should remain equal between groups. Model organisms are those that are amenable to experimentation such that they can be manipulated in experimental settings and what is learned from them could hold true for other organisms as well. In the case of an immunological model, a field generally dominated by concerns for human health, we would hope that the model organism also be close enough to humans (evolutionarily) for some of the results to translate back and that it be more convenient to work with than humans.

Zebrafish have these traits as well as others (noted below) that make it relatively ideal over other model organisms for certain types of experiments. As compared to some other model organisms (mouse, rat, fruitfly) the zebrafish genome has been fully sequenced (project 2001-2004), and in its ninth assembly (Zv9 - April 2010) has reached a reasonable quality.

The zebrafish was originally chosen for methods development by Dr. George Streisinger over other fishes (medaka, puffer, goldfish); zebrafish can be easily bred year-round within the laboratory; a female can lay hundreds of eggs per week; eggs are fertilized externally; haploid development of embryos until the larval stage; and transparent embryos develop externally to a swimming larvae within 48 hours. In addition, a new generation of zebrafish can be ready for breeding in just a few months and several zebrafish can be kept in a single liter of tank water. Furthermore, advanced tank systems have been engineered that require relatively little maintenance compared to what is needed for rodents. [Grunwald and Eisen, 2002]

Many protocols have been developed for the use of zebrafish in genetic research, further perpetuating their popularity as a model organisms. Random mutagenesis experiments have been popular in zebrafish since its inception; Christiane Nusslein-Volhard earned the Nobel Prize in 1995 for her contributions to developmental biology, largely using EMU zebrafish mutants. The techniques employed are either random insertion mutagenesis via virus [Chen et al., 2002] or chemical (EMS, ENU, or TMP) (reviewed in [Sullivan and Kim, 2008]). Mutagenesis experiments enabled research and discovery of the genes controlling phenotypes years before any significant amount of the zebrafish genome was deciphered.

Increased knowledge of the zebrafish genome has greatly enabled reverse-genetics, where genes are chosen to be targeted to investigate possible phenotypic changes. Morpholinos enable the temporary interruption (or "knock-down") of the transcription of a targeted gene. While the externally developing zebrafish embryos are at the single cell stage one can inject these specific sequences (up to 25 bp) as a means of interrupting the translation (or splicing) of a gene by anti-sense complementation [Bill et al., 2009]. A recent method to stop the function of a gene in a heritable fashion (in this case "knock-out") is through the use of engineered

zinc-fingered nucleases (ZFNs) that cause double stranded breaks in the desired DNA, effectively acting as a point mutation in a specified gene [Meng et al., 2008].

The *Tol2* transposase system allows insertion of DNA (a "knock-in" of up to 11 kb) by means of injection of a plasmid containing the desired DNA and a transposase gene isolated from the Japanese rice fish, medaka (*Oryzia latipes*) [Kawakami et al., 1998, Kawakami and Shima, 1999]. This method has been used to successfully insert genes that are heritable and in some cases with organ specific promoter regions [Kawakami, 2007].

1.1.2 Practical applications

As the world population increases, so does our demand for fish as a food source. Unfortunately, almost every natural population of fish is in drastic decline. Aquaculture could hold a solution (to the food shortage), and the leading problem with aquaculture is death from infection [Agnew and Barnes, 2007, Gauthier and Rhodes, 2009]. Generally fish farms grow many individuals of the same species in pens at very high densities leading to stress, damage, and a relatively ideal environment for opportunistic pathogens [Ford and Myers, 2008]. A great deal of time and money is being spent on vaccinations, antibiotics, and probiotics [Burr et al., 2005, Nayak, 2010]. Wild populations of fishes are also of great concern as a food source and as indicators of the health of an eco-system. Knowledge of their immune systems such that we might monitor their health more closely is of great importance to our economy and environment.

1.2 Comparative immunology

1.2.1 An evolutionary context

The progenitor of the most rudimentary phagocytic cell type, the macrophage, was first discovered by Metchnikoff while he was observing a myriad of invertebrate species in search of his newfound cell type that appeared to have dual functions in digestion and host defense. His "eureka" moment came during an

experiment in 1882 where he had "introduced" a wooden splinter into the transparent body of a sea-star larva and his trademark cells began surrounding the splinter [Tauber, 2003]. For his extensive work with phagocytic cells, development of the "phagocytosis theory", and his remarkable contributions to evolutionary biology (book - [Metchnikoff, 1905]), he was awarded the Nobel prize for medicine in 1908 [Schmalstieg and Goldman, 2008].

Metchnikoff's phagocytic cells, macrophages, are key immune effector cells in invertebrates. The immune systems of invertebrates are dominated by what is termed the "innate" or "non-specific" immune system, implying that the cells and proteins of the system are not fine-tuned to recognize pathogens with increasing precision but instead rely on the innate capabilities of the organism. This notion only really makes sense when one knows of the "adaptive" or "specific" immune system, which is so termed because of its ability to specifically adapt to new pathogens and functionally "remember" them to expedite resolution of future encounters. [Murphy et al., 2008]

The innate immune system is found in all organisms, including vertebrates which also possess an adaptive immune system. Through numerous studies, the adaptive immune system has been hypothesized to have emerged in vertebrates, shortly before the evolution of the jaw. Jawless vertebrates (agnathans: lamprey and hagfish) possess an adaptive immune system different from the rest of the vertebrates; it contains cells similar to those we see in the jawed vertebrates but with novel type of leucine rich repeat (LRR) receptors that function very similarly to the immunoglobulin/antibody receptors of the jawed vertebrates; a likely case of convergent evolution [Pancer and Cooper, 2006] [Lieschke and Trede, 2009]. This notion is being reconsidered in light of the recently sequenced amphioxus genome (a pre-agnathan) that shows evidence of domain shuffling among innate immune receptors that could lead to an increase in their binding repertoire of ligands. [Zhang et al., 2008b] [Litman et al., 2010]

In mammals, most immune cells develop within the bone marrow and then circulate in the blood system, continue their development in primary lymph nodes such as the thymus, or to reside in secondary lymph nodes until needed. A key

difference between fishes and mammals that affects both immune systems is that fishes lack bone marrow and appear to lack dedicated lymph nodes, thus requiring other organs such as their liver, kidney, epigonal or leydig organs to produce the immune cells and in some cases house them. Another obvious difference between land mammals and fishes is that fishes live in water, meaning that they are exposed to a constantly changing environment with nearly one million bacteria per milliliter of water. Instead of skin they rely on scales (in some cases) and the slime that coats them contains immune factors detailed below. [Lieschke and Trede, 2009] [Zaccone, 2009]

1.2.2 Adaptive immunity

The principle of the adaptive immune system; indeed, the reason it is called adaptive, is the ability of the non-self sensing receptors of this system to be actively generated into a near infinite range of possibilities during the single lifetime of an organism. This is possible thanks largely to the recombination activating genes (RAG) that can reorder the genetic material used to create these recognition proteins into new combinations prior to translation. Once the specific recognition protein complexes for a particular pathogen have been created (and they become increasingly accurate over time), the time needed to resolve the infection shortens drastically. The main protein-complexes generated by this process are the immunoglobulins(Ig)/antibodies and the T-cell receptors (TCR). Fishes do create these protein-complexes with a key difference being the lack of what is called class-switching, the ability to change one class to another. For example, immunoglobulins/antibodies are given a single letter designation (IgA,IgD,IgE,IgG,IgM in mammals) and in mammals the isotype being produced by a particular cell can be changed, which does not appear to be the case in fishes. Additionally, teleost fishes have a differing set of immunoglobulins/antibodies (IgD,IgM,IgT) with the new type (T for teleost) that is found primarily in the gut. [Zhang et al., 2011]

1.2.3 Innate immunity

The innate immune system could be most abstractly defined as any part of an organism that may hamper the success of a pathogen and is not part of the adaptive immune system. This can include a great many things such as physical barriers (skin), antimicrobial peptides found on the skin (also in sweat and tears), patrolling cells within tissues and blood (macrophages, neutrophils, dendritic cells, mast cells, etc), and unbound proteins circulating within the blood (complement components, pentraxins, lectins, etc). A more functional definition for the context of this thesis is that the innate immune system is an interacting system of cells and proteins that uses Pathogen Recognizing Receptors (PRRs) to target Pathogen Associated Molecular Patterns (PAMPs) on broad classes of pathogens. That is the terminology, the specific difference is that while the adaptive immune system can recombine the DNA of its receptors to specifically bind to a particular ligand, the innate immune system is presently known to rely on germ-line encoded DNA (no recombination) to make proteins that will less-specifically bind to conserved molecular patterns (PAMPs) found on broader classes of pathogens (it is also called the non-specific immune system).

Another important difference is the amount of time needed to enact an adaptive response (generally one week in humans without prior exposure, as many as five weeks in fishes) is much greater than is required for an innate response (nearly instantaneous). Indeed, inflammation and the acute phase response (see Chapters 2 and 3) can garner an effective response in a matter of hours. Most pathogens never get past the innate immune system, those that do may simply be combatted by the innate system until T cells and B cells can more specifically resolve the situation. Pathogens that frequently do overcome the innate system become well known as they cause death or a great deal of discomfort until the adaptive immune system can lead to their clearance.

The interplay between the two systems is substantial. Originally the innate response was considered to be a secondary response [Janeway, 1989], this is because recognition complexes of the adaptive immune system are known to illicit pathways of the innate immune system that will lead to clearance of the pathogen. This

Table 1.1: Copy-count of Toll-like receptors in selected mammal and teleost species, data adapted from [Kawai and Akira, 2010] [Palti, 2011] [Kasamatsu et al., 2010]. Across the top is the Toll-like receptor (TLR) designation, subsequent rows are the number of copies of that gene found within the organisms' genome, and the bottom row is whether the expressed receptor is found on the external surface of the cell (denoted by S), or internally on the membrane of a vesicle (denoted by V).

TLR	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Pufferfish	1	1	1		2		1	1	1					1							1	1	1
Zebrafish	1	1	1	2	2			2	1			1		1				1	1	1	1	1	
Frog	1	2	1		1	2	1	2	1			1		4							1	1	
Chicken	1	1	1	1	1	1	1								1						1		
Mice	1	1	1	1	1	1	1	1	1		1	1	1										
Humans	1	1	1	1	1	1	1	1	1	1													
Surface/Vesicle	S	S	V	S	S	S	V	V	V	S	S												

process is how the adaptive system clears many pathogens but we now know that the innate immune system does not require the adaptive to become activated [Medzhitov, 2009]. Recent studies are even beginning to show that the innate immune system may possess a great deal more diversity in its ability to recognize pathogens than ever considered [Litman et al., 2010].

During the first four days of development, a zebrafish embryo will not express any genes of the adaptive immune system [Traver et al., 2003], allowing study of the innate system without any interaction or intervention by the adaptive system. This makes the developing zebrafish a specialized model of the vertebrate immune system, perhaps enhancing studies of the immediate innate immune response (i.e. inflammation) as the developing fish is completely reliant on the innate immune system (and any maternally transferred materials [Swain and Nayak, 2009]) to defend itself.

The Toll-like receptors (TLRs) are pathogen recognition receptors that have a remarkable diversity in their binding ability to PAMPs, particularly in fishes though they are well conserved among organisms. They were so named because they greatly resemble the Toll receptors of fruitflies that are important developmental regulators [Anderson et al., 1985] and take part in defense against bacterial and fungal pathogens [Lemaitre et al., 1996]. Most of them (TLR1,2,4,5,6,11) have a transmembrane domain with extracellular Leucine Rich Repeats (LRRs) to recog-

nize PAMPs and Toll/IL-1 receptor (TIR) domains for intracellular signalling (often through a MyD88 dependent pathway). Others TLRs (TLR3,7,8,9) act in similar manners but are found upon internal membranes of vesicles such as endosomes or phagosomes. Additional TLRs have been identified in fishes, some of which are fish specific and may have multiple copies within the genome (Table 1.1). TLR4 is a special exception as it has only been identified in zebrafish and the two copies in zebrafish (TLR4a,b) appear to be paralogs that do not bind to lipopolysaccharides (LPS) like their mammalian counterparts [Sullivan et al., 2009].

1.2.4 Complement

The complement system (or complement cascade) is another example of a well conserved and highly diversified facet of the innate immune system, and serves as a pivotal link to the adaptive immune system (reviews [Kishore and Reid, 2000], [Kishore et al., 2004], [Nayak et al., 2010]). The complement cascade is an orchestrated cascade of pro-proteins (inactive precursors) that bind, are proteolytically cleaved (thus activating them), and enzymatically lead to the next stage in the pathway. The proteins are referred to with the letter C and then a number (C1-C9) signifying the proteins' stage in the classical pathway. The complement cascade is known to have three pathways of activation. The first pathway, and the most pertinent to this thesis, is the classical pathway. Complement component 1 (C1=C1q+C1s+C1r), particularly the binding portion of it (termed C1q) is known to bind to a diverse set of ligands (detailed below and by [Kishore et al., 2004]) including antibodies. The ability of C1q to bind to antibodies and functionally "complement" the antibody response is the reason for the name.

Once C1q has bound to something, it can lead to the cleavage/activation of C2 and C4, which in turn leads to the cleavage/activation of C3. The lectin pathway begins with the binding of Mannose Binding Lectin (MBL) binding to mannose on the surface of a pathogen and also leads to cleavage/activation of C2 and C4 (which can activate C3). Finally, the alternative pathway is in a sense the auto-activation of C3 (also possible with other factors). At this point the three pathways have largely converged and can lead to the Membrane Attack Complex

(MAC), where proteins C5-C9 continue to form a pore in the membrane of the target cell, leading to its destruction.

Additionally, and of greater note to this thesis on the genes/proteins similar to the secreted globular head of C1q (sghC1q), when the initiating proteins of these pathways (particularly C1q and C3) are deposited on a cell surface they can act as opsonins. Opsonins are effectively "immune tags", marking a ligand (such as a pathogen) for additional processing by the immune system. Receptors to C1q and C3 are found on phagocytic cells and their activation can lead to phagocytosis of the opsonized cells as well as other formative processes of innate and adaptive immunity [Hosszu et al., 2010]. The C1q protein complex is a hexamer of heterotrimers (total of 18 peptides) in which each heterotrimer is a globular head of C1q A, B and C chains held together by coils of collagen fibers (Figure 2.1). There are separate receptors for both the the globular head domain (gC1q or ghC1q binds to gC1q-R) and the collagen tail domains (cC1q binds to cC1q-R); each conferring specific responses though generally being able to induce phagocytosis, chemotaxis, and monocyte differentiation. These receptors are not thought to be directly attached to cells as they lack transmembrane domains but they are known to dock to other receptors (such as β1 integrin and CD91); thus, acting as intermediaries between cell surface receptors and the collagenous or globular head domains of C1q (and likely C1q domain containing proteins). [Ghebrehiwet et al., 2001] [Ghebrehiwet and Peerschke, 2004] [Vegh et al., 2006] [Peerschke and Ghebrehiwet, 2007]

All three pathways of complement system activation are present in teleosts [Boshra et al., 2006]. As with the TLRs, there appears to be a great deal of diversity among complement components, with varying numbers of homologs for each of the components (particularly C1) [Hu et al., 2010] [Nakao et al., 2011]. Teleosts are the most diverse group of vertebrates, owed to whole genome duplications and intrachromosomal gene duplications, they are known to have evolved more rapidly than the rest of the vertebrates [Ravi and Venkatesh, 2008]. Given this and the immunologically challenging environment in which they live, we should not be surprised that the innate immune systems of teleosts have evolved to be extremely

diverse.

1.3 The sghC1q family

The central theme of this thesis is the study of the evolution of the secreted globular head C1q (sghC1q) gene family. This gene family has been found to be much larger than previously thought, both throughout the Metazoan lineage (as defined in [Schierwater et al., 2009]) and within the zebrafish genome. The metazoan evolution of this gene family and its closely related genes are discussed in Chapter 2. The differential transcription of these genes during an infection and during development, as well as the evolution and radiative expansion of these genes within zebrafish is discussed in Chapter 3. The study of this gene family necessitated the development of a bioinformatic tool; presented in Chapter 4. The findings of these studies are summarized along with future directions and broader impacts in Chapter 5.

Chapter 2

The C1q domain containing proteins: Where do they come from and what do they do?

2.1 Abstract

The gene sequence encoding an N-terminal collagen stalk followed by a globular complement 1q domain (gC1q), an architecture that characterizes the C1q A, B and C chains of the first complement component (C1), did not become prevalent until the cephalochordates and urochordates. However, genes encoding only the globular complement 1q domain (ghC1q) are more ancient as they exist within many lower vertebrate and invertebrate genomes, and are even present in the prokaryotes. These genes can be divided into two groups, the first, which appears to be the more ancient form, encodes proteins that are not secreted (cghC1q). The second group encodes proteins in which the globular domain is preceded by a signal peptide indicating secretion (sghC1q). In this review we examine bioinformatic evidence for C1q domain containing (C1qDC) genes in many organisms and integrate these observations with research performed and published on the biochemistry and functions of this fascinating set of proteins.

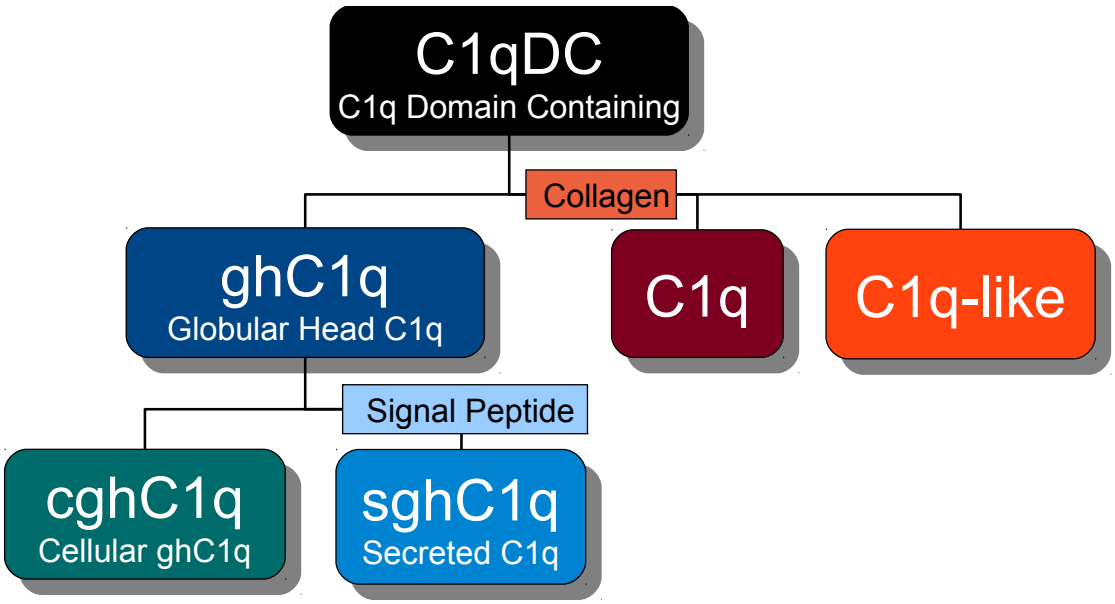


Figure 2.1: Flowchart outlining the relationship of C1qDC to C1q-like proteins, C1q, ghC1q, sghC1q and cgHC1q proteins. C1qDC = C1q domain containing; C1q-like proteins = peptide that has a collagen domain preceding a gC1q domain; C1q = first complement component consisting of C1q A, B and C chains; ghC1q = globular head C1q; cghC1q = globular head C1q domain protein containing no signal peptide, probably intracellular function. sghC1q = globular head C1q domain protein that contains a signal peptide, probably extracellular function.

Table 2.1: Table outlining definitions for abbreviations used in this chapter.

C1qDC	C1q domain (gC1q) containing proteins refer to all proteins that contain a C1q domain. Includes proteins with and without collagen
gC1q	Globular C1q domain. Structural term that refers to the amino acid sequence that folds into the jelly-roll topology
C1q	Complement component 1, subcomponent q. C1q forms a hexamer of heterotrimers (total of 18 peptides) in which each heterotrimer is a globular head of C1q A, B and C chains
C1q - like	A peptide that contains a collagen portion and a gC1q
ghC1q	Globular head C1q protein. Protein that contains only a gC1q and a short N-terminal that does not form a special motif. As exemplified by the precerebellins and CAPRINs
cghC1q	A (cellular) globular head C1q protein that does not contain a signal peptide. Exemplified by CAPRINs
sghC1q	A (secreted) globular head C1q protein that contain a signal peptide. Exemplified by precerebellin and precerebellin-like protein

2.2 Introduction

Genes encoding complement component 3 (C3) have been investigated within invertebrate genomes and traced in evolutionary history to the cnidarian radiation [Dishaw et al., 2005] [Nonaka and Kimura, 2006] [Pinto et al., 2007]. Presently the genes encoding for proteins containing a C1q domain (C1qDC) (Table 2.1, Figure 2.1, Figure 2.2) have been partially investigated from an evolutionary perspective [Dodds and Matsushita, 2007]. These genes exist within many of the sequenced mammalian, lower vertebrate and invertebrate genomes and functions have been described for some of these C1qDC proteins. However, many have not been characterized at all. For example, within the human genome, 32 open reading frames encoding C1qDC proteins have been found [Tom Tang et al., 2005] while within the zebrafish genome at least 52 exist [Mei and Gui, 2008]. In this review we will broadly cover all known C1qDC proteins found within the metazoa as well as suggest a comprehensive set of abbreviations with which to refer to them (Table 2.1 and Figure 2.1). The focus of this review will be primarily on the C1qDC proteins wherein the globular domain is preceded only by a short N-terminal amino acid sequence (as exemplified by the precerebellin-like proteins) (Figure 2.2) [Gerwick et al., 2000] and it will also contain a brief discussion of the C1q-like proteins that contain an N-terminal collagen portion (Figure 2.2).

The C1qDC proteins are a large group of proteins with many members that have been organized into groups on several occasions. In 2005, Tom Tang et al. [Tom Tang et al., 2005], divided the human C1qDC into three sub-families based on their sequence homology. In 2007 this scheme was basically agreed upon by Ghai et al. [Ghai et al., 2007] with an alignment of the human C1q proteins that divided them into two families with subgroups; the larger family containing the C1q-like and cerebellin-like subgroups while the smaller family was composed of EMILINs and multimerins. This was largely reiterated phylogenetically by Mei in 2008 [Mei and Gui, 2008] using the zebrafish C1qDC proteins. In 2008, using the mouse genome, Yuzaki [Yuzaki, 2008] further divided what had been established as sub-family B into what where referred to as the Cbln (precerebellin) and C1ql (C1q-like) groups. Two C1qDC proteins, the human C1q glob-

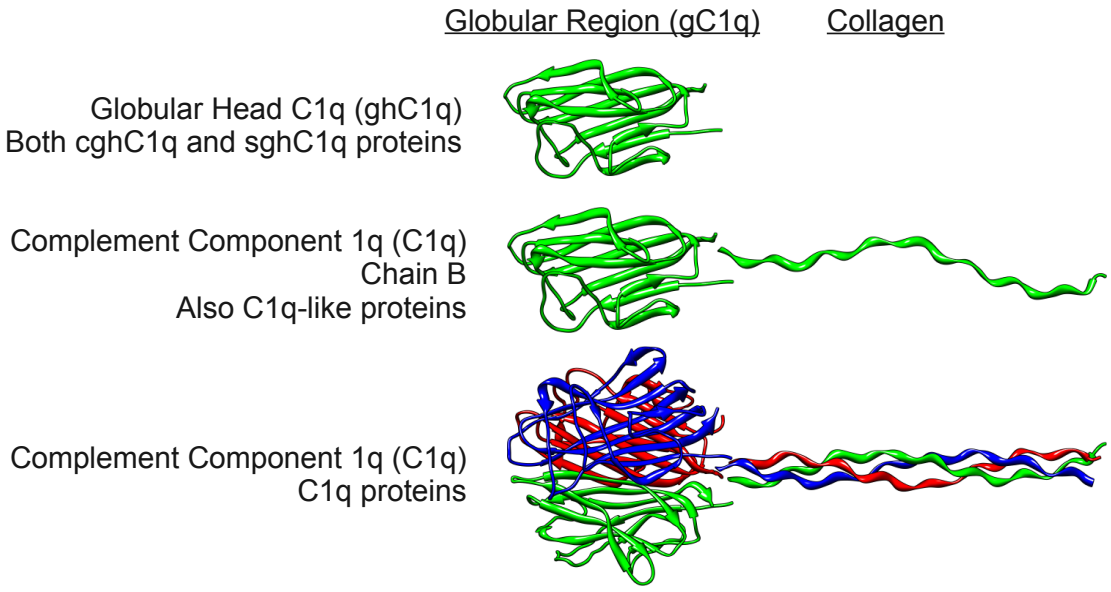


Figure 2.2: Protein crystallographic structures of the gC1q domain [PDB:1PK6] as it appears in the sghC1q proteins (including Cblnl and Cbln), Chain B of C1q and the entire C1q molecule. Collagen [PDB:1CAG] is found attached to C1q (Chain B or the entire protein) but not the cblnl proteins. Graphics were created with the Chimera viewer [Pettersen et al., 2004].

ular domain (PDB:1PK6) [Gaboriaud et al., 2003] and adiponectin (ACRP 30) (PDB:1C3H) [Shapiro and Scherer, 1998], have been crystallized and their X-ray structures determined to reso-lutions of 1.9 and 2.1 angstroms, respectively. From these crystal structures the 3D conformations have been deduced revealing that the C1q domain is characterized by its ability to fold into a jelly roll topology of five pairs of anti-parallel β-strands creating two β-sheets, generally referred to as the globular domain (gC1q) [Gaboriaud et al., 2003, Shapiro and Scherer, 1998, Jones et al., 1989].

2.3 C1q structure and function

Of all the C1qDC proteins, the mammalian first complement component (C1) has been the most thoroughly studied, both structurally and functionally. Sub-component q of C1 (C1q) forms a hexamer of heterotrimers (total of 18 peptides) in which each heterotrimer is a globular head of C1q A, B and C chains

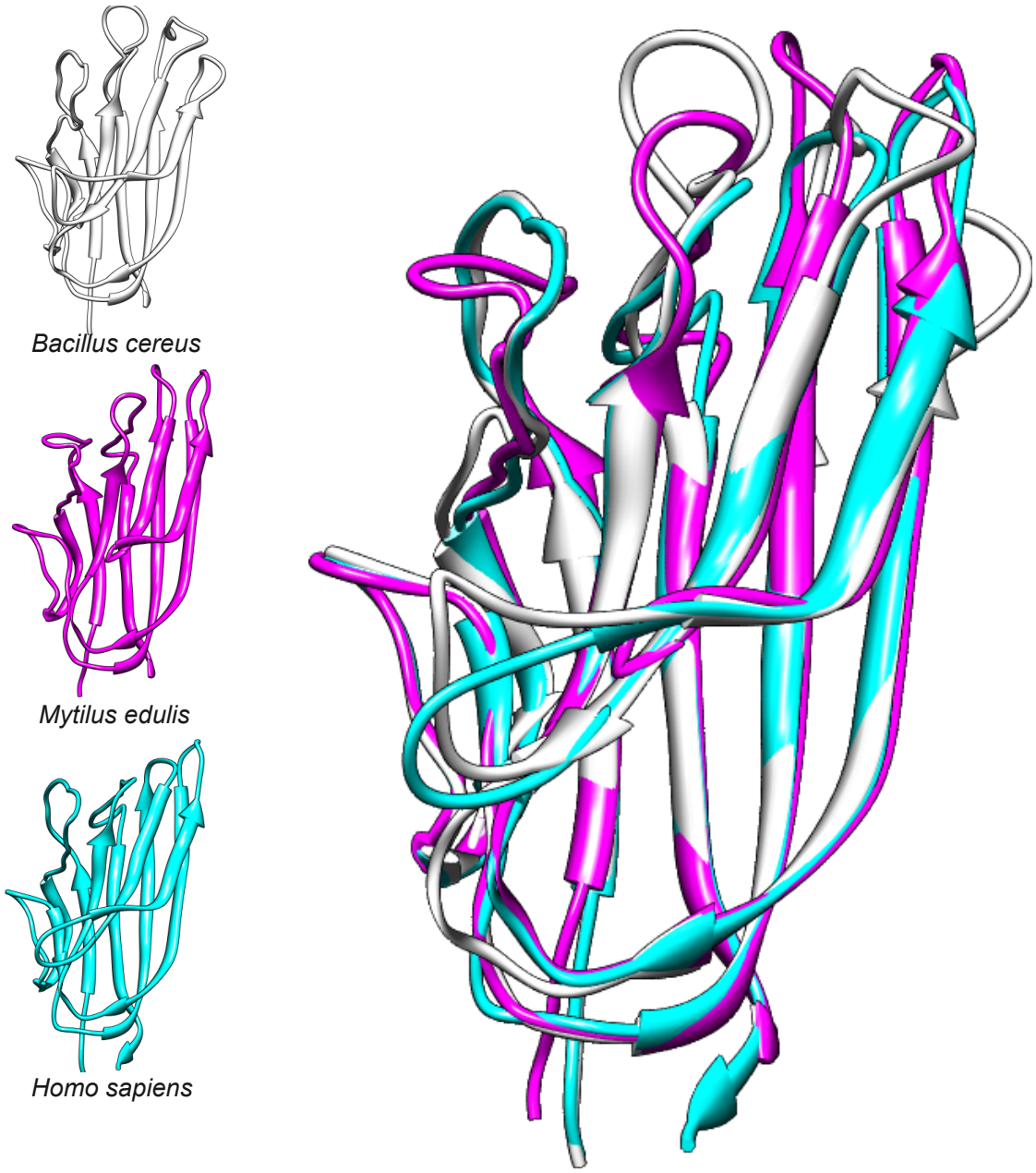


Figure 2.3: Protein modeling of three gC1q homologs from the bacterium *B. cereus* (white), the mussel *M. edulis* (pink), and the vertebrate *H. sapiens* (blue). To the left they are viewed singularly and to the right they are seen superimposed upon each other to exhibit how similar their deduced structures are. Modeling was done with the M4T server [Rykunov et al., 2008] and graphics were created with the Chimera viewer [Pettersen et al., 2004].

(Figure 2.1). C1q associates with C1s and C1r to form the C1 complex. This complex is the initiator of the classical complement pathway in which it binds to IgM, IgG or C-reactive protein (CRP) on the cells surface, thus activating C4. This initiates the formation of the membrane attack complex and subsequent breaching of the cell membrane [Ghai et al., 2007, Sjwall et al., 2007]. C1q has also been studied for its ability to interact with a diverse set of molecules including ligands on the surfaces of pathogens. These interactions have been mapped to different binding sites on the C1q globular head [Kishore and Reid, 2000, Kishore et al., 2004]. The nature of this binding appears to be a charged pattern recognition between the C1q peptides and theligand, however, no specific amino acid motif has been identified which promotes this interaction [Ghai et al., 2007].

2.4 C1q-like proteins

A C1q-like gene containing a 5' nucleotide sequence that encodes the amino acid repeat Gly-Pro-X, a feature which forms the collagen helix, in which X can be any of the other amino acids, and a 3' end that encodes the amino acids needed to form the globular C1q domain (Figure 2.2) have been detected in the medicinal leech *Hirudo medicinalis* [Tahtouh et al., 2009]. There are at least three receptors that can interact with C1qDC proteins: CR1, gC1qR, and $\alpha 2\beta 1$ integrin [Bohlson et al., 2007]. One of them, the gC1qR, interacts with the globular C1q domain [Ghebrehiwet et al., 1994]. This ligand receptor interaction was exploited when it was found that the leech C1q-like peptide elicited chemotactic behavior that could be blocked by the use of a human antibody towards the gC1q receptor [Tahtouh et al., 2009]. Experiments using both human and murine mast cells have also shown that gC1qR is involved in chemotaxis [Peerschke et al., 2004]. The results from the study of the leech C1q-like protein indicate that the gC1qR must be highly conserved since a human gC1qR antibody appears to be able to block the leech receptor [Tahtouh et al., 2009].

C1q-like gene copies have also been found in the urochordate *Ciona intestinalis* (sea squirt) and the cephalocordate *Branchiostoma floridae* (Florida lancelet)

Table 2.2: C1qDC genes and their structural characteristics.

Phylum	Binome	NCBI identifier	AA length	Collagen	SignalP	Exons
Firmicutes	*Bacillus cereus*	225785986	182			1
		52143142	178			1
Arthropoda	*Locust migratoria*	55889132	93			
	Caligus clemensi	225718376	220			
Platyhelminthes	*Schistosoma mansoni*	256090616	198			
Annelida	*Hirudo medicinalis*	184186854	320	Y	∼Y	
Nematoda	*Brugia malayi*	170580241	272			
	Cepaea hortensis	38043955	159		Y	
	Chlamys farreri	153793266	178		Y	
	Mytilus edulis	46395578	213			
		38635428	236		Y	
Echinodermata	*Strongylocentrotus purpuratus*	72123773	139		Y	3
		115925109	402		Y	6
Hemichordata	*Saccoglossus kowalevskii*	187143440	137			
Chordata	*Branchiostoma* floridae	260818350	266	Y	Y	1
		219442714	220		Y	4
	Ciona intestinalis	198430309	674	Y	Y	1
		18132947	254			
	Squalus acanthias	56843967	126			
		115393629	123		Y	
	Lethenteron japonicum	48675340	240	Y	Y	
	Danio rerio	167555053	336	Y	Y	3
		158534007	226		Y	3
	Tetraodon nigroviridus	47219370	283	Y	Y	
		56239997	200		Y	
	Xenopus tropicalis	114108311	244	Y	Y	2
		147905600	215		Y	4
	Mus musculus	6753220	253	Y	Y	2
		56744247	193		Y	3
	Homo sapiens	87298828	253	Y	Y	2
		4757922	193		Y	3

but it can be expected that more C1q-like genes will be found as more sequencing information becomes available (Table 2.2). However, the isolated case of a C1q-like peptide in the leech is especially interesting since none of the sequenced platyhelminth, nematode, molluscan or echinoderm genomes contain C1q-like genes. Several open reading frames in the echinoderm genome contain 5 codons coding for glycine and proline residues but not in the systematic Gly-Pro-X repeated fashion of collagen. This gene motif, as mentioned above, does appear in transcriptomes of amphioxus, lamprey, and several teleost fishes. Few of these putative C1q-like proteins have been characterized; however, some of these proteins appear to bind to a variety of carbohydrates; hence they may function as lectins [Matsushita et al., 2004].

As mentioned, a C1q-like protein was isolated from lamprey (an agnathan) that has lectin properties indicated by its isolation using an N-acetyl-d-glucosamine-SepharoseTM affinity column. Furthermore, the lamprey C1q-like protein has a mass of 480 kDa under native and non-reducing conditions, indicating that it exists as an 18 peptide multimeric protein, identical to the structure of mammalian C1q [Matsushita et al., 2004]. In addition to being able to bind to N-acetyl-d-glucosamine, the lamprey C1q-like protein, when co-purified with MASP-A, was able to cleave the C3 molecule also isolated from lamprey serum. In conclusion, C1q-like genes did not become common in genomes or transcriptomes until the evolution of the urochordates and cephalochordates. The exception, at this time, appears to be the medicinal leech; however, it is not known if this C1q-like protein forms a multimeric complex similar to the C1q complex seen in mammals [Tahtouh et al., 2009].

2.5 ghC1q proteins (precerebellin, precerebellin-like, CAPRIN)

The globular head C1q (ghC1q, see Table 2.1), protein structure differs from that of the C1q-like proteins in that it lacks the collagen region and instead has only a short N-terminal amino acid sequence with no particular motif followed

by the C1q domain. Most of these proteins contain a 5 signal peptide indicative of secretion (sghC1q). The precerebellin (cbln) and precerebellin-like proteins (cblnl) are subgroups of the sghC1q proteins. Copies without signal peptides (cghC1q, exemplified by CAPRINs) also exist [Yuzaki, 2008].

Four different precerebellin genes (Cbln1, Cbln2, Cbln3, and Cbln4) have been found in the human and mouse genomes, and homologs of these can be traced to many invertebrate and teleost genomes and transcriptomes. However, in many of these genomes and transcriptomes more than four homologs can be found. Presently, the evolutionary outliers appear to be those found in bacteria of the genus *Bacillus*, although these lack an identifiable signal peptide (cghC1q) (Table 2.2). Other cghC1q genes have also been found in other phyla, including the platyheminthes and arthropods (Table 2.2). These genes align well with the mammalian proteins termed cell cycle associated proteins (CAPRINs, or alternatively C1qdc1 or EEG-1L). The CAPRINs are involved in intracellular processes [Solomon et al., 2007], as opposed to the sghC1q proteins which are likely secreted from the cell.

The possibility that the ghC1q protein in *Bacillus sp.* resulted from horizontal gene transfer was investigated and rejected using the Dark Horse program [Podell and Gaasterland, 2007]. A computational protein modeling (via M4T [Rykunov et al., 2008]) of the putative gC1q *Bacillus* proteins revealed the stereotypical gC1q jelly roll topology with 10 β-strands (Figure 2.3). This finding substantiates that the gC1q domain has an ancient evolutionary history.

Among molluscs, some of the ghC1q genes encode for a signal peptide (sghC1q), indicating a potential expansion or change in the function(s) of the proteins containing this very utilitarian fold. The sghC1q gene continues to occur in organisms from sea urchins (*Strongylocentrotus purpuratus*), amphioxus (*B. floridae*), sea squirts (*C. intestinalis*), and all the way to the mammals.

The functions of the precerebellin proteins (sghC1q) have been investigated in mice and humans with notable progress. The cerebellin peptide, a 16 amino acid peptide contained within the precerebellin1 protein (Cbln1), has been located by immunohistochemistry within cerebellar Purkinje cells and cartwheel neurons

in the mammalian dorsal cochlear nucleus [Mugnaini and Morgan, 1987]. A cbln1 knockout mouse showed lack of synaptic plasticity and integrity in the cerebellar parallel fiber-Purkinje cell synapses [Hirai et al., 2005]. In addition, a mutant mouse that lacks the delta 2 glutamate receptor exhibits the same phenotype as a cbln1$^{-/-}$ mouse, suggesting that cbln1 and the delta 2 glutamate receptor may be part of the same pathway [Yuzaki, 2009]. The cbln1 gene is also expressed in the adrenal gland were it appears to be involved in secretion of corticosteroids [Rucinski et al., 2009].

Most investigations of the cbln genes in mammals have been restricted to early developmental stages, so questions regarding their function(s) during adulthood remain unanswered. It has been suggested that they may serve as transneuronal cytokines within the brain [Yuzaki, 2008]. It is interesting to note that expression of the cbln transcripts can be found in many different tissues and that the expression of cbln1 does not overlap with cbln2, 3 or 4 [Yanai et al., 2005]. Further studies are needed to determine the functions of all four cbln genes as they may be temporal and tissue specific.

2.6 Immune response and sghC1q proteins

In 2000, Gerwick et al. [Gerwick et al., 2000] isolated and characterized a sghC1q protein (NCBI gi:100135893) that was present in rainbow trout plasma after injection with *Vibrio anguillarum*. In addition, Murai et al. [Murai et al., 1990], while searching for fish C-reactive protein, isolated a trout protein, eluted from a C-polysaccharide affinity column, that was subsequently called trout C-polysaccharide binding protein (TCBP). Its N-terminal was sequenced by Edman degradation and determined to be the N-terminal of a sghC1q sequence found during sequencing of the clones from a subtractive suppressive library constructed from liver RNA (NCBI gi:100136613) [Bayne et al., 2001].

The ability of an sghC1q protein to function as a lectin was confirmed in the surf perch (*Neoditrema ransonnetti*) in which a sghC1q was isolated from plasma using a fucose affinity column [Nakamura et al., 2009]. In addition, this isolated

23 kD surf perch fucose-binding protein appeared as multiple bands on an SDS-PAGE gel, in increments of 23 kD, under non-reducing conditions, indicating that multimers were formed. The authors also showed that the transcript for this gene was up-regulated after exposure to an inflammatory stimulus. Furthermore, a sialic acid binding protein has been detected in the snail (*Cepaea hortensis*) and an LPS binding protein found in the Zhikong scallop (*Chlamys farreri*) and in both cases the isolated protein fits the description of a sghC1q protein [Gerlach et al., 2004, Zhang et al., 2008a]. Additionally, when rainbow trout were injected intracranially with *V. anguillarum* (Gram negative), *Carnobacterium piscicola* (Gram positive) and Freunds incomplete adjuvant, semi-quantitative PCR revealed the presence of the sghC1q transcript indicating that sghC1q was up-regulated in the brain during the ensuing inflammatory response [Gerwick et al., 2005]. These were the first indications that some of the sghC1q homologs might be involved in the innate immune response.

Several features of the trout Cblnl protein (NCBI gi:100135893) may be interpreted to imply immune-type functions. First, transcription of its gene is induced in the liver and brain following an inflammatory stimulus. Second, the cblnl gene that is transcribed and translated in the liver is subsequently released into the plasma [Murai et al., 1990]. Finally, the protein contains a functional domain of the complement component C1q. It is notable that its upregulation appears to be non-specific, since both Gram negative and Gram positive bacteria, as well as an irritant (Freunds incomplete adjuvant) increased the amount of cblnl transcript [Gerwick et al., 2005].

Continued research, using zebrafish as a model organism for the ghC1q type proteins has revealed the surprising result that its genome contains 27 copies of the ghC1q genes, seven of which do not contain a signal peptide (unpublished data). Eleven, so far, of the sghC1q paralogs have been investigated at the transcriptional level. Five of these genes are transcribed during early development, two are up-regulated during the inflammatory challenge while one (cbln1) is down-regulated. Interestingly, only one of the genes is expressed during both early development and the inflammatory challenge; however, it appears to be constitutively expressed. All

of the other genes investigated were expressed either during early development or during the inflammatory challenge [Carland et al., 2011].

2.7 Conclusions

The Animal Kingdom contains many C1qDC proteins. In this review, we have discussed the ghC1q proteins and the C1q-like proteins. The ghC1q proteins, which lack the N-terminal collagen portion seen in the C1q-like proteins, are broadly distributed from bacteria to the vertebrates. However, the ghC1q gene motif becomes more prevalent in the metazoan lineage starting with the protostomes and radiating towards the vertebrates. Reports regarding the existence of a ghC1q protein in *Bacillus cereus* were investigated and revealed, after modeling, that the gC1q in bacteria contained a β-barrel with 10 β-strands and no obvious signal peptide. The possibility that these ghC1q genes arrived in the *B. cereus* genome via horizontal gene transfer was also investigated but there was no evidence to support this conjecture. The *Bacillus* ghC1q domain motif could be the ancestral gene sequence that further evolved into the C1qDC family of proteins. Unfortunately, nothing is known about the function of these ghC1q domain proteins in bacteria.

The gene sequence that codes for a C1q-like peptide (N-terminal collagen with a C-terminal gC1q domain) has been discovered in the medicinal leech. Whether or not this is the ancestral gene of the human C1q gene will have to be determined when more sequence information is available. Lampreys contain the first C1q-like complex of 18 peptides but further investigation of the Ciona and amphioxus C1q-like genes could shed light on when the first C1q complex arose. It has been discussed in earlier reviews that the alternative complement pathway is more ancient than the classical [Dodds and Matsushita, 2007]. While this is true in terms of the pathway itself, the gC1q protein motif is more ancient than the C3 gene motif (no hits against the Joint Genome Institute microbial database).

Several of the sghC1q proteins have been found to have lectin activity. Further investigations will have to determine if other sghC1q proteins also have the

ability to act as lectins and if the different homologs within a species display affinity for different carbohydrates, thus creating a diverse set of pattern recognition molecules. Furthermore, the ability of the sghC1q proteins to influence chemotaxis as well as their ability to increase gC1qR mediated phagocytosis needs further investigation. The ghC1q proteins are a large group of proteins found in many organisms (including mammals) and their roles and importance in the innate immune response remains to be fully seen and demands further attention to increase our knowledge of the immune system and its evolution.

2.8 Acknowledgements

This publication was prepared by Tristan M. Carland and under NOAA Grant #NA04OAR4170038 California Sea Grant College Program Project #R/A-124, through NOAAs National Sea Grant College Program, U.S. Dept. of Commerce; and was supported in part by the California State Resources Agency. The statements, findings, conclusions and recommendations are those of the author(s) and do not necessarily reflect the views of California Sea Grant or the U.S. Dept. of Commerce. In addition, funding from the UCSD Academic Senate and the Ledger Benbough Foundation supported the project.

Chapter 2 is a full reprint of the publication: Carland, T. M. and Gerwick, L. (2010). The C1q domain containing proteins: Where do they come from and what do they do? Developmental and Comparative Immunology, 34(8):785-790, with permission from all coathors.

Chapter 3

Differential expression and intrachromosomal evolution of the *sghC1q* genes in zebrafish (*Danio rerio*)

3.1 Abstract

The secreted globular head C1q (*sghC1q*) genes can be characterized as a family of genetic loci encoding signal peptides followed by single complement component 1q globular (gC1q) motifs. Members of this family have been referred to as precerebellin-like (cblnl), C1q-like or ovary specific C1qlike factors, and are transcribed in response to infection and/or during early development. This study was primarily undertaken to identify the zebrafish *sghC1q* (or *DrsghC1q*) genes that increase their transcription in response to infection and to examine their transcriptional patterns during early development. Twenty *sghC1q* genes were found in the zebrafish (*Danio rerio*) genome (Zv9). Two of the examined twenty genes showed significant up-regulation within 24 h of infection with the fish pathogen *Streptococcus iniae*, and eleven of the examined twenty were expressed during early development. Due to the clustered nature of these genes on chromosomes two and

seven, intrachromosomal duplication events are hypothesized and explored.

3.2 Introduction

The immediate innate immune response (acute phase response) has been characterized to a certain extent in humans and mice [Murphy et al., 2008], and is beginning to be explored in several fish species through proteomic and transcriptional studies. These fish species include rainbow trout (*Oncorhyncus mykiss*) [Gerwick et al., 2007, Raida and Buchmann, 2009], catfish (*Ictaluras punctatus*) [Peatman et al., 2007], zebrafish (*Danio rerio*) [Hegedus et al., 2009], turbot (*Scophthalmus maximus*) [Pardo et al., 2008], Japanese flounder (*Paralichthys olivaceus*) [Dumrongphol et al., 2009], carp (*Cyprinus carpio*) [Gonzalez et al., 2007], tilapia (*Oreochromis mossambicus*) [Ndong et al., 2007], large yellow croaker (*Pseudosciaena crocea*) [Yan et al., 2009] and hybrid striped bass (*Morone saxatilis* x *M. chrysops*) [Pasnik and Smith, 2006] among others. Changes in gene transcription occur soon after infection, and as such, the newly transcribed genes can be considered to encode acute phase proteins. The acute phase response is an organism's response to physiological insult (infection, injury, etc.), or is alternatively defined as any time during which the organism is not in homeostasis [Bayne and Gerwick, 2001]. Acute phase proteins are those experiencing altered synthesis during the onset of non-homeostasis, as measured by changes in transcript or protein abundance. The onset of the altered transcription is likely variable between different acute phase genes. However, numerous studies have found transcription of these genes to be activated and peaking within the first 8-24 h post infection (hpi), sometimes lasting for days, though the response tends to subside within 24-48 h [Baumann and Gauldie, 1994, Cray et al., 2009]. Many acute phase proteins are secreted from the liver into the plasma (vertebrates) or from the hepatopancreas or equivalent organ into the hemolymph (invertebrates). In previous studies the hepcidin gene rapidly increased its transcription between 4 and 48 hpi [Bayne et al., 2001, Lauth et al., 2005], haptoglobin's transcription increased by 12 hpi [Giffen et al., 2003, Quaye, 2008], and mannose-binding lectin

was detected for multiple days with the transcription only occurring in the liver [Sastry et al., 1991]. In addition, several complement components, transferrin and the precerebellin-like protein (Cblnl) transcripts have been detected at 24 hpi [Peatman et al., 2007, Gerwick et al., 2007]. All these examples fit the definition of being acute phase genes [Baumann and Gauldie, 1994].

The Cblnl protein was first discovered as an acute phase protein in rainbow trout [Gerwick et al., 2000]. This protein was so named because it shared 53% amino acid sequence similarity to precerebellin (Cbln) [Urade et al., 1991], and in addition, shares 47% identity with the B-chain of the first complement component (C1qB). Precerebellin is considered a precursor to the neuropeptide cerebellin [Umrath and Silberbauer, 1967]. The Cbln and Cblnl proteins differ in sequence and structure from complement component 1q (C1q) as they lack N-terminal sequences that fold into alpha-helical collagen structures [Carland and Gerwick, 2010]. There are currently four cbln homologues (cbln1-4) in mammals. These genes (*cbln*) are largely expressed in the brain during development [Wei et al., 2007] [Rucinski et al., 2009] and appear to function in the formation and stabilization of synaptic contact and the control of functional synaptic plasticity between cerebellar granule cells and Purkinje cells [Hirai et al., 2005]. The four genes exhibit differential expression in the mouse brain during development and in adults [Miura et al., 2006]. All of these proteins are able to form homomeric and heteromeric trimers via their shared C1q domains and larger assemblies (dimers of trimers) by disulfide bonds from their respective dual cysteine residue motifs [Bao et al., 2005]. It has also been shown in mice that expression of cbln1 can modulate the trafficking of Cbln3 out of the endoplasmic reticulum [Iijima et al., 2007]. The Cbln and Cblnl proteins all fit within the secreted globular head C1q (*sghC1q*) protein family recently reviewed by [Carland and Gerwick, 2010].

The complement system is an innate defense mechanism that can lead to the eradication or opsonization of pathogens and damaged tissues. Complement can be triggered by recognition of substrates by complement component 1 (C1), lectin, or C3 tick-over. C1q is an important structural and binding component of C1, a protein complex consisting of C1r, C1s, and C1q. The C1 protein com-

plex initiates the complement cascade leading ultimately to the formation of the final pore-forming membrane attack complex. The C1q molecule itself interacts with several ligands including recognition molecules on pathogens and responds through either C1qR receptor mediated phagocytosis or via the deposition of C1r and C1s on the cell surface leading to activation of the complement cascade [Kishore and Reid, 2000]. More recently C1q expression was identified in postnatal neurons and found to mediate elimination of inappropriate synaptic connections during development [Stevens et al., 2007].

Like the Cbln proteins, the Cblnl proteins have a relatively high amino acid sequence identity to C1q, due to their shared globular C1q domain (gC1q); ten β strands (Figure 3.1) folding into a β-barrel formation [Ghai et al., 2007]. The C1q domain has been genetically conserved and replicated throughout vertebrate evolution as there exist at least 31 C1q-domain-containing proteins within the human (*Homo sapiens*) genome [Tom Tang et al., 2005], at least 52 in the zebrafish genome [Mei and Gui, 2008] and 75 in the amphioxus (*Branchiostoma floridae*) genome [Huang et al., 2008]. These C1q domain-containing proteins are considered a structural family and have a diverse range of functions. Kishore et al. 2004 [Kishore et al., 2004], grouped the C1q domain proteins into three subfamilies based on their sequence similarities to the A, B or C chains found in the heterotrimeric heads of the hexameric human C1q protein. The Cblnl family of proteins fit as members of the B group [Kishore et al., 2004]. We have recently proposed a more structurally based classification as a further clarification of the gC1q family [Carland and Gerwick, 2010]. The *sghC1q* group can be identified by their: (a) tendency to contain an N-terminal signal peptide, (b) C-terminal gC1q domain, (c) size of between 100 and 300 amino acids, and (d) lack of an N-terminal collagen-like region.

Studies of proteins containing C1q domains have recently been undertaken in a few aquatic organisms. In surfperch, (*Neoditrema ransonnetii*) the protein was isolated using a fucose affinity column. The C1q domain protein in the Zhikong scallop (*Chlamys farreri*) has the capacity to bind lipopolysaccharide [Nakamura et al., 2009, Zhang et al., 2008a]. Transcriptional work in the mussel

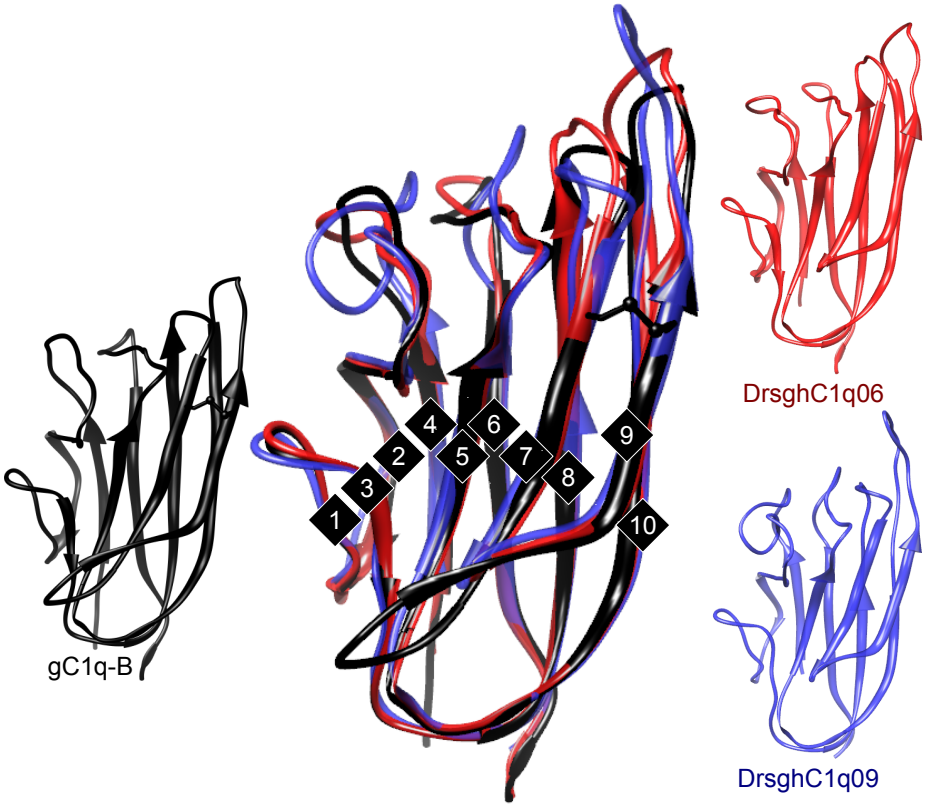


Figure 3.1: *DrsghC1q* Modeling. The models seen here are a combination of chain B of the globular portion of mammalian C1q (gC1q-B in black) as determined by X-ray crystallography and computational predictions of the structures for *sghC1q06* (red) and 09 (blue). The large model is a combination of the three, illustrating the conservation of the ten β-strands (numbered in black). Also shown in black is the disulfide bond known to mammalian C1q and some of the *DrsghC1q* genes. Graphics developed in the Chimera viewer [Pettersen et al., 2004].

(*Mytilus galloprovincialis*) has revealed significant upregulation of MgC1q after infection and the likelihood of adaptive molecular features based on positive selection analyses [Gestal et al., 2010]. The goldfish (C. auratus) ovary specific C1q-like (CaOC1q-like) protein contains a collagen region and appears to only be expressed in follicular epithelial cells [Mei et al., 2008a]. The previously identified C1q-like factor from zebrafish fully fits the characteristics of the Cblnl family of proteins and it was shown to inhibit p53-mediated apoptosis during head and craniofacial development [Mei et al., 2008b].

Zebrafish was chosen as a model organism for further functional studies of the *sghC1q* genes due to its extensive utility in functional genomic studies [Yoder et al., 2002] and as a model for studying extensive gene duplication events. In order to characterize the *sghC1q* genes upregulated in zebrafish during the inflammatory response, we identified twenty genes in the zebrafish genome that fit the description of an *sghC1q* gene [Carland and Gerwick, 2010]. These twenty genes were explored by transcriptional profiling during two different physiological conditions; the first in response to an inflammatory stimulus and the second during early development. Our study of the *sghC1q* genes includes the four previously identified *cbln* genes [Mei and Gui, 2008] as well as the previously studied C1q-like gene [Mei et al., 2008b].

3.3 Materials and methods

3.3.1 Zebrafish maintenance

Adult zebrafish (*Danio rerio*) were purchased from Aquatica Tropicals (Plant City, FL) and maintained at 28℃ in a three tier Table Top Rack (Aquaneering Inc. San Diego, CA) following standard husbandry procedures for care and feeding [Westerfield, 2007]. When not breeding, male and female fish were cohabitated to prevent harmful overproduction of eggs. Fish were fed a dry flake mix (57% Aquatox Flake, 19% Spirulina Flake, 8% Hikari Micropellet, 8% Cyclop-eeze, 4% Golden Pearl 300-500 and 4% Golden Pearl 500-800) and freshly grown brine shrimp two to three times daily.

3.3.2 Zebrafish breeding and egg collection

On the day prior to breeding, male and female fish were recollected and placed in a standard 2-way fish breeder (Petsmart, Inc.) in the afternoon and kept separated by a clear plastic divider during the night without tank water system flow. At daybreak (at least 16 h after separation) the divider was removed and fish were given privacy for 1 h. The adult fish were then removed and the eggs were collected and sorted with plastic transfer pipettes (VWR, Cat.#16001-174) and considered T=0. Eggs were kept in a salt solution with 0.001% methylene blue [Westerfield, 2007]. Eggs were collected at 12, 24, 36 and 48 hpf and placed in 1.5 ml microcentrifuge tubes with 100 µl TRIZol® Reagent (Invitrogen Cat.#15596-026). The collected eggs were flash frozen in liquid nitrogen followed by storage at -80℃. This experiment was repeated a second time.

3.3.3 Bacteria culture, strain and preparation

Streptococcus iniae strain K288 was isolated from the brain of a diseased hybrid striped bass at the Kent SeaTech aquaculture facility in Mecca, California [Buchanan et al., 2005]. *S. iniae* was grown at 30℃ in Todd Hewitt Broth (THB) or on Todd Hewitt Agar (THA) (Hardy Diagnostics). Prior to injection, an overnight culture of *S. iniae* was diluted 1:10 in fresh THB and grown to mid-log phase (optical density, OD600 = 0.40). A 1.0 ml aliquot of the culture was centrifuged at 3500xg for 5 min, washed once in an equal volume of phosphate buffered saline (PBS) (Gibco) (no calcium or magnesium), pelleted, and resuspended in PBS. Bacteria were then diluted in PBS to a final concentration of 3.5 x 10^5 CFU/ml and held on ice until injected.

3.3.4 Bacterial infection and zebrafish liver collection

Male zebrafish were placed in independent challenge tanks and allowed to acclimate for one week. The separate tanks had the same water parameters and the fish were given a reduced feed cycle (once daily). Fish were anesthetized in Tricaine (3-aminobenzoic acid ethyl ester, Sigma-Aldrich, St. Louis, MO) and challenged

via intraperitoneal injection of 10 µl of *S. iniae* (3.5 x 10^3 CFU) or PBS (control) using a 0.3 cc syringe and 29 g needle as previously described [Phelps et al., 2009]. At 12 and 24 h after the injection fish were snap frozen in liquid nitrogen and their frozen livers removed. Three livers were pooled for each sample in 200 µl of TRIZol® Reagent and kept frozen in liquid nitrogen. Each time point had two biological replicates and the entire experiment was repeated once. After injection the CFU were confirmed through serial dilution of the starting inocula and plating on THA. Fish challenges were carried out in an AAALAC certified facility following IACUC approved protocols.

3.3.5 Gene discovery and annotation

Using the previously characterized trout cblnl protein sequence [NCBI: NP001117737] as the query, tBLASTn and BLASTp alignments were performed against the zebrafish data available from TGIP (webref) and at NCBI (webref) as well as PSI-BLAST alignments (NCBI only). The resulting sequences (> 500) were built into contiguous sequences with CAP3 [Huang and Madan, 1999] where appropriate and their ORFs extracted via web-service pipeline EST Keeper (Carland and Gerwick, unpublished). The best alignments were screened for size (< 300 AA), absence or presence of a signal peptide using the program SignalP [Emanuelsson et al., 2007], and the presence of the β-strand motifs that characterize this family via protein modeling (detailed below). The locations of the genes on their corresponding chromosomes were discerned using BLAT and BLASTn alignments against the Zv9 version of the zebrafish genome from Ensembl (webref).

3.3.6 Primer design

Full-length cDNA sequences and their corresponding genomic sequences aligned in Spidey (webref) to determine exon location to guide primer design. Primer3 v. 0.4.0 (webref) was used to design primers that would cross or span intron/exon boundaries with amplicon sizes between 80 and 120 bp. Primers were obtained from Sigma-Aldrich. 2.7. Phylogenetic analysis All of the *sghC1q*

genes from zebrafish and one from amphioxus (as an out-group) were aligned by codon using MUSCLE [Edgar, 2004] and loaded into TOPALi [Milne et al., 2009] for subsequent model selection testing and Bayesian tree phylogenetic analysis [Ronquist and Huelsenbeck, 2003, Anisimova and Kosiol, 2009]. Codon substitution model settings (obtained from model selection testing) are as follows: 1,000,000 generations, 65% burn in, HKY substitution model for codon position one with gamma and invariable sites, HKY substitution model again for codon position two with gamma and without invariable sites, K80 substitution model for codon position three without gamma or invariable sites, no parameter linking across codon positions.

3.3.7 Protein modeling

Protein models were generated through submission of amino acid sequences to the M4T modeling server [Fernandez-Fuentes et al., 2007] with analysis performed using default settings. Models are displayed using the UCSF Chimera package [Pettersen et al., 2004].

3.3.8 RNA isolation and cDNA synthesis

The embryonic and liver tissues were homogenized by manual force using a plastic pestle (VWR Cat# KT749520-0500) in a 1.5 ml microcentrifuge tube and 200 µl of TRIZol®. Total RNA was isolated from the tissue following the TRIZol® protocol for isolating RNA from animal tissue (Invitrogen, Carlsbad, CA) followed by cDNA synthesis using SuperScript™III Reverse Transcriptase (Invitrogen, Carlsbad, CA). Total RNA (800 ng/sample) were incubated with Superscript™III at 55℃ for 1 h following manufacturer's protocol for First Strand Synthesis.

3.3.9 Semi-quantitative Reverse Transcriptase PCR

Upon obtaining cDNA samples corresponding to the embryonic time points collected, PCR reactions were performed using all combinations of embryonic time

point derived cDNA and cblnl primer pairs. PCR reactions were set up as follows: 8 µl 2x MasterMix (Promega, Cat #PAM7505), 0.5 µl of 25 mM Mg2+, 0.5 µl of both forward and reverse primers at 20 µM each, 1 µl cDNA template (100 ng/µl) at 1:50 dilution and water for a final reaction volume of 20 µl. The subsequent PCR reaction (Eppendorf gradient thermocycler) was programmed to run at 95 ℃ for 3 min followed by 32 cycles of 95 ℃ for 15 s, annealing of 57 ℃ for 15 s and extension of 72 ℃ for 15 s, with a final elongation step of 72 ℃ for 2 min. Gel electrophoresis was performed with 5 µl of each PCR product mixed with 1.5 µl 5x Loading Dye in a 1% agarose gel stained with 1x SYBR Gold (Invitrogen, Carlsbad, CA) at 90 Volts for 55 min in sodium borate buffer. All resulting amplicons were sequence verified.

3.3.10 Quantitative Reverse Transcriptase PCR (qRT-PCR)

After validating primers and optimizing PCR conditions by use of agarose gel electrophoresis (as mentioned above) to ascertain that amplification yielded a single product of the predicted size, qPCR reactions were set up as follows: 10 µl of SYBR® Premix Ex Taq™, 0.4 µl of a 50x ROXII™solution (Cat #RR041A, Takara Mirus Bio Inc.), with primers at 200 nM final concentration, 1 µl cDNA template at 1:50 dilution, and dH2O for a final reaction volume of 20 µl. An MX3000p qPCR thermocycler (Stratagene) was programmed to run at 95℃ for 30 sec, 40 cycles of 95℃ for 15 sec, annealing of 53℃ for 15 sec and extension of 72℃ for 15 sec. The qRT-PCR protocol for each target gene was validated by melting curve analysis to ensure the absence of primer-dimers or other unwanted amplicons. The relative expression levels of the *sghC1q* transcripts were calculated by using the delta delta cT method and normalized by the expression level of *ef1α*. The MxPro v4 (Stratagene) software package was used to analyze raw data and OpenOffice.org Calc (Sun Microsystems, Inc.) was used to perform Student's t-tests of the means from infected and control livers and to graph the resulting fold changes following established protocols [Schmittgen and Livak, 2008].

3.4 Results and discussion

3.4.1 Intrachromosomal duplications as deduced by wide expansion, phylogeny and conserved predicted protein structure

The databases of zebrafish cDNAs and ESTs available at The Gene Indices Project (TGIP) website [Lee et al., 2005] were investigated using the tBLASTn and BLASTp algorithms [Altschul et al., 1990] using the sequence encoding the conserved C1q domain from the trout Cblnl sequence [Gerwick et al., 2000] and the zebrafish Cbln sequences identified by [Mei and Gui, 2008] as a queries. All of the identified transcripts that fit the criteria for potentially encoding a secreted globular head C1q (*sghC1q*) protein [Carland and Gerwick, 2010] were kept. Next, the NCBI non-redundant (NR) and expressed sequence tag (EST) datasets were similarly queried. From these searches, 597 sequences of varying similarity were obtained. Many of these represented partial duplicate ESTs so a contiguous sequence building algorithm [Huang and Madan, 1999] was employed to combine any exact duplicates, and then their open reading frames (ORFs) were extracted by a custom Perl script and this process was repeated using the EST Keeper program (Carland and Gerwick, unpublished). This process brought the number of sequences down to 45, consisting largely of the 52 C1qDC found previously [Mei and Gui, 2008]. After careful removal of sequences encoding a collagen motif, sequences without a signal peptide motif, sequences including intron, and sequences that were deemed too long (greater than 300 amino acids), the candidate list narrowed to 21 sequences that could be accurately located on the zebrafish genome using the BLAT algorithm [Kent, 2002]. One of these sequences appeared to be an alternative splice variant of another, thus leaving twenty genetic loci that were identified in the zebrafish genome. The TC number of each gene from TGIP (webref), their Ensembl identification number, their name in any previous studies and any NCBI accession numbers are listed in Table 3.1 along with their basic attributes and proposed formal family names. To bring order and clarity to the many different genes encountered, acronym based names are assigned to each of the genes that exist in

Table 3.1: Directory of the *DrsghC1q* genes according to Zv9. This table contains the proposed formal family names of the *sghC1q* genes in zebrafish as well as a listing of their previous names and identifiers from Ensembl and NCBI. Also shown are the chromosomes and number of exons of the genes.

Naming					
Formal	Previous Names	Chr	Exons	Ensembl	NCBI
DrsghC1q01	Cbln1	18	3	ENSDARG00000057296	50540101
DrsghC1q02	Cbln2a	2	3	ENSDARG00000074601	190337513, 192452522
DrsghC1q03	Cbln2b	24	3	ENSDARG00000077151	148922966
DrsghC1q04	Cbln4	23	3	ENSDARG00000061240	160333645
DrsghC1q05a/b	C1ql1, TC312919	2	4	ENSDARG00000053802	165970362, 187960124
DrsghC1q06	TC316425	7	3	ENSDARG00000035718	292616584, 166796410
DrsghC1q07	TC321847	2	3	ENSDARG00000095040	42406703, 47776018
DrsghC1q08	TC310692	7	3	ENSDARG00000019294	158534006
DrsghC1q09	TC326038	2	3	ENSDARG00000030254	92097198, 186910331, 213624837, 213624839
DrsghC1q10	TC310182	2	3	ENSDARG00000023157	70887628, 94536921, 120538649, 124001534
DrsghC1q11	C1q4L, TC341552	7	3	ENSDARG00000086654	146350791, 292615669
DrsghC1q12		2	3	ENSDARG00000068232	24459827, 42406648
DrsghC1q13		2	4	ENSDARG00000026904	158254221, 176866358
DrsghC1q14		2	4	ENSDARG00000053845	26984632, 78183339
DrsghC1q15		24	6	ENSDARG00000091278	42406703, 47776018
DrsghC1q16		2	3	ENSDARG00000088624	157154298, 292616586
DrsghC1q17		2	5	ENSDARG00000088911	32362346, 90954827, 117957400
DrsghC1q18		7	4	ENSDARG00000090969	1888528553, 292621116
DrsghC1q19		10	3		57163716, 83308939, 292613885, 46935016
DrsghC1q20		15	3	ENSDARG00000087476	188528553, 189528471

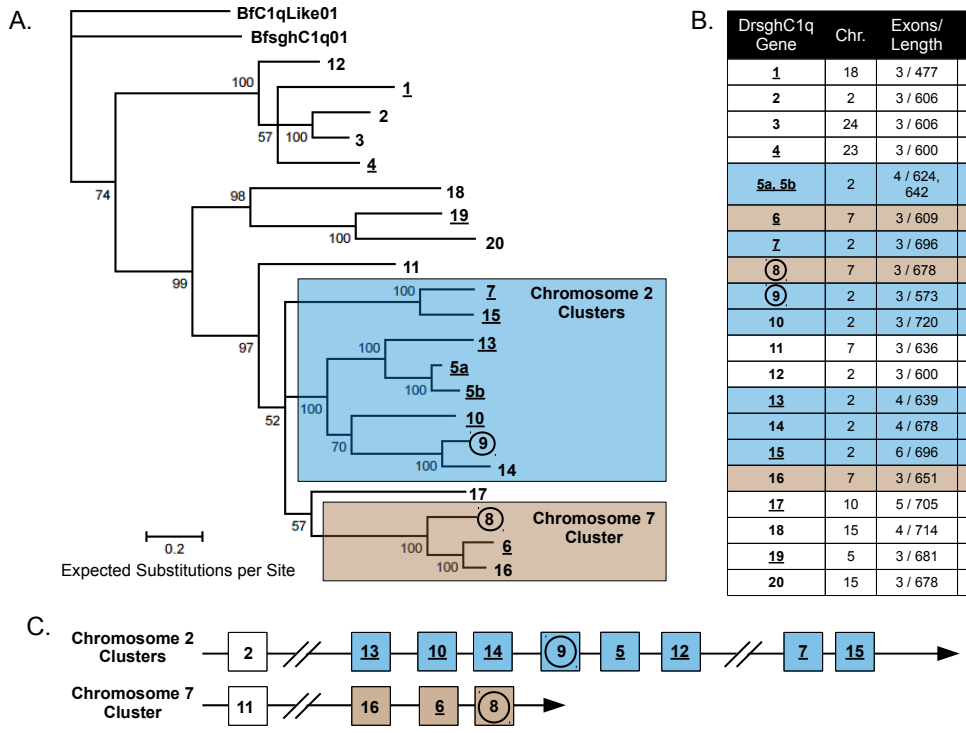
the genome. For example, the formal family name of Cbln1 [NCBI: gi50540102] is *DrsghC1q01*, the Dr is for *Danio rerio* and will not be a necessary part of the gene/protein name for the remainder of this text; *sghC1q01* will suffice.

The degree of amino acid sequence identity among the *sghC1q* proteins ranges from 23% to 95%. Despite this wide range of identity, the family members maintain a preserved predicted ten stand β-barrel configuration (Figure 3.1). To illustrate the relationships of these genes, a Bayesian codon phylogram was chosen as it can take into account prior distributions related to both sequential and biochemical relatedness of amino acids and codons [Ronquist and Huelsenbeck, 2003] [Anisimova and Kosiol, 2009]. The ensuing phylogram (Figure 3.2a) has three groups of proteins that clade according to their location on the same chromosome. Two of these groups are on chromosome two but occur at distant reaches of that chromosome (Figure 3.2c). Within the clustered regions on the chromosomes, the

sghC1q genes can be found close together and surrounded by pseudo-genes copies or gene-fragments of themselves (not shown). The relatedness of the clustered genes on these chromosomes and this arrangement of gene-fragments and pseudo-genes are indicative of intrachromosomal gene duplication events [Peatman et al., 2007, Bennetzen, 2007]. Additionally, there has been almost no mention of splice variants in this gene family. Only *sghC1q05* appears to have an immediately identifiable splice variant in the form of an alternate donor site within the first exon. It has been observed that while alternative splicing may occur in more than 50% of mammalian genes, it is less frequent among genes that have been recently duplicated [Su et al., 2006]. Barely any of the twenty genes in this family (in zebrafish) appear to undergo alternative splicing, thus it is likely that the duplications that created this family are relatively recent. In addition, polymorphisms have been noted in certain EST datasets for these genes making the need for careful analysis of the different sequences even greater.

3.4.2 Differential expression in response to infection and during early development

Experiments to study the potential differential regulation of the *sghC1q* genes were conducted in two ways. Firstly, to determine which (if any) of the identified *sghC1q* genes would be transcribed during an acute infection adult male zebrafish were injected intraperitoneally with 3,500 colony-forming units of the aquatic bacterial pathogen *Streptococcus iniae* (*S. iniae*). Only adult males were used in an effort to reduce potential gender specific variation in transcription. The fish were euthanized at 12 and 24 hpi, as was done in previous experiments [Gerwick et al., 2007], and the livers removed for subsequent RNA extraction and qRT-PCR analysis. Of the twenty *sghC1q* genes examined, only the transcripts from the genes corresponding to *sghC1q08* and *sghC1q09* were significantly ($p<0.05$) up-regulated at 23- and 13-fold respectively by 24 h post-infection (Figure 3.3). Previous studied *sghC1q05a* (*C1q-like*) indicated the presence of the transcript at 12 and 24 h post infection in embryos, however no significant difference between the treatment and the control fish were seen. (p-values of 0.19 and 0.90).



DrsghC1q Gene	Chr.	Exons/ Length	Expression
1	18	3 / 477	Development
2	2	3 / 606	
3	24	3 / 606	
4	23	3 / 600	Development
5a, 5b	2	4 / 624, 642	Development
6	7	3 / 609	Development
7	2	3 / 696	Development
8	7	3 / 678	Infection
9	2	3 / 573	Infection
10	2	3 / 720	
11	7	3 / 636	
12	2	3 / 600	
13	2	4 / 639	Development
14	2	4 / 678	
15	2	6 / 696	Development
16	7	3 / 651	
17	10	5 / 705	Development
18	15	4 / 714	
19	5	3 / 681	Development
20	15	3 / 678	

Figure 3.2: Phylogeny and chromosomal clusters of the *sghC1q* genes in Zv9. This figure illustrates the clustered nature of the *sghC1q* genes, particularly on chromosomes two and seven. The circled genes indicated up-regulation during inflammation and the underlines genes indicated transcription during early development. A.) The phylogram depicts the evolutionary relationships of this family with shaded portions illustrating the chromosomal clusters. Bf indicates *Brachiostoma floridae* which was used as the out-group for the analysis. B.) Table of all of the chromosomal locations of the *DrsghC1q* genes, complete with their exon counts, lengths, and expressions found in this study. C.) Graphical representation of the chromosomal clusters on two and seven, illustrating the clustered nature of the clades even apart from other clustered clades on the same chromosome (in the case of chromosome 2).

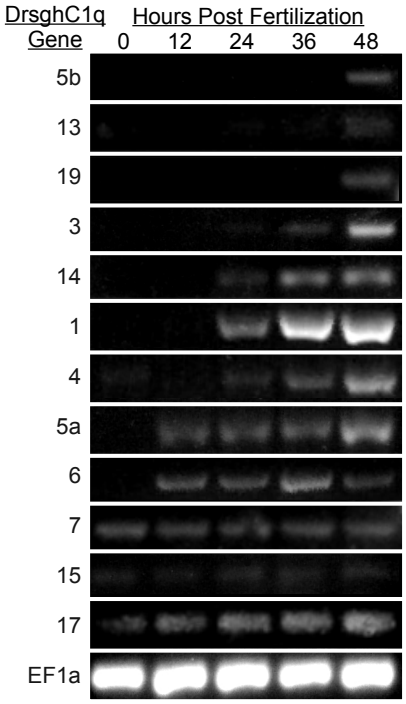


Figure 3.3: *DrsghC1q* expression during early development of *D. rerio*. A series of electrophoresis gels depicting PCR amplification performed on cDNA obtained from reverse transcription of RNA taken 0, 12, 24, 36 and 48 h post fertilization (HPF) of zebrafish embryos. Gene *ef1α* was used as a reference gene and hepcidin as a negative control (not shown). Of the twenty genes examined, eleven *sghC1q* genes were expressed, one of which showed expression of both of its alternative splice variants (*sghC1q05a* and *sghC1q05b*).

These results neither confirm nor discount previous studies [Mei et al., 2008b] that found this gene to reach its maximum up-regulation at 4 h post infection in zebrafish embryos. Regulation of gene *sghC1q05b* was similarly inconsequential yet notably different from its alternate splice variant. Gene *sghC1q01* exhibited significant down-regulation 24 h post infection. The expression patterns of the other *sghC1q* genes were highly variable between individuals and generally exhibited down-regulation (not shown). Due to the high sequence identity of the *sghC1q* transcripts, all amplicons were verified by direct sequencing to ensure that no cross priming occurred during the PCR reactions.

Secondly, to examine the potential absence or presence of the identified *sghC1q* transcripts during early development, zebrafish embryos were collected at 0, 12, 24, 36 and 48 h post fertilization; followed by RNA extraction and semi-quantitative PCR. Of the twenty genes examined, a total of eleven *sghC1q* genes (1, 3, 4, 5a/b, 6, 7, 13, 14, 15, 17, 19) exhibited expression within 48 h of fertilization (Figure 3.4). Gene *sghC1q05* has two splice variants, both of which were expressed during development though at different times. *sghC1q* genes 7, 15, and 17 were transcribed throughout the first 48 h of development and may represent maternally transferred transcripts. This correlates in function with a finding in goldfish [Mei et al., 2008a] where a C1q-like protein (CagOC1q-like) was discharged from maternal ovaries into egg envelopes. Genes *sghC1q05a* and *sghC1q06* begin to be expressed starting around 12 h post fertilization, corresponding to the early segmentation period (6-somite). This finding matches previous observations of *sghC1q05a* (C1q-like) by Mei and colleagues [Mei et al., 2008b]. Interestingly, *sghC1q05b* exhibited a different pattern from its splice variant. Genes *sghC1q01* and *sghC1q04* (which correspond to cbln1 and cbln4 respectively) as well as *sghC1q*14 appear to begin transcription around 24 h post fertilization corresponding with the transition to the pharyngula period. Genes *sghC1q*5b, 13, 19, and 3 are transcribed at 36 to 48 h after fertilization. Hepcidin, a known acute phase protein was used as a negative control (data not shown) because this gene is expected to show little or no expression, unless the embryos were subjected to bacterial infection [Lauth et al., 2005, Gerwick et al., 2007]. The *ef1α* gene was used as the reference gene and it exhibited constant expression for all of the embryonic time points collected. Again, due to the high sequence identity found among the *sghC1q* transcripts, all amplicons were verified by direct sequencing to ensure that no cross priming occurred during the PCR reactions.

Expression patterns in all instances did not appear to correlate with chromosomal location or any other basic characteristics of the genes (Figure 3.2). An interesting finding is that the *sghC1q* genes were either transcribed during early development or during the response to infection but never during both of these two physiological conditions. This observation contrasts with findings in mice

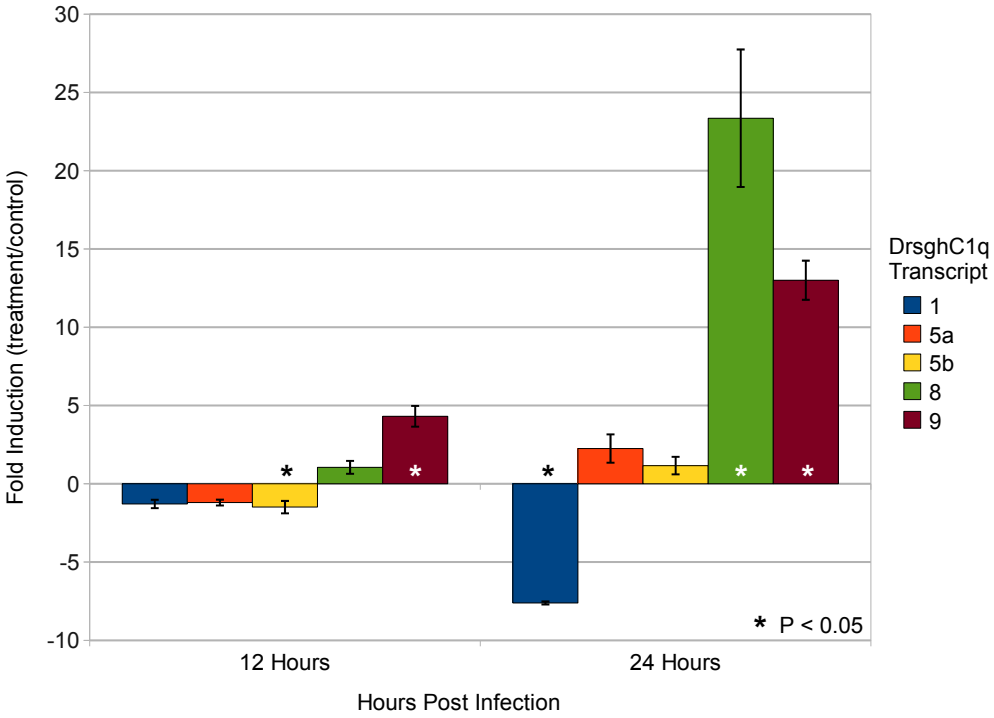


Figure 3.4: *DrsghC1q* expression during infection of *D. rerio* with *S. iniae*. Fold induction results from qRT-PCR on non-infected vs. infected liver RNA at 12 and 24 h post- infection with*S. iniae*for five genes of interest. Twenty *DrsghC1q* genes were found in the zebrafish genome and qRT-PCR was performed on all them but only *sghC1q01* (blue), *sghC1q05b* (yellow), *sghC1q08* (green), and *sghC1q09* (red) exhibited significant regulation during these times (data from the other genes not shown). Hepcidin (not shown) was chosen as a positive control and it was up-regulated four and forty fold at 12 h and 24 h respectively. Only *sghC1q09* exhibited up-regulation at 12 hours post infection while the others exhibited regulation at 24 hours (all up-regulation except for Cbln1). * indicates significance at a p-value <0.05.

[Stevens et al., 2007] and in zebrafish [Mei et al., 2008b] where C1q and *sghC1q05a* do appear to function during both of these physiological conditions. We postulate that the zebrafish *sghC1q* gene duplications could have allowed the genes to become subfunctionalized to the point of participating in one or the other of the two physiological responses. Alternatively, the ability for *sghC1q* genes to act during both conditions may have simply evolved in other organisms separately. Genes *sghC1q05*, 8 and 9 that were upregulated during the inflammatory response are now being further studied to determine what function the encoded proteins have during the inflammatory response.

3.5 Conclusions

The *sghC1q* genes have radiated intrachromosomally in the zebrafish genome, primarily on chromosomes two and seven. The expression patterns found in this study for the *sghC1q* genes do not correlate with the chromosomal locations of these genes, meaning that the functions of these genes likely changed in a manner unrelated to their evolutionary radiation. We have shown that the *sghC1q* genes identified are expressed during the innate response to infection and/or early development. These findings, coupled with the expansive radiation of these genes in zebrafish, lead us to hypothesize that the dual functionality has been lost in favor of subfunctionalization. Pairs of *sghC1q* genes that display the same temporal expression pattern can be found during early development as well as during the response to infection. These co-expressed genes may operate in concert, forming multimeric protein complexes, much like C1q or Cbln [Bao et al., 2005]; however, this remains to be examined.

3.6 Acknowledgements

This publication was prepared by Tristan M. Carland and under NOAA Grant #NA04OAR4170038 California Sea Grant College Program Project #R/A-124, through NOAA's National Sea Grant College Program, U.S. Dept. of Com-

merce; and was supported in part by the California State Resources Agency. The statements, findings, conclusions and recommendations are those of the author(s) and do not necessarily reflect the views of California Sea Grant or the U.S. Dept. of Commerce. This work was also supported by the University of California at San Diego Academic Senate and the Scripps Institution of Oceanography Graduate Department. The authors would also like to thank Francisco Villa, Monica Brunneto, Shumpei Maruyama, and Winnie Trieu, for assistance with the zebrafish colony and experiments.

Chapter 3 is a full reprint of the publication: Carland, T. M., Locke, J. B., Nizet, V., and Gerwick, L. (2011). Differential expression and intrachromosomal evolution of the sghC1q genes in zebrafish (Danio rerio). Developmental and Comparative Immunology, (In press), with permission from all coathors.

Chapter 4

EST Keeper: a Flash based web-tool for extracting complete and non-redundant ORFs from BLAST alignment sequence hits

4.1 Summary

One of the most basic and common bioinformatics tasks performed today is a BLAST alignment search. When using BLAST to search for sequences, sorting through the results can be very time consuming as there are likely to be redundant and partial sequences among the hits. EST Keeper is a user-friendly tool that takes a set of nucleotide BLAST sequence hits, assembles the sequences into contiguous sequences (contigs) using CAP3, searches for open reading frames (ORFs) among them and then returns the non-redundant set of ORFs with defined similarity to the original BLAST alignment input sequence. A good example of when to use this tool would be a search for how many homologs of a particular gene (or gene family) exist within a given genome. The user would perform a BLAST alignment of a representative sequence against the database available for that organism and use the BLAST results (and BLAST query sequence) as input for this tool. The

output from this tool is a FASTA file containing the non-redundant set of ORFs from within those sequences that meet criterion for size and similarity to the query sequence.

4.2 Introduction

The Basic Local Alignment Search Tool (BLAST) [Altschul et al., 1990] is the workhorse program for sequence identity searches today but sometimes the results generated from one of these searches can be difficult to manage due to the large number and redundancy of the resulting "hits". At the National Center for Biotechnology (NCBI) website (www.ncbi.nlm.nih.gov) BLAST alignments can be performed against several databases including NR (non-redundant), EST (expressed sequence tag), and chromosome (genomic), among many others. Sequence data are constantly pouring into these databases which results in repeated sequences. Some of these repeats are necessary and some can be avoided and must be purged from individual sets of results. Additionally, NR contains some very long sequences, such as sequenced BAC (bacterial artificial chromosome) clones and assembled linkage groups, which might not be desirable to the user in their entirety depending on the situation. Dealing with these issues can lead to confusion and take additional time to complete, particularly for a task such as finding the number of homologs of a gene or gene family in the genome of an organism. Problems especially arise when trying to identify the number of novel genes among results from a BLAST alignment against an EST database, as there will necessarily be many copies of the same EST sequence due to the nature of EST sequencing. Indeed, this problem is well noted and has been addressed in some cases by The Gene Indices Project (TGIP) [Quackenbush et al., 2001], whose website functions as a repository for pre-processed data similar to what EST Keeper generates promptly for new (and much smaller) data sets.

The EST Keeper software tool was constructed to find the non-redundant set of complete ORFs among fresh BLAST results by using one user-friendly program on one simple website. It is presented as a free web-service. Open-source

code and instructions for installation are also available.

4.3 Implementation

EST Keeper was designed to take an assortment of shorter (less than 10,000 bp) and longer nucleotide sequences in FASTA format with a limitation of 30 MB (megabytes or megabases) of total input. At its core, EST Keeper is a pipeline of two freely available open source tools and a custom ORF finding tool:

CAP3 assembles contigs of given sequence [Huang and Madan, 1999]

- Using default parameters
- Output viewable in a custom made parser

findORF.pl finds open reading frames (ORFs) in a given sequence (or sequences)

- Returns only complete (start-stop) coding regions
- Standard codon table
- Output is directly comparable to NCBI ORF-Finder
- Custom-made; see Appendix A.2

BLAST compares sequences, used in this instance for removing completely unrelated sequences [Altschul et al., 1990]

- tBLASTn or BLASTn depending on query

The CAP3 parameters were left at their defaults to allow for a broad use in dealing with both pre-compiled sequences and raw ESTs. The Perl script findORF.pl was made to generate complete ORFs from eukaryotic sequences, though further options (bacteria, mitochondrial, etc) are a consideration for future development. This custom ORF finding script was written instead of using freely available programs as most of those remove the stop codons from the resultant ORFs (a feature the authors decided against). BLAST parameters are set to default as the utility of BLAST in this event is not to find similar sequences among others

of near similarity but simply to distinguish between random/unrelated ORFs with very poor e-values and those that have already been found using BLAST. Thus, the default cutoff is relatively low.

Beyond these tools, EST Keeper contains other custom Perl scripts that manage the data. EST Keeper was written in three different languages, each with its own purpose. Actionscript (Flash) is used for the user interface and calls PHP web-service pages that in turn run the applications or Perl scripts. In this manner the user is confronted only with a basic interface where input options are presented when needed.

To begin, the user uploads a FASTA file that contains as many sequences as desired and that will fit within the 30 MB file limit. The first step is to remove any redundant sequences and this is accomplished with CAP3. CAP3 will require an excess amount of system memory for any sequences longer than 10,000 bp; as such these longer sequences are divided out for preprocessing. The longer sequences are reduced in size substantially by keeping only the predicted ORFs (scanned for start-stop segments using the standard codon table), and are then checked with BLAST alignments against a user supplied sequence to make sure they are the desired ORFs from those sequences. This subset is then combined with the shorter sequences and the entire set is assembled using CAP3 (with default parameters) to both remove redundancy and further compile any incomplete sequences. The results are processed through the custom made ORF finder and a BLAST check is performed, similar to what was done with the longer set of sequences. The results are then put through another round of CAP3 $\rightarrow$ findORF $\rightarrow$ BLAST to ensure that no duplicate ORFs remain (Fig. 1). The final output file contains the non-redundant set of complete ORFs and other output files are available after each step for potential troubleshooting.

4.4 Summary and future direction

This program works quickly and effectively at condensing BLAST alignment results into non-redundant ORF sequences. The resulting data are tailored

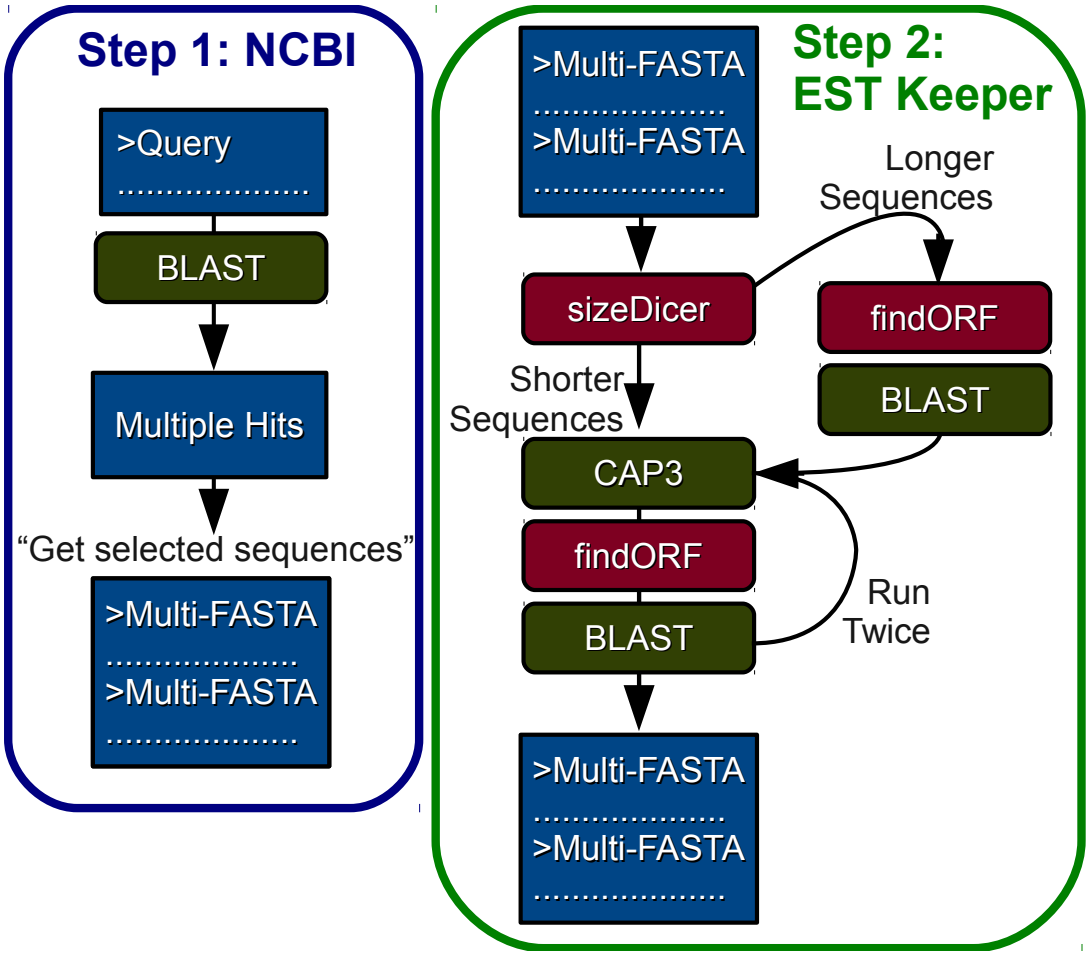


Figure 4.1: The EST Keeper program takes a FASTA file (preferably from NCBI-BLAST output) as input and returns another as output. After user input of a file, the file is cut into longer and shorter sequences as to not overload CAP3, the longer of which are scanned for ORFs followed by a BLAST alignment against a given sequence and the hits are recombined with the shorter sequences. The results are then put through a pipeline of CAP3, findORF, and BLAST twice before final results are generated.

to phylogenetic analyzes of protein coding sequences and can also be helpful for quickly assessing the number of novel copies of a gene or gene family in a given genome of an organism. This program has been successfully tested on data from many organisms (including mammals, fishes, worms, etc) and the results were found comparable to subsets of data found by the Gene Indices Project.

EST Keeper will be maintained and updated for improvement although its functions are not expected to greatly increase as they are considered most modular at this time. Future improvements will include more user defined parameters and a function to deduce the number of complete replications and locations of the condensed sequences in a given genome.

4.5 Acknowledgments

T.M.C. would like to thank Sheila Podell, William Shipman, members of the Gerwick laboratory, and family members.

Funding: This work was supported by the Ledger Benbough Foundation to L.G., the NSF program for Ocean Technology under grant OCE-0728305, and the Scripps Institution of Oceanography Graduate Department.

Chapter 4 is a modified version of a manuscript in preparation for submission under the title "EST Keeper: a Flash based web-tool for extracting complete and non-redundant ORFs from BLAST alignment sequence hits" with permission from coauthor Dr. Lena Gerwick.

Chapter 5

Conclusions and perspectives

5.1 Synopsis

The secreted globular head C1q (sghC1q) genes are a family of genetic loci encoding a signal peptide followed by a complement component 1q motif (Figure 2.2) [Carland and Gerwick, 2010]. Their namesake, C1q, is an established link between the innate and adaptive immune systems with additional functions during neuronal development, cellular chemotaxis, adhesion, and differentiation [Nayak et al., 2010]. The globular domain of C1q (gC1q) is a prevalent domain throughout the metazoa that gives rise to all of the C1q domain containing (C1qDC) genes that radiate from prokaryotes (notably *Bacillus sp.*) to the recently emerged vertebrates; cellular ghC1q (cghC1q) appearing first in *Bacillus* bacteria (not by horizontal gene transfer); secreted ghC1q (sghC1q) appearing later in the nematodes; C1q-like genes (with an added collagenous portion or cC1q) appearing in worms, but not becoming common until the vertebrates [Carland and Gerwick, 2010]. The prevalence of this domain is likely due to its versatile structure that enables binding to a diverse range of ligands [Ghai et al., 2007].

There are seven resolved crystal structures of C1q available revealing a ten-stranded β-sandwich of two five-stranded anti-parallel β-sheets. The first two are the first elucidations of the gC1q structure [PDB:1C28] [Shapiro and Scherer, 1998] [PDB:1PK6] [Gaboriaud et al., 2003] and the next five are C1q-like proteins or C1q

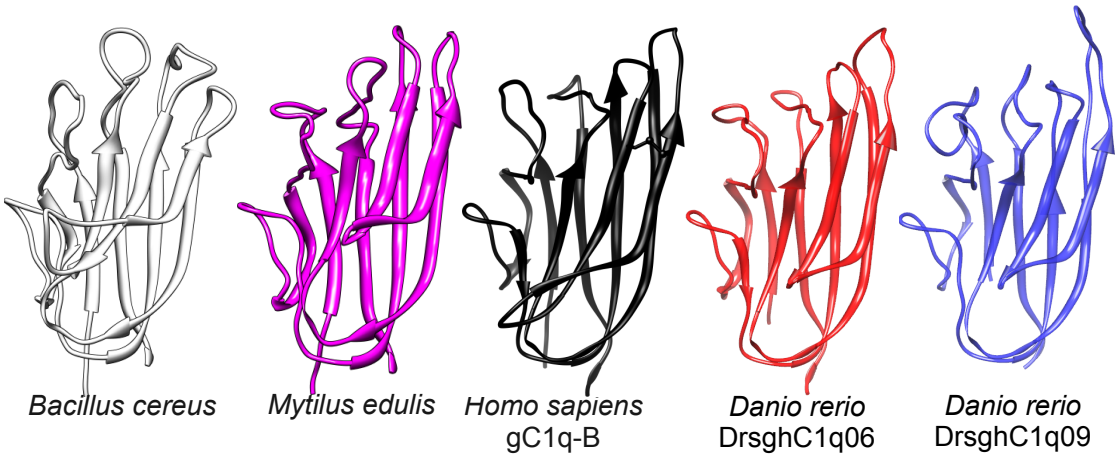


Figure 5.1: Protein modeling of five gC1q homologs from a bacterium *Bacillus cereus* (white), a mussel *Mytilus edulis* (pink), a mammal *Homo sapiens* (black), and a fish *Danio rerio* (red and blue). Modeling done with the M4T server [Rykunov et al., 2008] and graphics created with the Chimera viewer [Pettersen et al., 2004].

bound to a ligand: ACRP30 [PDB:1C3H], gC1q bound to phosphatidyl-serine [2JG8] [Padassi et al., 2008], Emilin-1 [PDB:2OII], gC1q bound to deoxyribose [PDB:2WNV] [Garlatti et al., 2010], and gC1q bound to heparan [PDB:2WNU] [Garlatti et al., 2010].

The many C1q proteins found within the metazoa and within the different genomes exhibit great sequence variation, for example the zebrafish genes can have as low as 20% amino acid (AA) identity to one another, yet are computationally predicted to maintain the aforementioned ten-stranded β-sandwich (Figure 5.1) [Carland and Gerwick, 2010] [Carland et al., 2011].

Genes with the conserved gC1q structure, C1qDC genes, are heavily replicated within certain genomes; 32 open reading frames (ORFs) encoding C1qDC proteins have been found in humans [Tom Tang et al., 2005], at least 52 exist in zebrafish [Mei and Gui, 2008], and as many as 75 have been found in amphioxus [Huang et al., 2008]. Twenty sghC1q ORFs exist within the current (Zv9) zebrafish genome, of which two are significantly (p-value<0.05) up-regulated within the first 24 h of the response to an infection with *Streptococcus iniae* and eleven

others are expressed during the first 48 h of development. They appear mostly on chromosomes two and seven, and are predicted to have radiated to those positions by intrachromosomal duplication events more recent than the whole genome duplication events of the early teleost lineage. [Carland et al., 2011].

The functions of the sghC1q proteins remain largely unknown. Because they are related to mammalian C1q they were hypothesized to have an immune function. This was the hypothesis tested in this thesis when the transcription of these genes was measured during the response to infection. Mammalian C1q has many other functions as well, particularly during development, and the sghC1q proteins might also. This was the hypothesis tested in this thesis when the transcription of these genes was assayed during early development. During both physiological conditions, zebrafish sghC1q genes were transcribed. Transcriptionally, the sghC1q gene family is in this case analogous to mammalian C1q.

5.2 The alarming sghC1q

Mammalian C1q is the archetype initiator of the complement system, both as a link to the adaptive immune system and through its own binding to various ligands [Nayak et al., 2010]; beyond which it has been found to function as a sort of pruning molecule during neuronal development, removing improperly formed optic neurons [Stevens et al., 2007]. Under the Danger Model (detailed below), this developmental activity could also be considered an immune function. The *Danger Model* philosophy of the immune system dictates that the primary concern of the immune system is that which may cause harm to the organism, as opposed to simply recognizing that which is "non-self" [Matzinger, 2002]. In one sense the Danger Model only formalizes the semantics of what is and is not part of the immune system by broadening the definition to include signaling molecules that do not directly handle pathogens; importantly it allows such an involvement as C1q pruning optic neurons to also fall under the mandate of the immune system. Perhaps most importantly it dictates that the immune system will allow many microbes to exist within an organism as useful symbionts because they simply are

not “Dangerous”.

Given that sghC1q genes are transcribed in response to an infection, under the Danger Model, those genes could be considered immune genes. Other sghC1q genes in zebrafish are transcribed during early development. Should those sghC1q genes be found to be acting in a manner at all similar to C1q during development, the same classification could plausibly be extended to them.

To elicit an immune response, when studying infections or administering a vaccine, the “immunologists dirty little secret” was adjuvant [Janeway, 1989]. Adjuvant, often an oil emulsion and small dose of toxin, was used to damage the tissue to further simulate a standard infection to which the immune system (particularly the innate) would respond. In the Danger Model, this would be considered a danger or alarm signal [Matzinger, 2002]. When cells and tissues are damaged, components of them (free collagen, DNA, etc) that are not normally found outside of a cell (or organelle) are found floating around in the plasma (or cytoplasm); the immune system will sense these and elicit a response. One protein complex that can bind to DNA, and is found in high concentrations in peripheral blood serum, is C1q [Garlatti et al., 2010].

C1q is a danger signal. It can bind to antibodies (specific danger signals), many pathogen ligands, and a suite of altered self ligands indicative of cellular damage. C1q can lead to the membrane attack complex (MAC) that lyse cells or be recognized by intermediary receptors gC1q-R (receptor for the globular portion of C1q) or cC1q-R (receptor for the collagenous portion of C1q) that are then recognized by immune response cells such as macrophages, immature monocytes, and dendritic cells. Indeed, monocyte derived dendritic cells have been shown interact first with gC1q and then with cC1q as they differentiate. [Nayak et al., 2010] [Hosszu et al., 2010]

A teleostean C1q has been found in zebrafish with conserved synteny with mammals that interacts with zebrafish IgM, human IgG and IgM [Hu et al., 2010]. Hu and colleagues go on to hypothesize that the C1qA, C1qB, and C1qC genes may all be duplications of an original C1qB. This fits well with our findings that the ghC1q genes far out-date C1q, and that they are most akin to C1qB (over A

or C) [Carland and Gerwick, 2010].

5.3 Why so many sghC1q genes in zebrafish?

If zebrafish has classical C1q genes, why does it have so many sghC1q genes? The answer to this question is yet to be elucidated but hypotheses can be made. The null hypothesis in this case would be that the sghC1q genes contain the gC1q domain simply by chance and that they have no function and may not even be transcribed. These studies have disproved the null hypothesis by showing that 13 of the 20 sghC1q genes in zebrafish are transcribed during early development and the response to infection. An alternative hypothesis is that the sghC1q genes in zebrafish may have functions that are not related to the immune system, but share the gC1q domain simply because it is a versatile motif for ligand binding. The two zebrafish sghC1q genes that were up-regulated during an infection (DrsghC1q08 and DrsghC1q09), do not support that hypothesis. Granted though, we do not know if the proteins from these genes are translated, much less take part in the resolution of the infection, though it probably isn't overstepping to assume that they are translated and functional studies with those proteins are underway.

A second alternative hypothesis to be presented here concerns the matter of ligand binding diversity. In mammals, as mentioned above and throughout this tome, C1q binds to a host of ligands including pathogen associated molecular patterns and a host of modified self ligands (DNA, free collagen, etc). The list of C1q ligands is further augmented by opsonizing antibodies that are specifically generated for a particular ligand, and to these antibodies C1q may also bind to begin the complement cascade or signal an immune effector cell. In a sense, the ability of mammalian C1q to bind to antibodies implies that it has a near infinite ligand binding repertoire so long as the adaptive immune system (specifically B-cells) remains active. Zebrafish does have C1q and does have an antibody generating system but the function of these systems has not been fully resolved. Thus far, fishes are known to have a reduced set of antibodies (compared to mammals) to which zebrafish C1q has only been shown to bind to one of (IgM), and teleosts

lack class (or isotype) switching. Might the sghC1q proteins be compensating for a lacking adaptive immune system? Zebrafish are known to habitate swamps fraught with pathogens [Westerfield, 2007]. Another factor is that teleosts have an additional antibody isotype (IgT). Could the sghC1q proteins interact with IgT? Possibly, and this hypothesis should be investigated.

The sghC1q proteins could create a larger repertoire of danger sensing signals by binding to pathogens, damaged tissues, or maybe even antibodies; and then interacting with the well conserved gC1q-R receptors on immune effector cells. An increased ability to sense cellular damage, caused by anything (pathogens or injuries), would be of great benefit to the immune system as defined by the Danger Model of immunity. The sghC1q genes are more diverse in fishes than in mammals, and appear to be even more diverse in amphioxus [Huang et al., 2008]. Other prototypic chordates (echinoderms and tunicates) do not appear to have great diversity in their sghC1q genes but instead appear to have developed duplicates of the other initiating complex of the complement system, the mannose binding lectin [Endo et al., 2006]. The sghC1q genes are far more ancient than the prototypic chordates, but may have expanded in their diversity around the same time as the foundations of the adaptive immune system. Alternatively, the sghC1q genes could be very diverse in other invertebrate eukaryotes and we have to sequence those organisms to answer this question.

5.4 Diversification of sghC1q genes

Teleosts are the most diverse group of vertebrates, owed to whole genome duplications and intrachromosomal gene duplications, they are known to have evolved more rapidly than the rest of the vertebrates [Ravi and Venkatesh, 2008]. The number of sghC1q genes found in zebrafish (20) is greater than the four generally found in mammals, and preliminarily appears to be greater than other sequenced fishes (medaka and puffer) [Carland and Gerwick, 2010]. Why (and how) are there so many sghC1q genes in zebrafish?

5.4.1 Duplications and accelerated evolutionary rates

Teleosts are theorized to have been through three whole genome duplications, one at the onset of the jawless fishes, a second at the radiation of the cartilaginous fishes, and a third at the radiation of the teleosts [Dehal and Boore, 2005] [Ravi and Venkatesh, 2008]. Two of these would have affected mammals as well. These proposed whole genome duplication events (or polyploidization), lead to an entire second set of genetic material and have experimentally been shown to be more transmissible to future generations because the organisms with them can become reproductively isolated (reviewed by [Levasseur and Pontarotti, 2011]). Genes from whole genome duplications are found to tolerate ten times more amino acid changes than a single copy gene [Wagner, 2008]. The fate of duplicate genes is a topic of great debate, though it is generally accepted that soon after a duplication event there is a period of relaxed constraint allowing evolution that may persist for a few million years (species dependent) before genomic contractions begin [Lynch and Conery, 2003].

Teleosts are also theorized to be evolving at a higher overall rate than their jawless, cartilaginous, or mammals relatives [Ravi and Venkatesh, 2008]. This is evidenced by a high level of gene-linkage disruptions and highly conserved non-coding elements (CNEs) that are known to act as tissue specific enhancers being mutated nearly beyond recognition in certain fish lineages [Lee et al., 2011]. These studies have also found a higher rate of evolution (hypothesized to be adaptive evolution) specifically in zebrafish as compared to other teleosts. Additionally, it has been found in studies of gene families that immune gene families tend to expand or contract at a rate significantly above neutral [Hahn et al., 2005]. Smaller scale evolutionary events leading to interchromosomal rearrangements have been found in zebrafish when compared to medaka [Kasahara et al., 2007]. Duplications of chromosomal segments are found to be somewhat common between human individuals and are hypothesized to be caused by cellular stress that may cause an increase in gene copy number of a single cell (through non-homologous repair of a double stranded break) as an adaptive measure [Hastings et al., 2009].

This last means of diversification is an especially interesting concept given

the nature of stress and danger to which immune cells respond (some of which synthesize sghC1q). Could liver hepatocytes or macrophages have private collections of duplicated immune genes (especially highly replicated ones like sghC1q genes) that they diversify to greater extents under the stress of inflammation?

5.4.2 Alternative splicing

A method of diversification that does not require the same variety of genomic evolution, and strains the old definition of a gene is alternative splicing. Before the central dogma (DNA $\rightarrow$ RNA $\rightarrow$ Protein) became the central dogma [Crick, 1958, Crick, 1970], the notion of "one gene, one protein" (then "one gene-one enzyme") arose [Beadle and Tatum, 1941]. This notion was ahead of its time and found to be an oversimplification as more than 70% of human genes are predicted to undergo alternative splicing [Su et al., 2006], a process that allows one gene to produce more than one protein (sometimes a dozen) [Black, 2003]. Even today this fact is considered a justification for the fact that vertebrates (especially humans) have far fewer genes than previously expected compared to other organisms [Pennisi, 2005]. This process is not widely known to affect the sghC1q genes, but is present in at least one instance among zebrafish [Carland et al., 2011].

5.5 Closing remarks

Through these studies we have increased the knowledge of the C1q domain containing genes throughout the metazoa and to a greater detail, the sghC1q genes within zebrafish. These findings have implications for the understanding of the function(s) and evolution of these genes and further our understanding of gene families and adaptive evolution within the zebrafish. Additionally, the bioinformatic tool developed for these studies will be useful for future use in similar studies.

Appendix A

Bioinformatic Scripts and Programs

The following appendix is a collection of some of the scripts and programs that were written during the course of the studies in this thesis. They have been selected from a much larger set as those that may be of most use, and are small enough to be reproduced in this format. Each script/program has been altered to fit within the required margins.

A.1 fixFasta.pl A beginner script that all bioinformatic programmers should have ready. It is used to prepare FASTA files for processing by removing extraneous lines/spaces, putting all DNA/RNA/AA on one line per sequence, and making all sequence code upper-case. It reads from STDIN and writes to STDOUT. The proper usage would be "cat your_old_file.fasta |perl fixFasta.pl >your_new_file.fasta". What this will do is to 'cat' or concatenate your fasta file (which means to print it to screen), the '|' will "pipe" those data into the script that is run by 'perl fixFasta.pl'. The output from that program is written (via '>') to the file on the end, even if there is a file already with that name.

A.2 findORF.pl A Perl script that will parse through a FASTA file (single or multiple sequences) and return all open reading frames that meet the input size specifications. It does not look for introns.

A.3 translate.pl A Perl script that will take a FASTA file of DNA sequence and translate it in its current frame to an amino acid sequence in an efficient manner.

A.1 fixFasta.pl

```
#!/usr/local/bin/perl
# removes spaces, cases and broken lines from any FASTA piped in
# Author: Tristan M. Carland (UCSD–SIO) October 13, 2005

my $line = ""; # creates an empty string variable
while(<STDIN>) # parses by line, as long as there is input
{
  chomp;     # chomp removes leading/trailing spaces and breaks
  if(/^>/) {  # if the line starts with >, meaning header lines
    prnt($line);       # call the prnt method below
    $line = "$_"."\n";  # $line is the current line plus a break
  }
  else {    # when not on a header line
    s/\d//g;    # replace all digits (0-9) with nothing
    s/\s//g;    # replace all whitespaces with nothing
    $line .= uc$_;  # append current line (uppercase) to $line
  }
}
prnt($line);  # when input is done, call prnt last time

sub prnt {
  $fastaSeq = @_[0];       # $fastaSeq is param variable 0
  unless($fastaSeq =~ /^$/) # ^$ is beginning next to end
  {
    print "$fastaSeq\n";    # print the seq, with a line break
  }
}
```

A.2 findORF.pl

```
#/usr/local/bin/perl
# Returns full open reading frames from seqs
# Author: Tristan M. Carland (Summer 2010, SIO-UCSD)
use warnings; use strict; use strict 'vars';  # a good practice

# input variables
unless (scalar @ARGV == 4)
{
  print STDERR "Usage: $0 InFile OutFile MinOrfSizeNT MaxOrfSizeNT\n";
  exit(0);
}

# input files
my $inFile  = $ARGV[0]; # file to get data from
my $outFile = $ARGV[1]; # file to print output to
open(INFILE, "$inFile")
  or die "Could not open input file $inFile,  $!\n";
open(OUT, ">$outFile")
  or die "Could not open output file $outFile,  $!\n";

# input size ranges
my $minSize = int($ARGV[2]);  # minimum orf size, in internal codons
my $maxSize = int($ARGV[3]);  # maximum orf size
  if( $minSize > $maxSize)
  {
    die "Minimum ORF size is larger than maximum, try again\n";
  }

# variables to work with, including regular expressions
my($header, $seq) = ("","");
my $start = qr/ATG/; # start codons
my $stop  = qr/TAA|TAG|TGA/; # stop codons

# non-stop codons - was trying an alternative approach
# my $coder = qr/TC\w|TAT|TAC|TGT|TGC|TGG|A\w\w|G\w\w|C\w\w/;

# pan through input file in the usual perl style
while (my $line = <INFILE>)
```

```
{
  chomp($line); # remove spaces and such

  if ($line =~ /^>/)  # if this is a header line
  {
    findORFs(); # look for orfs
    ($header, $seq) = ($line, ""); # clear the variables, start over
  }
  else  # else we are in mid-sequence
  {
    $seq .= uc$line;  # so add the new sequence to what we have
  }
}
findORFs(); # one more time after the last sequence
close OUT;
close INFILE;

# process a sequence (check for orfs, reverse compliment, try again)
sub findORFs
{
  # unless it's blank (beginning and end are next to each other)
  unless( $header =~ /^$/ )
  {
    # set the header to just the first ">word " from the old header
    if ($header =~ /^>+(\S+)\s*/)
    {
      $header = $1;
    }
    # if the header is still more than 20 chars
    if($header =~ /^>*(\S){20}(\S){+}\s/)
    {
      $header = $1."..."; # take first 20 and put ... on the end
    }
    $seq =~ tr/U/T/;    # change any 'U' to a 'T'
    orfCheck('+',1,$seq); # check for orfs in first positive frame
    orfCheck('+',2,(substr $seq, 1));
    orfCheck('+',3,(substr $seq, 2));
    my $qes = reverse $seq; # reverse the strand
```

```
    $qes =~ tr/ATGC/TACG/;  # find the compliment
    orfCheck('-',1,$qes); # check for orfs in first negative frame
    orfCheck('-',2,(substr $qes, 1));
    orfCheck('-',3,(substr $qes, 2));
  }
}
# manages the frames on a strand
sub orfCheck
{
  my $strand  = $_[0];
  my $frame = $_[1];
  my $search  = $_[2];
  my @starts  = ();
  my @stops = ();
  my ($begin,$end,$length) = (0,0,0);

  # seems we still need to remove formatting from the sequence
  $search =~ s/[\t\n\r\f\a\e]//gi;

  # regular expressions, find all start and stop codons in this frame
  while( $search =~ /\G(\w\w\w)*?($start)/g )
  {
    push(@starts,$-[2]);  # add them to the array
  }
  while( $search =~ /\G(\w\w\w)*?($stop)/g )
  {
    push(@stops,$-[2]);
  }
  # flip the array so that pop will remove the smallest
  @starts = reverse(@starts);
  @stops  = reverse(@stops);

  # continue the iteration as long as the array has elements
  while (scalar(@starts) > 0)
  {
    # take the last entry from the array (smallest)
    $begin = pop(@starts);
```

```
    # only occurs if start is less than last stop or 0
    next if ($begin < $end);

    while (scalar(@stops) > 0)
    {
      # take the last entry from the array (smallest)
      $end = pop(@stops);

      # don't need it if the end is less than start
      next if ($end < $begin);

      # save the length, the last codon counts too
      $length = $end - $begin +3;

      if( ($length>=$minSize) && ($length<=$maxSize) ) # size params
      {
        print OUT "\n>$header";

        # unnecessary if this is just repeating another ORF find
        unless ( ($header =~ /:ORF/) && ($begin+$frame == "1") )
        {
          print OUT ":ORF".$strand.$frame."(".$length.")";

          if($strand =~ /\+/)
          {
            print OUT ($begin+$frame)."-".($end+$frame+2);
          } else {
            my $le = (length $search);
            print OUT ($le-$end-2)."-".($le-$begin);
          }
        }
        print OUT "\n".substr($search, $begin, $length)."\n";
      }
      last;
    }
  }
}
```

A.3 translate.pl

```
#/usr/local/bin/perl
# translate.pl - Translates DNA to amino acids in the first frame
# Author: Tristan M. Carland (Summer 2010, SIO-UCSD)
use warnings; use strict; use strict 'vars'; # good practice

# input variables
unless (scalar @ARGV == 1)
{
  print STDERR "Usage: $0 InFile\n";
  exit(0);
}

# input files
my $inFile  = $ARGV[0]; # file to get data from
  open(INFILE, "$inFile")
    or die "Could not open input file $inFile,  $!\n";

# variables to work with, including regular expressions
my($header, $seq) = ("","");

# start codons
my $start = qr/ATG/;

# amino acid / non-stop codons
my $coder = qr/TC\w|TAT|TAC|TGT|TGC|TGG|A\w\w|G\w\w|C\w\w/;

# stop codons
my $stop  = qr/TAA|TAG|TGA/;

# pan through input file in the usual perl style
while (my $line = <INFILE>)
{
  chomp($line); # remove leading/trailing spaces

  if ($line =~ /^>/)  # if this is a header line
  {
```

```
    translate(); # look for orfs
    ($header, $seq) = ($line, ""); # clear the variables to start over
  }
  else  # else we are in mid-sequence
  {
    $seq .= uc$line;  # so add the new sequence to what we have
  }
}
translate(); # one more time after the last sequence
close INFILE;

# translate the sequence
sub translate
{
  unless( $header =~ /^$/ ) # unless beginning is next to end (blank)
  {
    # put a blank space after each codon
    $seq =~ s/(\S{3})/$1 /g;

    # translate each codon - this could take a few steps
    $seq =~ s/TTT|TTC/F/g;
    $seq =~ s/TTA|TTG|CT\S/L/g;
    $seq =~ s/TC\S|AGC|AGT/S/g;
    $seq =~ s/TAT|TAC/Y/g;
    $seq =~ s/TAA|TAG|TGA/*/g;
    $seq =~ s/TGT|TGC/C/g;
    $seq =~ s/TGG/W/g;
    $seq =~ s/CC\S/P/g;
    $seq =~ s/CAT|CAC/H/g;
    $seq =~ s/CAA|CAG/Q/g;
    $seq =~ s/CG\S|AGG|AGA/R/g;
    $seq =~ s/ATA|ATC|ATT/I/g;
    $seq =~ s/ATG/M/g;
    $seq =~ s/AC\S/T/g;
    $seq =~ s/AAC|AAT/N/g;
    $seq =~ s/AAG|AAA/K/g;
    $seq =~ s/GT\S/V/g;
    $seq =~ s/GC\S/A/g;
```

```
    $seq =~ s/GAC|GAT/D/g;
    $seq =~ s/GAA|GAG/E/g;
    $seq =~ s/GG\S/G/g;

    # remove the spaces
    $seq =~ s/\s//g;  # by replacing them with nothing

    # print it out
    print STDOUT $header."\n".$seq."\n";
  }
}
```

Appendix B

Automated identification of conserved intergenic regions in vertebrates via genomic comparisons with *Takifugu rubripes*

B.1 Abstract

The tiger puffer (*Takifugu rubripes*) was one of the first fish genomes to be sequenced and is renown for being an extraordinarily small genome for a vertebrate while still maintaining a comparable gene set. This logically means that it has reduced non-coding regions, making it an excellent base from which to launch a comparative genomic study to identify very important intergenic areas that may have been functionally constrained among vertebrates. In this study the latest genomes and cDNA libraries available for the tiger puffer, zebrafish (*Danio rerio*), frog (*Xenopus tropicalis*) and human (*Homo sapiens*) were utilized to identify intergenic regions flanked by genes with orthologous counterparts in each organism exhibiting the same order and orientation. The processing was done bioinformati-

cally with several new programs written to handle fresh data in a parallel pipeline fashion (meaning the data was passed between a series of programs). Preliminary results (proof-of-concept) found only four regions fully conserved but since then the program pipeline has undergone a major overhaul to improve the theoretical framework and has been redone in several programming languages instead of just Perl. Preliminary analysis found hundreds of the intergenic regions (not shown) but this must be revisited as three years have passed and the genomes have been recompiled and updated (most notably the zebrafish genome).

B.2 Introduction

The tiger puffer (*T. rubripes*) presently has the smallest genome of all vertebrates. This is largely due to a genetic compaction as it has proved to have a far greater gene density than any other vertebrate tested [Brenner et al., 1993]. These facts, combined with its position in the lineage of fishes as a more evolutionarily derived (specialized) species branching far from the radiation of mammals, have led to its selection as a base organism for comparative genomic predictions of genes and non-coding sequences [Roest Crollius and Weissenbach, 2005].

Phylogenetic footprinting, a technique for finding conserved genomic elements constrained by evolution via comparative genomics, has been supported by functional studies and is now considered a viable technique for finding a variety of important elements in a genome [Sandelin et al., 2004]. The fundamental concept is as follows: if an element has been conserved in the genomes of two species that parted long ago (over 450 million years ago in the case of human and fugu), the element may play a critical role in the organism to have been kept through such an evolutionary distance [Venkatesh et al., 2000]. This concept is supported by studies such as Woolfe *et al.* whom in 2004 [Woolfe et al., 2004] found 1,400 conserved intergenic elements by whole genome screens between humans and tiger puffers, many of which regulate vertebrate development.

In this study, the concept of evolutionary constriction of functionality was taken further by including the genomes of other organisms radiating between hu-

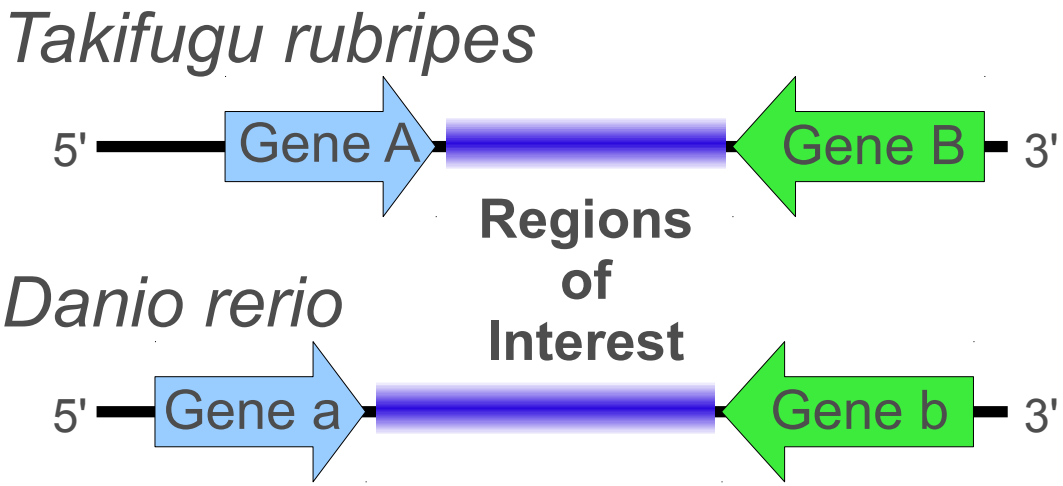


Figure B.1: Illustration depicting the Regions of Interest as being intergenic spacers between sets of orthologous genes in different genomes with conserved order and orientation.

mans and the derived tiger puffer, namely the less derived cyprinid zebrafish and an amphibian. The target areas of this study were intergenic regions flanked by genes found to have orthologues in each organism exhibiting the same order and orientation (Figure B.1).

These selection criteria ensured that the chosen sequences would not only be in areas of relative evolutionary stability, but also increased the chances that the intergenic regions selected would play a role in the transcription/translation of the flanking genes in such ways as micro-RNA or transcription factor binding sites. In future studies the desired intergenic regions can be tested for functional elements.

B.3 Materials and methods

B.3.1 Genomes and cDNA libraries

Whole genomes and full cDNA libraries of the chosen organisms were each obtained from the Ensembl website (http://www.ensembl.org). Table B.1 lists all the versions of each dataset from Ensembl at the time of download.

B.3.2 Employed programming languages

BLAST All BLAST (Basic Local Alignment Search Tool) alignments in this study were performed using the hardware accelerated Tera-BLASTTM (Active Motif Inc., Carlsbad, CA) implementation of the original BLAST algorithm [Altschul et al., 1990]. This algorithm utilizes specially designed and manufactured TimeLogic computer boards that are built specifically for performing BLAST calculations. This amplifies the speed of these calculations by a few orders of magnitude.

Perl The high level scripting and reporting language "Perl" was used for most of the string comparison portions of the pipeline due to the ease and the speed at which it can parse through large text files, mostly thanks to its regular expressions. [Wall et al., 2000]

XML The extensible markup language "XML" was used as a means of passing data from the Perl scripts to the Java programs. It was chosen for being a fast and standard way to abstractly store data in a customized flat file. [Bray et al., 2006] This standard format increases the universality of the pipeline.

Java The object oriented programming language "Java" (Sun Microsystems, Santa Clara, CA) was used for the central logic as it is object oriented in a very abstract sense, making it a good choice for creating data structures based on the nature of the data being processed. [Arnold et al., 2000]

MySQL The relational database management system "MySQL" (Sun Microsystems, Santa Clara, CA) was used for storage and manipulation of data processed by the Java portion of the pipeline. Its web-application oriented nature and ubiquitousness across most system architectures make it a reasonable choice for any program seeking wide distribution. It was particularly selected also for it's ease of interface with Java and PHP. [DuBois, 2008]

PHP The dynamic web page scripting language "PHP" (Personal Home Page) was used to create a dynamic report page for viewing the data. It was se-

lected for its dynamic capabilities and ease of use when dealing with MySQL. [Lerdorf et al., 2006]

B.3.3 Intergenic identification pipeline

Finding intergenic regions flanked by orthologous genes with conserved order and orientation required that the following steps be repeated for each genome (Figure B.2). First the exact locations of all putative genes, represented by the cDNAs, had to be determined as well as their order and orientation tracked. This was accomplished by BLAST alignment of the total cDNA library available for a given organism with its genome. The second step was to determine which genes in a genome had orthologs in the tiger puffer genome. Orthologous genes were determined by reciprocal BLAST alignments (alignment of one cDNA library with another and then back again). Genes were considered orthologous when their cDNAs exhibited a two-way best matching scenario (Figure B.2). The BLAST alignments took nearly thirty minutes for each set when using two TimeLogic accelerator boards simultaneously, and would have taken weeks on what is presently a state of the art quad-core machine.

All four sets of the above BLAST results were then used as input for *regScouter.pl* for consideration of two- way best matching, scaffold location and orientation information. This program would seek pairs of contigs that potentially held a region of interest and generate an XML file representing the data that was fed into *regHunter.java*. This program would load all the information into custom abstract objects (Scaffold, Airoi, GeneRegion, Loci) and finish the logic programming in an intuitive manner. If the conserved intergenic regions of interest were present, the data identifying them would be loaded into a MySQL database where it could be viewed by *tableView.php* through any web-browser.

Once this process was completed for each additional genome, the joined results could be viewed via *tableView.php*.

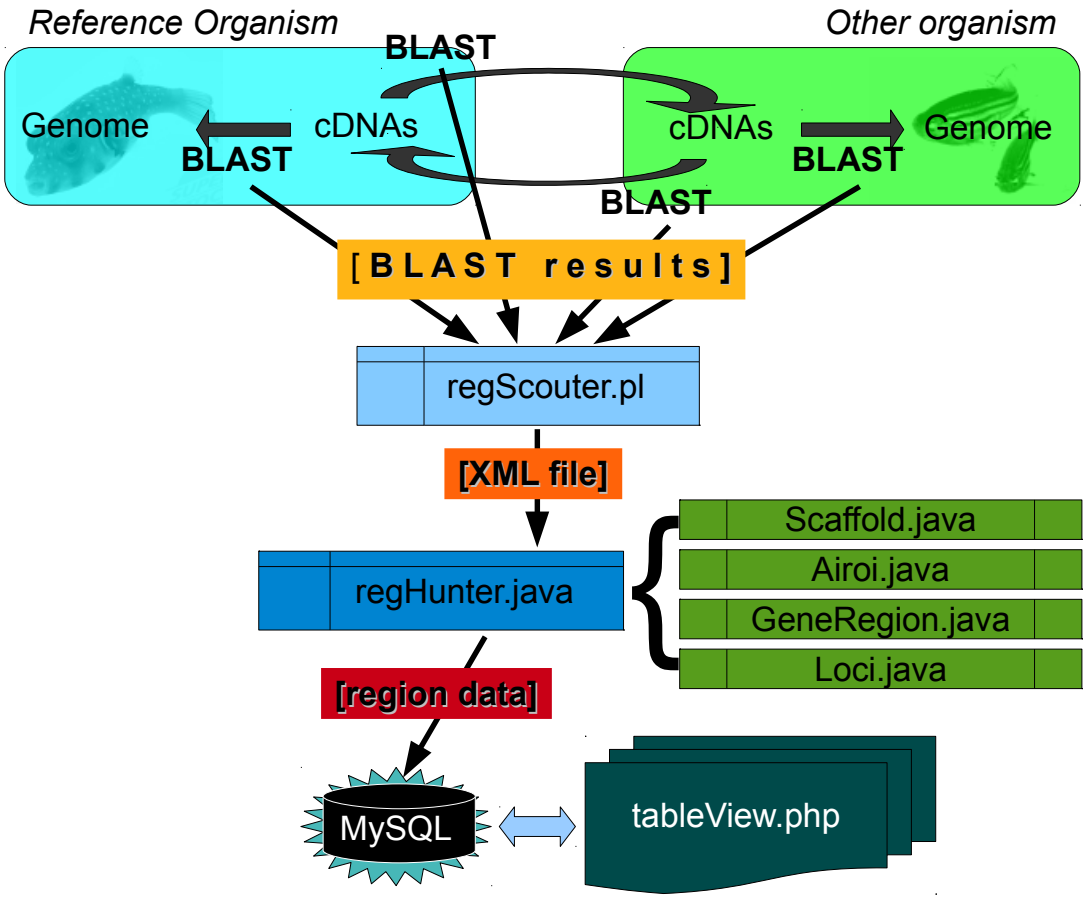


Figure B.2: Work-flow diagram of pipeline to find intergenic regions of interest.

B.4 Preliminary results

Presently the pipeline has been massively updated but not run since 2006. Many bugs were addressed and subsequent runs were of better quality though incomplete. The following is a reduced report of the previous findings to exhibit the kind of data that can come from the pipeline as it was devised. As with all programs, the results depend entirely on the input data so we will begin there.

B.4.1 Relative state of the input data

Publicly available sequence data for the genomes used in this project were of variable quality. The genomes and cDNA libraries from Ensembl illustrate the relative assembly status of each genome (Table B.1, columns 1, 2, 4, 5). Were a genome completely assembled it would only have one scaffold per chromosome, and the highest chromosome count in this study was 25 pairs. *T. rubripes* had a genome size of 393 megabases spread across 7,213 scaffolds, when compared to that of its close relative *Tetraodon nigroviridis* (7th version) it became clear that the construction of *T. nigroviridis* was not nearly as complete as *T. rubripes*. The *D. rerio* genome appears relatively well assembled (or closed) with a size of 1,626 megabases and only 6,653 scaffolds, *X. tropicalus* was not quite so with 1,511 megabases and 19,501 scaffolds, *H. sapiens* had 27,233 scaffolds but given the genome size (3,433 Mb) this figure was acceptable, and finally *Caenorhabditis elegans* and Ciona intestinalis were in relatively similar states with sizes of 101 Mb and 173 Mb and scaffold counts of 3,268 and 4,390 respectively.

B.4.2 cDNA libraries and locations

Complete sets of cDNA sequences (in FASTA format), including all confirmed and apriori (gene-finder algorithm based) cDNA sequences, where a staple input of this program. It was a necessary step to compute the locations of the cDNAs on the downloaded genomes. This non-trivial step required a great deal of fine-tuning but was necessary due to the lack of standard and full tables of gene locations. The number of cDNAs actually located on the genome via parsing of

Table B.1: Data concerning species, genomes, cDNAs and regions of interest. (Spring 2006)

Species	Version	%GC	Genome (Mb)	Scaffolds	cDNAs	Located (%)	Two-way best matches w/fugu	Regions of interest
T. rubripes	4.0	46	393	7213	22102	22092 (100%)	21932	18123
T. nigroviridis	7	46	402	25773	183041	173592 (95%)	7647	169
D. rerio	6	37	1626	6653	32143	31922 (99%)	11486	665
X. tropicalus	4.1	42	1511	19501	28324	28306 (100%)	7149	197
H. sapiens	36	41	3433	27223	53233	53208 (100%)	9101	147
C. elegans	150	35	101	3268	27006	27005 (100%)	785	0
C. intestinalis	2	36	173	4390	14182	14180 (100%)	1438	0

the BLAST output closely reflected the number downloaded in most cases, the nearest thing to an exception was in *T. nigroviridis* where only 95% of the cDNAs could be placed on the genome. In the rest of genomes, 99% or 100% of the genes could be placed. The remaining errors could be related to low sequence complexity, incomplete genomes or cDNA sequencing error.

B.4.3 Orthologous genes

Determining the orthologous relationships of the genes was required because all-inclusive data was not available to deduce gene orthology. Orthologous genes were determined via two-way best matching of their representative cDNAs (Figure B.3). Some genes were lost that were exact duplicates of another (the other being chosen in its stead during the matching). This can be seen in the "two-way best matches w/fugu" column of Table B.1, where the *T. rubripes* cDNA library was BLAST aligned to itself and 160 were lost. This may account for why *T. nigroviridis* had far fewer orthologues with *T. rubripes* than *D. rerio* despite *T. nigroviridis* being a fellow Tetraodontid (*puffer family*). Two thousand more or-

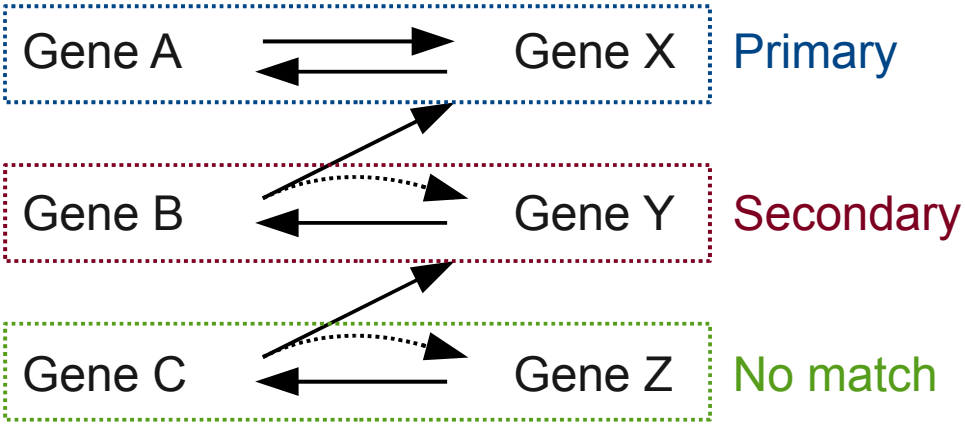


Figure B.3: Orthologs were assigned based on reciprocal BLAST alignments. When two genes were the best-hit of each other; or when the best-hit of one was taken, the two were considered orthologous.

thologues were found between *T. rubripes* and *H. sapiens* than with *X. tropicalus*, which could be attributed to the higher number of cDNAs available with *H. sapiens* than with *X. tropicalus*. A reciprocal BLAST alignment of *C. elegans* cDNAs followed by two-way best matching produced 24,773 two-way best matches, and the same for *C. intestinalis* produces 14,070 best matches. The great deal of duplicate genes in *C. elegans* (up to 2,233) may be why it had so few orthologues with *T. rubripes* compared to *C. intestinalis*.

B.4.4 Intergenic regions of interest

Using the beta (Perl-only) version of the pipeline, the number of intergenic regions found in a genome to be flanked by a pair of genes with orthologues in *T. rubripes* exhibiting the same order and orientation can be found in the final column of table B.1. Once again *T. rubripes* was tested reciprocally to provide the maximum score. *T. nigroviridis* continued to show signs of being a problematic dataset by having only 169 regions, less than the frog (*X. tropicalus*) at 197 and far less than *D. rerio* at 665. *T. nigroviridis* aside, this column is evolutionarily sensible. The downward trend of regions found moving away from *T. rubripes* within the vertebrates speaks of evolutionary divergence, and the lack of regions

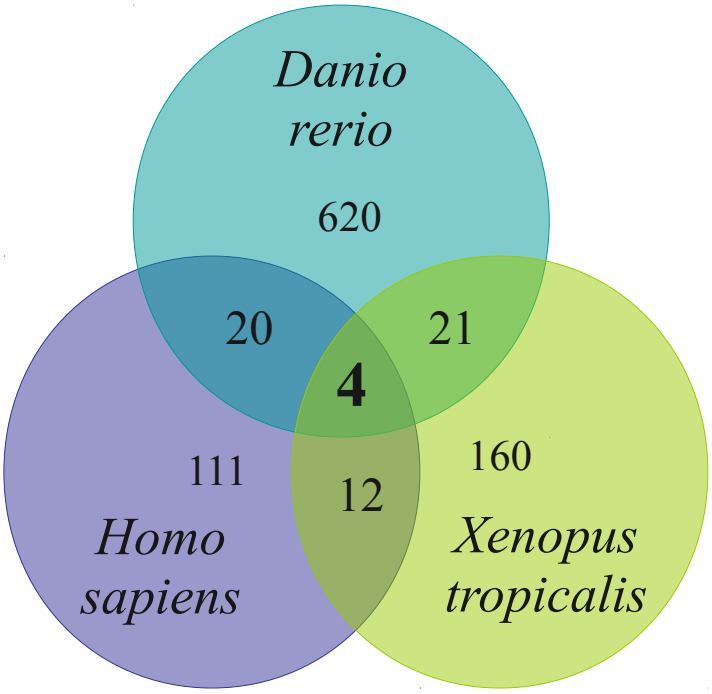


Figure B.4: Venn diagram of how the regions of interest overlap between organisms.

shared with the two invertebrates hints that these regions may be fundamentally vertebrate.

The Venn diagram in Figure B.4 illustrates intersecting regions between *D. rerio*, *X. tropicalus* and *H. sapiens*. The diagram illustrates the conservation of the sets of orthologous flanking genes with conserved order and orientation between species. Between *D. rerio* and the others, the values were similar (20 and 21). Between *X. tropicalus* and *H. sapiens* there were only 12, though this is likely because they had far fewer regions of interest than *D. rerio*. A set of 4 regions appeared to be conserved through all four genomes. This final step was the one most greatly improved upon when the program pipeline was rewritten.

B.5 Discussion

The primary concept of this study was that of the minimal vertebrate genome. *T. rubripes* has the smallest genome of all tested vertebrates yet maintains a similar gene repertoire, thus its tendency to eliminate unnecessary genome coding may have left *T. rubripes* with little more than what is strictly required [Venkatesh et al., 2000].

The beauty of this concept is not so much the elucidation of the possible minimal set of genes, but the ability to more easily pinpoint essential non-coding elements of the genome. The process of trial and error to find non-coding elements of a genome would be very laborious in the lab; fortunately *T. rubripes* undergone a great deal to remove such much of its non-essential intronic and intergenic regions from its genome. The next logical step will be to comparatively locate the regions that are shared between multiple organisms. This has been done with humans [Gilligan et al., 2002] and frogs [Stapleton et al., 2004] by more manual methods with incomplete datasets.

The approach of this study was to sweep the entire genome looking for pairs of genes with a span of uninterrupted (by gene region) intergenic region between them, and if they had orthologues in another genome that shared the same order and orientation, the region was flagged and held. This required a great deal of programming to be accomplished as none of the data used in this study came with the information necessary to locate the regions. Unfortunately, most programs have bugs and the programs written for (and used in) this study were not immaculate. Four sets of intergenic regions conserved in four genomes across 450 million years of evolution were still found. This bodes very well for the future of this study and the concept employed. In addition, now that the proof-of-concept version of this program has been rewritten in a more robust manner, more results are likely to be obtained.

B.6 Future studies and development

In the other studies that followed similar logic researchers used whole genome comparisons or would manually follow orthologous genes through multiple genomes to find flanking regions similar to what this pipeline finds. This powerful program pipeline can be run nearly autonomously and has great potential as it is compartmentalized in such a way that its parts can easily be reused for other purposes. Not long after this program was originally tested, several papers came out describing similar techniques. Most of those data have became very outdated and are limited to higher vertebrates [Visel et al., 2007].

The future goal of this project will be to rerun all the new program versions on the updated genomes. A sensible expansion would be to have a second stage that takes the large sets of sequences from the regions of interest and looks for conserved pieces among sets of regions and all regions using some of the techniques that other researchers have now begun to refine. This program can have great application in all genomes as all it needs for input are putative cDNA sets and genomic contigs. In the near future we would like to apply it to the original organisms in search of conserved regions near immune genes, the lower vertebrates and to cyanobacterial genomes in search of conserved regions near biosynthetic clusters of use to make secondary metabolites.

Appendix C

Fish Phylogenetics Activity

Hands-on dichotomous keys, phylogenetics, and fish diversity

C.1 Abstract

This activity yields experience with dichotomous keys and a fundamental understanding of DNA based evolutionary trees (phylogenetic trees). The students observe a genetic algorithm at work, sort fishes based on their morphology, formalize this into an evolutionary tree, and this is used as a transition point to DNA based trees (of which there is a live example).

LEVEL 9-12th Grade Biology Students

DURATION One 50 minute class period, though more than 60 minutes is better

LEARNING OBJECTIVES At the end of this activity students should be able to use a dichotomous key to identify species, infer evolutionary relationships from observations of morphology, draw representative cladograms and understand the basis for molecular studies of evolution.

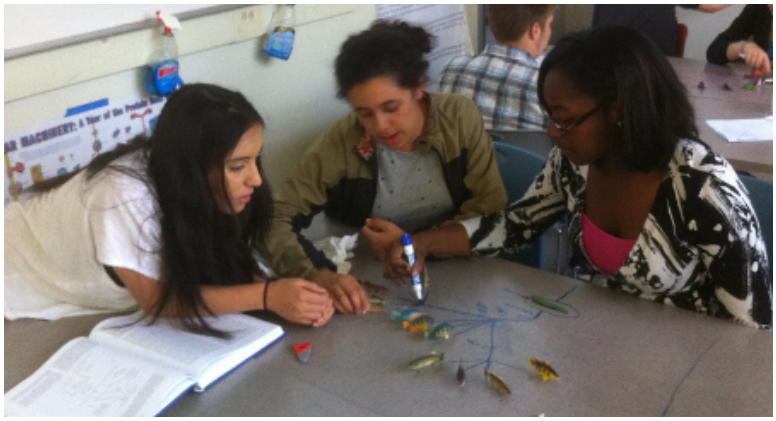


Figure C.1: High School students debating the finer points of fish evolution.

C.2 Educational Standards Addressed

C.2.1 California Science Standards

EVOLUTION

(7) The frequency of an allele in a gene pool of a population depends on many factors and may be stable or unstable over time. As a basis for understanding this concept:

(a) Students know why natural selection acts on the phenotype rather than the genotype of an organism.

(c) Students know new mutations are constantly being generated in a gene pool.

(d) Students know variation within a species increases the likelihood that at least some members of a species will survive under changed environmental conditions.

(8) Evolution is the result of genetic changes that occur in constantly changing environments. As a basis for understanding this concept:

(b) Students know a great diversity of species increases the chance that at least some organisms survive major changes in the environment

(f*) Students know how to use comparative embryology, DNA or protein sequence comparisons, and other independent sources of data to create a branching diagram (cladogram) that shows probable evolutionary relationships.

Investigation and Experimentation

(1) Scientific progress is made by asking meaningful questions and conducting careful investigations. As a basis for understanding this concept and addressing the content in the other four strands, students should develop their own questions and perform investigations. Students will:

(a) Select and use appropriate tools and technology (such as computer-linked probes, spreadsheets, and graphing calculators) to perform tests, collect data, analyze relationships, and display data.

(d) Formulate explanations by using logic and evidence.

(g) Recognize the usefulness and limitations of models and theories as scientific representations of reality.

(l) Analyze situations and solve problems that require combining and applying concepts from more than one area of science.

For reference - letters and numbers above correlate to listed California educational standards.

C.2.2 National Science Education Standards

Life Science Standards: Molecular basis of heredity

- In all organisms, the instructions for specifying the characteristics of the organism are carried in DNA, a large polymer formed from subunits of four kinds (A, G, C, and T). The chemical and structural properties of DNA explain how the genetic information that underlies heredity is both encoded in genes (as a string of molecular "letters") and replicated (by

a templating mechanism). Each DNA molecule in a cell forms a single chromosome.

- Changes in DNA (mutations) occur spontaneously at low rates. Some of these changes make no difference to the organism, whereas others can change cells and organisms. Only mutations in germ cells can create the variation that changes an organism's offspring.

Life Science Standards: Biological evolution

- Natural selection and its evolutionary consequences provide a scientific explanation for the fossil record of ancient life forms, as well as for the striking molecular similarities observed among the diverse species of living organisms.
- The millions of different species of plants, animals, and microorganisms that live on earth today are related by descent from common ancestors.
- Biological classifications are based on how organisms are related. Organisms are classified into a hierarchy of groups and subgroups based on similarities which reflect their evolutionary relationships. Species is the most fundamental unit of classification.

Science and Technology Standards

Abilities of technological design

EVALUATE THE SOLUTION AND ITS CONSEQUENCES. Students should test any solution against the needs and criteria it was designed to meet. At this stage, new criteria not originally considered may be reviewed.

Understanding about science and technology

Scientists in different disciplines ask different questions, use different methods of investigation, and accept different types of evidence to support their explanations. Many scientific investigations require the contributions of individuals from different disciplines, including engineer-

ing. New disciplines of science, such as geophysics and biochemistry often emerge at the interface of two older disciplines.

C.3 Background Information

Natural Selection The process by which certain organisms manage to pass on their genetic code while others do not due to a varying level of fitness. A good example of this for the sake of discussion is the Genetic Algorithm used in this activity. Once run, a random 2D car or bike will be generated that will attempt to drive over a hilly terrain. The blue circles represent the wheels and the red circles represent weights that must not touch the ground (black line). Bikes will be randomly generated at first (20 bikes per generation), and at the end of each generation, the two bikes that drove the furthest will be mated to create the genetic basis for the next generation. The next generation will contain 20 bikes that are mutations of the mating of those two previous bikes. This process will continue for as long as the program is allowed to run.

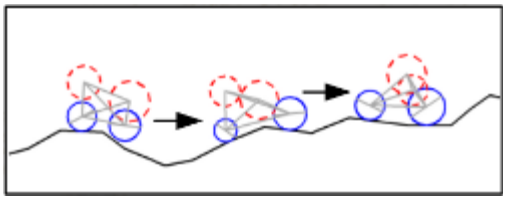


Figure C.2:

The genetic algorithm demonstration accurately simulates natural selection and genetic inheritance in some ways, but certainly does not in others. These are more fully covered in the attached presentation.

Classification by Morphology The Linnaean system of naming organisms is rooted in the sorting of organisms based on their physical characteristics (morphology). This has in turn allowed us to design guides to finding the

names of organisms based on their morphology. These keys (called dichotomous keys) rely on a system of yes/no questions in a type of choose your own adventure book.

Evolution Descent with modification is what has led to the wide array of organisms found on our planet. Based on the principle of natural selection leading to favorable changes to morphology that allow increasingly favorable changes to morphology over large time spans we can theorize over the common ancestry of different groups of organisms.

Phylogenetics Theories of common ancestry between groups of organisms can be generated and tested (to an extent) using DNA sequences taken from the organisms in question. Much the way evolutionary trees (cladograms) can be built upon tables of physical characteristics among organisms, comparisons between nucleotides of highly conserved genes can lead to very accurate evolutionary trees called phylogenetic trees.

C.4 Research Applications

LABORATORY OF PROFESSOR LENA GERWICK

Phylogenetics as a tool, has many applications in biological research today. At its core, it is simply a way to present the relative relatedness of a series of DNA sequences. This is both useful for presenting the data and the analysis used to build the tree is important for making hypothesis about the evolution of the DNA sequences themselves. This was historically used and developed to identify the evolution of one species from another but more recently has also been applied to evolution of genes within a species' entire genetic code (or genome). Below is a phylogenetic tree that I generated using gene sequences of the 20 sghC1q genes (labeled 1-20 in the tree) found in the zebrafish genome.

As you may know, the branches that come from common bifurcations are theoretically from common ancestors. In this figure we show that the genes found on Chromosomes two and seven in the zebrafish genome are more similar to them-

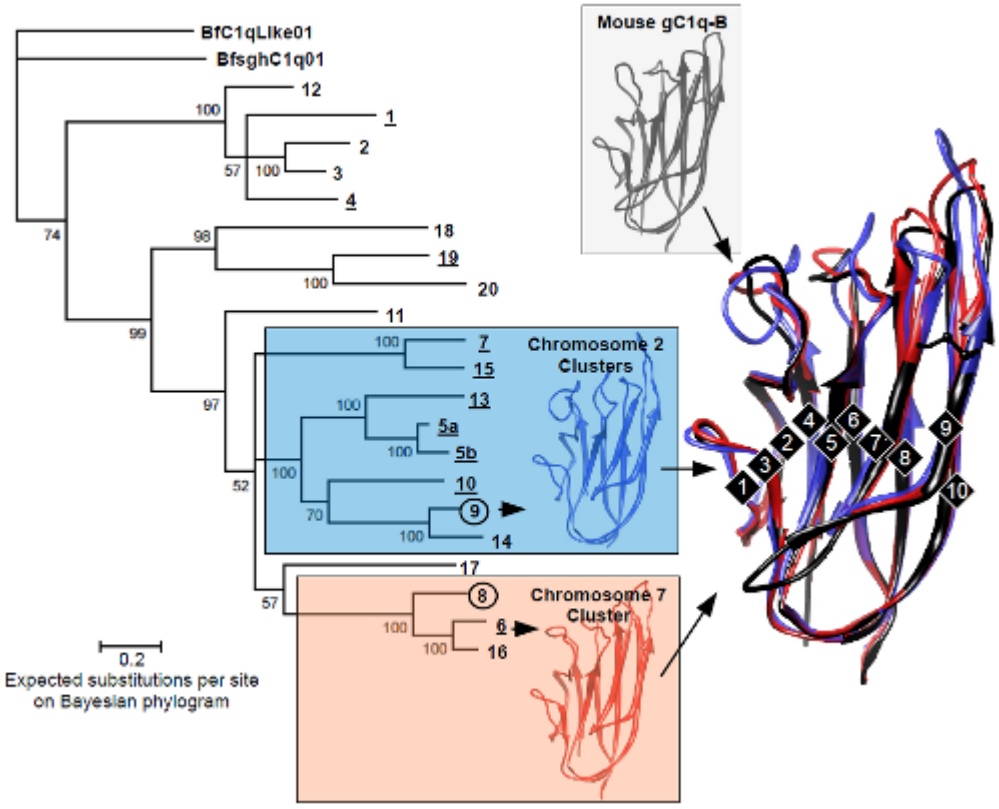


Figure C.3: [Carland et al., 2011]

selves than the genes from other chromosomes, meaning that they probably radiated within chromosomes instead of across (such as in a whole genome duplication). The protein structures for selected genes are shown, overlayed to the right with a similar protein from mice, and they all overlap very well, showing that the shape of this protein has been conserved throughout evolution (fish to mice at least).

C.5 Implementation Guide

ENGAGE

Activity Informal discussion of Genetic Algorithm (with guided notes)

Key Questions What is happening? How does this model Natural Selection?

Discussion How is this incorrect from true natural selection?

Potential Misconceptions Natural selection requires a measure of fitness, the algorithm only allows the best two bikes to "procreate" while a generation of real population may allow all but a few.

EXPLORE

Logistics During this phase, leave the genetic algorithm running and begin the presentation. Go until you get to the slide with the shark morphology guide.

Activity Shark-dichotomous-key handout, do a few (or all) of the sharks.

Introduce There is a great diversity of life on this planet, particularly among the fishes. In the presentation file there are a few slides of some interesting fishes to incite interest. How we keep track of the different types and name them follows a logical system. This phase practices that system.

Guide Make sure that the guide to shark morphology is on display. (in presentation)

Key Questions What are the family names of these sharks? How do we distinguish between them? Can we just google them? (try it, fishbase and wiki do a good job)

EXPLAIN

At this point it is good to run over a few of the answers and have fun pointing out any of the sharks that may exist in the local area. The presentation contains sharks local to the San Diego/La Jolla area. Keep the focus on morphology and how we can use that to name things. Any mention of how organisms have evolved to survive in particular habitats and styles of predation (adaptations) is especially useful here. Stop when you get to the slide titled Fish / Oceanic evolution.

ELABORATE

Activity Pass out the fish models and instruct the students to organize them as they see fit, however that may be (most will go by morphology, that's the idea). Given some time most students will get close to correct. Try to walk around the room and have each group explain their system. Transition to sorting the groups in terms of an evolutionary tree with large paper and markers or dry-erase if possible.

Discuss Why did they choose what they chose?

Sample Questions Where does the eel go? Which came first, turtles or penguins?

Key Points Evolution doesn't always go from simple to complex, sometimes simple looking creatures are actually the product of millions of years of honed simplifications. Application/Elaboration: Have the groups discuss as much as possible, and after a certain point (a strong tree) give them a reference. Most editions of Campbell/Reece have a good section on vertebrate evolution for any groups with the marine organisms pack. The freshwater fishes groups will have a harder time finding a reference but generally seem to do better without one because they have viewer preconceived misconceptions.

EVALUATE Explain/Discuss After discussing key question 1, it is time to make a phylogenetic tree. A simple and imperfect method is to submit the provided sequences to the one-click phylogeny website. It will align the sequences and provide a fairly strong phylogenetic tree. If you are feeling a tad more daring, try showing the sequences to the students in an alignment program such as mega or clustalw, to help them visualize DNA. This goes especially well if they have learned anything about cladistic tables, as a DNA alignment is basically a large cladistic table.

Key Questions

1. What other data could we use to discern the evolution of these animals? Most of the time the students will eventually guess DNA. Might be a good time to drive home Phenotype vs Genotype as they have been

using phenotype but now we can use the genotype.

2. Are the student made trees correct? (Compare to the one you make)

3. Is the tree made by the website/program correct? It will probably be very close.

Table C.1: The 5E Model

5Es At-A-Glance	Activity	Key Ideas/Questions	Timing
ENGAGE	While the class is getting organized, leave the genetic algorithm running on the board to garner inquisitive attention. Then (after introductions) have an informal discussion of what the animation is trying to do.	How does natural selection work?	15 min
EXPLORE	Transition (by speaking about morphology) to the dichotomous key phase. Have the students discern 3-4 of the sharks within the shark-dichotomous-key activity.	How do we classify fishes? Naming is related to morphology.	20 min
EXPLAIN	Tie together what they are doing to the concepts of morphology from evolutionary processes.	What other ways can fishes be sorted and organized?	10 min
ELABORATE	Pass out the fishes and have student groups (of 3-4) arrange them based on their morphologies. If the students are prepared for it, transition this into drawing evolutionary trees of the groups.	How should we sort these fishes? How did these fishes evolve?	20 min
EVALUATE	Use the "one click phylogeny" website and COI sequences from the organisms to generate a phylogenetic tree. Compare and contrast to student trees. If time permits, final portion of the lecture explains how these are not always right but very powerful.	How does these trees stand up to your trees?	10-30 min

C.6 Materials and Methods

Table C.2: Materials

Quantity	Material	Where to Order
2-3 per station	Large sheets of paper	
1-2 per station	Markers	
1 set per station	Models: Midwest Tackle Box, Ocean Toob	www.replicatoyfish.com www.safariltd.com
1 per class	Computer + Projector	
1 per student	Handout	

C.6.1 Teacher preparation instructions

- Have a computer attached to a projector with:
  - Presentation
  - Web-browser opened to genetic algorithm
  - Web-browser also opened to "one-click phylogeny" website
  - Optional - the sequence files loaded into an alignment viewing program (MEGA is a good and simple one)
- Each student should have a handout
- Fish models should be organized and ready to deploy (later)

C.6.2 Implementation strategies

- Four students per station/group
- Each group shares the large papers and collections of fishes
- List of Teacher Materials

- For the demonstration (explain phase) the teacher should run through the presentation and be ready to make a simple phylogenetic tree using the sequences provided. See the supplementary video for specifics.

C.6.3 Resources and References

California Science Standards Science Content Standards for California Public Schools, Kindergarten Through Grade Twelve. 1998. California State Board of Education. Available online: http://www.cde.ca.gov/be/st/ss/

National Science Education Standards National Science Education Standards. 1996. Center for Science, Mathematics, and Engineering Education. National Academy Press, Washington DC. Available online: http://www.nap.edu

Genetic Algorithm http://www.qubit.devisland.net/ga/ - Pete Shadbolt, University of Bristol

[**Carland et al., 2011**] See Chapter 3 of this thesis or bibliography

Dichotomous Key Activity Available all over the web as shark dichotomous key activity

C.6.4 Guided Notes for Genetic Algorithm / Phylogenetic Activity

1. A genetic algorithm is a _____ that uses the concepts of _____ to solve a problem.

2. The genetic algorithm shown is inaccurate because only _____ individuals are allowed to pass on their genes.

3. After a few generations, draw one of the bikes that you see, then wait and draw another.

4. What is a dichotomous key?

5. Name one morphological characteristic that you used to find the name of a shark.

6. Draw another one of your bikes, what has changed?

7. In your own words:

   What is convergent evolution?

   What is divergent evolution?

8. Which group of organisms did your group get? (circle) Ocean Creatures / Freshwater Fishes

9. Describe two of your evolutionary groups

10. Which organism did you decide is the most:

    "primitive/ancient"?

    "derived/recent"?

11. Currently, scientists use _____ and _____ to discern species and evolutionary relationships.

12. Did the computer program agree with your evolutionary deductions?

C.6.5 Presentation Guide

1. This presentation brought to you by a graduate student in Marine Biology at the Scripps Institution of Oceanography. Seen here are our five primary ships.

2. Fish vs Fishes. The pet peeve of many ichthyologists (fish researchers) is the common misconception of the word fishes. Single or multiple individuals of the same species are called fish, multiple species are collectively referred

to as fishes. This slide is also an easy way to get the audience involved by asking them to name the stars of Finding Nemo.

3. This is a large and strange fish. Ask the students if they know what it is. To their surprise, it is a freshwater fish, found in lakes/rivers, mostly in the Northeastern United States. Formally Esox masquinongy, it is the muskellunge or musky.

4. Once again, what is this? This is a shark, shorter in length (about a foot) that we collected in the fall of 2010 on a cruise in the Santa Barbara basin. This deep sea shark has no sharp teeth and can appear an almost translucent purple, it's name is the filetail catshark. The question to bring up now is how do you find out the name of an organism?.

5. This is a picture of myself (Tristan) and my colleague (Daniel Conley) aboard the R/V Melville trying to discern the name of the fish in my hand. In front of us are prepared photos of organisms we were likely to catch. In Daniel's hand, is a book called a dichotomous key. This particular one contains all the near-shore fishes of the Southern California coast.

6. This guide contains part of the morphological characteristics that will help when identifying the fish.

7. This is the rest of them.

8. These are some basic characteristics needed to complete the provided hand-out, identifying the family names of the sharks. Stop here until they've completed the keys.

9. The next several slides are of sharks that can (though not often) be found around Southern California. Saying something of the diversity that can be attained from evolution would be good around here, and this is a fine time to take a look at how the genetic algorithm has been doing.

10. Fish / Oceanic evolution (actually 18): This is where you pass out the replica fishes. Presentation should stop here until they are done with this part.

11. This slide is to wrap up what they have been doing with the replicas. If they haven't been asked to arrange them into trees, now is a good time to do so. Now is also when you should input the sequences into the phylogenetic program of your choice.

12. Probably won't find the time to use these slides, but they are here to discuss an overturned understanding. It was once thought that amphioxus was more related to vertebrates than tunicates, and that turned out to be incorrect. The last three slides are backup info.

Bibliography

[Agnew and Barnes, 2007] Agnew, W. and Barnes, A. C. (2007). Streptococcus iniae: An aquatic pathogen of global veterinary significance and a challenging candidate for reliable vaccination. *Veterinary Microbiology*, 122(1-2):1–15.

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

[Anderson et al., 1985] Anderson, K. V., Jrgens, G., and Nsslein-Volhard, C. (1985). Establishment of dorsal-ventral polarity in the drosophila embryo: Genetic studies on the role of the toll gene product. *Cell*, 42(3):779–789.

[Anisimova and Kosiol, 2009] Anisimova, M. and Kosiol, C. (2009). Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Molecular Biology and Evolution*, 26(2):255–271. PMID: 18922761.

[Arnold et al., 2000] Arnold, K., Gosling, J., and Holmes, D. (2000). *Java(TM) Programming Language, The (3rd Edition)*. Prentice Hall PTR, 3 edition.

[Bao et al., 2005] Bao, D., Pang, Z., and Morgan, J. I. (2005). The structure and proteolytic processing of cbln1 complexes. *Journal of Neurochemistry*, 95(3):618–629. PMID: 16135095.

[Baumann and Gauldie, 1994] Baumann, H. and Gauldie, J. (1994). The acute phase response. *Immunology Today*, 15(2):74–80.

[Bayne and Gerwick, 2001] Bayne, C. J. and Gerwick, L. (2001). The acute phase response and innate immunity of fish. *Developmental & Comparative Immunology*, 25(8-9):725–743.

[Bayne et al., 2001] Bayne, C. J., Gerwick, L., Fujiki, K., Nakao, M., and Yano, T. (2001). Immune-relevant (including acute phase) genes identified in the livers of rainbow trout, oncorhynchus mykiss, by means of suppression subtractive hybridization. *Developmental & Comparative Immunology*, 25(3):205–217.

[Beadle and Tatum, 1941] Beadle, G. W. and Tatum, E. L. (1941). Genetic control of biochemical reactions in neurospora. *Proceedings of the National Academy of Sciences of the United States of America*, 27(11):499–506. PMID: 16588492.

[Bennetzen, 2007] Bennetzen, J. L. (2007). Patterns in grass genome evolution. *Current Opinion in Plant Biology*, 10(2):176–181.

[Berg, 1996] Berg, R. D. (1996). The indigenous gastrointestinal microflora. *Trends in Microbiology*, 4(11):430–435.

[Bill et al., 2009] Bill, B. R., Petzold, A. M., Clark, K. J., Schimmenti, L. A., and Ekker, S. C. (2009). A primer for morpholino use in zebrafish. *Zebrafish*, 6(1):69–77. PMID: 19374550 PMCID: 2776066.

[Black, 2003] Black, D. (2003). Mechanisms of alternative pre-messenger RNA splicing. *ANNUAL REVIEW OF BIOCHEMISTRY*, 72:291–336.

[Bohlson et al., 2007] Bohlson, S. S., Fraser, D. A., and Tenner, A. J. (2007). Complement proteins c1q and MBL are pattern recognition molecules that signal immediate and long-term protective immune functions. *Molecular Immunology*, 44(1-3):33–43.

[Boshra et al., 2006] Boshra, H., Li, J., and Sunyer, J. (2006). Recent advances on the complement system of teleost fish. *Fish & Shellfish Immunology*, 20(2):239–262.

[Bray et al., 2006] Bray, T., Paoli, J., Sperberg-McQueen, C., Maler, E., and Yergeau, F. (2006). Extensible markup language (XML) 1.0 (Fourth edition). http://www.w3.org/TR/2006/REC-xml-20060816/.

[Brenner et al., 1993] Brenner, S., Elgar, G., Sanford, R., Macrae, A., Venkatesh, B., and Aparicio, S. (1993). Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome. *Nature*, 366(6452):265–268.

[Buchanan et al., 2005] Buchanan, J. T., Stannard, J. A., Lauth, X., Ostland, V. E., Powell, H. C., Westerman, M. E., and Nizet, V. (2005). Streptococcus iniae phosphoglucomutase is a virulence factor and a target for vaccine development. *Infect. Immun.*, 73(10):6935–6944.

[Burr et al., 2005] Burr, G., Gatlin III, D., and Ricke, S. (2005). Microbial ecology of the gastrointestinal tract of fish and the potential application of prebiotics and probiotics in finfish aquaculture. *Journal of the World Aquaculture Society*, 36(4):425–436.

[Carland and Gerwick, 2010] Carland, T. M. and Gerwick, L. (2010). The c1q domain containing proteins: Where do they come from and what do they do? *Developmental & Comparative Immunology*, 34(8):785–790.

[Carland et al., 2011] Carland, T. M., Locke, J. B., Nizet, V., and Gerwick, L. (2011). Differential expression and intrachromosomal evolution of the sghC1q genes in zebrafish (Danio rerio). *Developmental & Comparative Immunology*, In Press, Accepted Manuscript.

[Chen et al., 2002] Chen, W., Burgess, S., Golling, G., Amsterdam, A., and Hopkins, N. (2002). High-Throughput selection of retrovirus producer cell lines leads to markedly improved efficiency of germ Line-Transmissible insertions in zebra fish. *J. Virol.*, 76(5):2192–2198.

[Clack, 2002] Clack, J. A. (2002). *Gaining Ground: The Origin and Early Evolution of Tetrapods*. Indiana University Press.

[Cray et al., 2009] Cray, C., Zaias, J., and Altman, N. H. (2009). Acute phase response in animals: A review. 59(6):517–526. PMID: 20034426 PMCID: 2798837.

[Crick, 1970] Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563.

[Crick, 1958] Crick, F. H. (1958). On protein synthesis. *Symposia of the Society for Experimental Biology*, 12:138–163. PMID: 13580867.

[Dehal and Boore, 2005] Dehal, P. and Boore, J. L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology*, 3(10). PMID: 16128622 PMCID: 1197285.

[Dishaw et al., 2005] Dishaw, L. J., Smith, S. L., and Bigger, C. H. (2005). Characterization of a c3-like cDNA in a coral: phylogenetic implications. *Immunogenetics*, 57(7):535–548.

[Dodds and Matsushita, 2007] Dodds, A. W. and Matsushita, M. (2007). The phylogeny of the complement system and the origins of the classical pathway. *Immunobiology*, 212(4-5):233–243.

[DuBois, 2008] DuBois, P. (2008). *MySQL*. Addison-Wesley Professional, 4 edition.

[Dumrongphol et al., 2009] Dumrongphol, Y., Hirota, T., Kondo, H., Aoki, T., and Hirono, I. (2009). Identification of novel genes in japanese flounder (Paralichthys olivaceus) head kidney up-regulated after vaccination with streptococcus iniae formalin-killed cells. *Fish & Shellfish Immunology*, 26(1):197–200.

[Edgar, 2004] Edgar, R. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5(1).

[Ellis, 2001] Ellis, R. (2001). *Aquagenesis: The Origin and Evolution of Life in the Sea*. Viking Adult, first edition edition.

[Emanuelsson et al., 2007] Emanuelsson, O., Brunak, S., von Heijne, G., and Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protocols*, 2(4):953–971.

[Endo et al., 2006] Endo, Y., Takahashi, M., and Fujita, T. (2006). Lectin complement system and pattern recognition. *Immunobiology*, 211(4):283–293. PMID: 16697920.

[Falkowski et al., 2000] Falkowski, P., Scholes, R. J., Boyle, E., Canadell, J., Canfield, D., Elser, J., Gruber, N., Hibbard, K., Hgberg, P., Linder, S., Mackenzie, F. T., Moore III, B., Pedersen, T., Rosenthal, Y., Seitzinger, S., Smetacek, V., and Steffen, W. (2000). The global carbon cycle: A test of our knowledge of earth as a system. *Science*, 290(5490):291–296.

[Fernandez-Fuentes et al., 2007] Fernandez-Fuentes, N., Madrid-Aliste, C. J., Rai, B. K., Fajardo, J. E., and Fiser, A. (2007). M4T: a comparative protein structure modeling server. *Nucleic Acids Research*, 35(Web Server issue):W363–368. PMID: 17517764 PMCID: 1933164.

[Finnegan et al., 2011] Finnegan, S., Bergmann, K., Eiler, J. M., Jones, D. S., Fike, D. A., Eisenman, I., Hughes, N. C., Tripati, A. K., and Fischer, W. W. (2011). The magnitude and duration of late OrdovicianEarly silurian glaciation. *Science*, 331(6019):903–906.

[Flajnik and Kasahara, 2001] Flajnik, M. F. and Kasahara, M. (2001). Comparative genomics of the MHC: glimpses into the evolution of the adaptive immune system. *Immunity*, 15(3):351–362.

[Ford and Myers, 2008] Ford, J. S. and Myers, R. A. (2008). A global assessment of salmon aquaculture impacts on wild salmonids. *PLoS Biol*, 6(2).

[Gaboriaud et al., 2003] Gaboriaud, C., Juanhuix, J., Gruez, A., Lacroix, M., Darnault, C., Pignol, D., Verger, D., Fontecilla-Camps, J. C., and Arlaud, G. J. (2003). The crystal structure of the globular head of complement protein c1q provides a basis for its versatile recognition properties. *Journal of Biological Chemistry*, 278(47):46974–46982.

[Garlatti et al., 2010] Garlatti, V., Chouquet, A., Lunardi, T., Vivs, R., Padassi, H., Lortat-Jacob, H., Thielens, N. M., Arlaud, G. J., and Gaboriaud, C. (2010). Cutting edge: C1q binds deoxyribose and heparan sulfate through neighboring sites of its recognition domain. *The Journal of Immunology*, 185(2):808–812.

[Gauthier and Rhodes, 2009] Gauthier, D. T. and Rhodes, M. W. (2009). Mycobacteriosis in fishes: A review. *The Veterinary Journal*, 180(1):33–47.

[Gerlach et al., 2004] Gerlach, D., Schlott, B., and Schmidt, K. (2004). Cloning and expression of a sialic acid-binding lectin from the snail cepaea hortensis. *FEMS Immunology and Medical Microbiology*, 40(3):215–221.

[Gerwick et al., 2007] Gerwick, L., Corley-Smith, G., and Bayne, C. J. (2007). Gene transcript changes in individual rainbow trout livers following an inflammatory stimulus. *Fish & Shellfish Immunology*, 22(3):157–171.

[Gerwick et al., 2005] Gerwick, L., Corley-Smith, G. E., Nakao, M., Watson, J., and Bayne, C. J. (2005). Intracranial injections induce local transcription of a gene encoding precerebellin-like protein. *Fish Physiology and Biochemistry*, 31(4):363–372.

[Gerwick et al., 2000] Gerwick, L., Reynolds, W. S., and Bayne, C. J. (2000). A precerebellin-like protein is part of the acute phase response in rainbow trout, oncorhynchus mykiss. *Developmental & Comparative Immunology*, 24(6-7):597–607.

[Gestal et al., 2010] Gestal, C., Pallavicini, A., Venier, P., Novoa, B., and Figueras, A. (2010). MgC1q, a novel c1q-domain-containing protein involved in the immune response of mytilus galloprovincialis. *Developmental & Comparative Immunology*, 34(9):926–934.

[Ghai et al., 2007] Ghai, R., Waters, P., Roumenina, L. T., Gadjeva, M., Kojouharova, M. S., Reid, K. B. M., Sim, R. B., and Kishore, U. (2007). C1q and its growing family. *Immunobiology*, 212(4-5):253–266.

[Ghebrehiwet et al., 2001] Ghebrehiwet, B., Lim, B., Kumar, R., Feng, X., and Peerschke, E. I. B. (2001). gC1qR/p33, a member of a new class of multifunctional and multicompartmental cellular proteins, is involved in inflammation and infection. *Immunological Reviews*, 180(1):65–77.

[Ghebrehiwet et al., 1994] Ghebrehiwet, B., Lim, B. L., Peerschke, E. I., Willis, A. C., and Reid, K. B. (1994). Isolation, cDNA cloning, and overexpression of a 33-kD cell surface glycoprotein that binds to the globular "heads" of c1q. *The Journal of Experimental Medicine*, 179(6):1809–1821.

[Ghebrehiwet and Peerschke, 2004] Ghebrehiwet, B. and Peerschke, E. I. B. (2004). cC1q-R (calreticulin) and gC1q-R/p33: ubiquitously expressed multiligand binding cellular proteins involved in inflammation and infection. *Molecular Immunology*, 41(2-3):173–183.

[Giffen et al., 2003] Giffen, P. S., Turton, J., Andrews, C. M., Barrett, P., Clarke, C. J., Fung, K. W., Munday, M. R., Roman, I. F., Smyth, R., Walshe, K., and York, M. J. (2003). Markers of experimental acute inflammation in the wistar han rat with particular reference to haptoglobin and c-reactive protein. *Archives of Toxicology*, 77(7):392–402.

[Gilligan et al., 2002] Gilligan, P., Brenner, S., and Venkatesh, B. (2002). Fugu and human sequence comparison identifies novel human genes and conserved non-coding sequences. *Gene*, 294(1-2):35–44.

[Gonzalez et al., 2007] Gonzalez, S. F., Huising, M. O., Stakauskas, R., Forlenza, M., Lidy Verburg-van Kemenade, B. M., Buchmann, K., Nielsen, M. E., and Wiegertjes, G. F. (2007). Real-time gene expression analysis in carp (Cyprinus carpio l.) skin: Inflammatory responses to injury mimicking infection with ectoparasites. *Developmental & Comparative Immunology*, 31(3):244–254.

[Grunwald and Eisen, 2002] Grunwald, D. J. and Eisen, J. S. (2002). Headwaters of the zebrafish - emergence of a new model vertebrate. *Nat Rev Genet*, 3(9):717–724.

[Hahn et al., 2005] Hahn, M. W., De Bie, T., Stajich, J. E., Nguyen, C., and Cristianini, N. (2005). Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Research*, 15(8):1153–1160. PMID: 16077014 PMCID: 1182228.

[Hastings et al., 2009] Hastings, P. J., Lupski, J. R., Rosenberg, S. M., and Ira, G. (2009). Mechanisms of change in gene copy number. *Nat Rev Genet*, 10(8):551–564.

[Hegedus et al., 2009] Hegedus, Z., Zakrzewska, A., goston, V. C., Ordas, A., Rcz, P., Mink, M., Spaink, H. P., and Meijer, A. H. (2009). Deep sequencing of the zebrafish transcriptome response to mycobacterium infection. *Molecular Immunology*, 46(15):2918–2930.

[Hirai et al., 2005] Hirai, H., Pang, Z., Bao, D., Miyazaki, T., Li, L., Miura, E., Parris, J., Rong, Y., Watanabe, M., Yuzaki, M., and Morgan, J. I. (2005). Cbln1 is essential for synaptic integrity and plasticity in the cerebellum. *Nat Neurosci*, 8(11):1534–1541.

[Hobbie et al., 1977] Hobbie, J. E., Daley, R. J., and Jasper, S. (1977). Use of nuclepore filters for counting bacteria by fluorescence microscopy. *Applied and Environmental Microbiology*, 33(5):1225–1228. PMID: 327932 PMCID: 170856.

[Holland and Garcia-Fernndez, 1996] Holland, P. W. H. and Garcia-Fernndez, J. (1996). HoxGenes and chordate evolution. *Developmental Biology*, 173(2):382–395.

[Hosszu et al., 2010] Hosszu, K. K., Santiago-Schwarz, F., Peerschke, E. I. B., and Ghebrehiwet, B. (2010). Evidence that a C1q/C1qR system regulates monocyte-derived dendritic cell differentiation at the interface of innate and acquired immunity. *Innate Immunity*, 16(2):115–127. PMID: 19710097.

[Hu et al., 2010] Hu, Y., Pan, X., Xiang, L., and Shao, J. (2010). Characterization of c1q in teleosts. *Journal of Biological Chemistry*, 285(37):28777–28786.

[Huang et al., 2008] Huang, S., Yuan, S., Guo, L., Yu, Y., Li, J., Wu, T., Liu, T., Yang, M., Wu, K., Liu, H., Ge, J., Yu, Y., Huang, H., Dong, M., Yu, C., Chen, S., and Xu, A. (2008). Genomic analysis of the immune gene repertoire of amphioxus reveals extraordinary innate complexity and diversity. *Genome Research*, 18(7):1112–1126.

[Huang and Madan, 1999] Huang, X. and Madan, A. (1999). CAP3: a DNA sequence assembly program. *Genome Research*, 9(9):868–877. PMID: 10508846 PMCID: 310812.

[Iijima et al., 2007] Iijima, T., Miura, E., Matsuda, K., Kamekawa, Y., Watanabe, M., and Yuzaki, M. (2007). Characterization of a transneuronal cytokine family cbln regulation of secretion by heteromeric assembly. *European Journal of Neuroscience*, 25(4):1049–1057.

[Janeway, 1989] Janeway, C. (1989). Approaching the asymptote? evolution and revolution in immunology. *Cold Spring Harbor Symposia on Quantitative Biology*, 54:1–13.

[Jones et al., 1989] Jones, E. Y., Stuart, D. I., and Walker, N. P. C. (1989). Structure of tumour necrosis factor. *Nature*, 338(6212):225–228.

[Kasahara et al., 2007] Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., Yamada, T., Nagayasu, Y., Doi, K., Kasai, Y., Jindo, T., Kobayashi, D., Shimada, A., Toyoda, A., Kuroki, Y., Fujiyama, A., Sasaki, T., Shimizu, A., Asakawa, S., Shimizu, N., Hashimoto, S.-i., Yang, J., Lee, Y., Matsushima, K., Sugano, S., Sakaizumi, M., Narita, T., Ohishi, K., Haga, S., Ohta, F., Nomoto, H., Nogata, K., Morishita, T., Endo, T., Shin-I, T., Takeda, H., Morishita, S., and Kohara, Y. (2007). The medaka draft genome and insights into vertebrate genome evolution. *Nature*, 447(7145):714–719.

[Kasamatsu et al., 2010] Kasamatsu, J., Oshiumi, H., Matsumoto, M., Kasahara, M., and Seya, T. (2010). Phylogenetic and expression analysis of lamprey toll-like receptors. *Developmental & Comparative Immunology*, 34(8):855–865.

[Kawai and Akira, 2010] Kawai, T. and Akira, S. (2010). The role of pattern-recognition receptors in innate immunity: update on toll-like receptors. *Nat Immunol*, 11(5):373–384.

[Kawakami, 2007] Kawakami, K. (2007). Tol2: a versatile gene transfer vector in vertebrates. *Genome Biology*, 8 Suppl 1. PMID: 18047699.

[Kawakami et al., 1998] Kawakami, K., Koga, A., Hori, H., and Shima, A. (1998). Excision of the tol2 transposable element of the medaka fish, oryzias latipes, in zebrafish, danio rerio. *Gene*, 225(1-2):17–22.

[Kawakami and Shima, 1999] Kawakami, K. and Shima, A. (1999). Identification of the tol2 transposase of the medaka fish oryzias latipes that catalyzes excision of a nonautonomous tol2 element in zebrafish danio rerio. *Gene*, 240(1):239–244.

[Kent, 2002] Kent, W. J. (2002). BLATthe BLAST-like alignment tool. *Genome Research*, 12(4):656–664. PMID: 11932250.

[Kishore et al., 2004] Kishore, U., Gaboriaud, C., Waters, P., Shrive, A. K., Greenhough, T. J., Reid, K. B. M., Sim, R. B., and Arlaud, G. J. (2004). C1q and tumor necrosis factor superfamily: modularity and versatility. *Trends in Immunology*, 25(10):551–561.

[Kishore and Reid, 2000] Kishore, U. and Reid, K. B. M. (2000). C1q: Structure, function, and receptors. *Immunopharmacology*, 49(1-2):159–170.

[Lauth et al., 2005] Lauth, X., Babon, J. J., Stannard, J. A., Singh, S., Nizet, V., Carlberg, J. M., Ostland, V. E., Pennington, M. W., Norton, R. S., and Westerman, M. E. (2005). Bass hepcidin synthesis, solution structure, antimicrobial activities and synergism, and in vivo hepatic response to bacterial infections. *Journal of Biological Chemistry*, 280(10):9272–9282.

[Lee et al., 2011] Lee, A. P., Kerk, S. Y., Tan, Y. Y., Brenner, S., and Venkatesh, B. (2011). Ancient vertebrate conserved noncoding elements have been evolving rapidly in teleost fishes. *Molecular Biology and Evolution*, 28(3):1205–1215.

[Lee et al., 2005] Lee, Y., Tsai, J., Sunkara, S., Karamycheva, S., Pertea, G., Sultana, R., Antonescu, V., Chan, A., Cheung, F., and Quackenbush, J. (2005). The TIGR gene indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucleic Acids Research*, 33(Database Issue):D71–74. PMID: 15608288 PMCID: 540018.

[Lemaitre et al., 1996] Lemaitre, B., Nicolas, E., Michaut, L., Reichhart, J., and Hoffmann, J. A. (1996). The dorsoventral regulatory gene cassette sptzle/-Toll/cactus controls the potent antifungal response in drosophila adults. *Cell*, 86(6):973–983.

[Lerdorf et al., 2006] Lerdorf, R., Tatroe, K., and MacIntyre, P. (2006). *Programming PHP*. O'Reilly Media, second edition edition.

[Levasseur and Pontarotti, 2011] Levasseur, A. and Pontarotti, P. (2011). The role of duplications in the evolution of genomes highlights the need for evolutionary-based approaches in comparative genomics. *Biology Direct*, 6(1).

[Lieschke and Trede, 2009] Lieschke, G. J. and Trede, N. S. (2009). Fish immunology. *Current Biology*, 19(16):R678–682.

[Litman et al., 2010] Litman, G. W., Rast, J. P., and Fugmann, S. D. (2010). The origins of vertebrate adaptive immunity. *Nat Rev Immunol*, 10(8):543–553.

[Lynch and Conery, 2003] Lynch, M. and Conery, J. S. (2003). The origins of genome complexity. *Science*, 302(5649):1401–1404.

[Matsushita et al., 2004] Matsushita, M., Matsushita, A., Endo, Y., Nakata, M., Kojima, N., Mizuochi, T., and Fujita, T. (2004). Origin of the classical complement pathway: Lamprey orthologue of mammalian c1q acts as a lectin. *Proceedings of the National Academy of Sciences of the United States of America*, 101(27):10127–10131.

[Matzinger, 2002] Matzinger, P. (2002). The danger model: A renewed sense of self. *Science*, 296(5566):301–305.

[Medzhitov, 2009] Medzhitov, R. (2009). Approaching the asymptote: 20 years later. *Immunity*, 30(6):766–775.

[Mei et al., 2008a] Mei, J., Chen, B., Yue, H., and Gui, J. (2008a). Identification of a c1q family member associated with cortical granules and follicular cell apoptosis in carassius auratus gibelio. *Molecular and Cellular Endocrinology*, 289(1-2):67–76.

[Mei and Gui, 2008] Mei, J. and Gui, J. (2008). Bioinformatic identification of genes encoding c1q-domain-containing proteins in zebrafish. *Journal of Genetics and Genomics*, 35(1):17–24.

[Mei et al., 2008b] Mei, J., Zhang, Q., Li, Z., Lin, S., and Gui, J. (2008b). C1q-like inhibits p53-mediated apoptosis and controls normal hematopoiesis during zebrafish embryogenesis. *Developmental Biology*, 319(2):273–284.

[Meng et al., 2008] Meng, X., Noyes, M. B., Zhu, L. J., Lawson, N. D., and Wolfe, S. A. (2008). Targeted gene inactivation in zebrafish using engineered zinc-finger nucleases. *Nature Biotechnology*, 26(6):695–701. PMID: 18500337.

[Metchnikoff, 1905] Metchnikoff, E. (1905). *Immunity in infective diseases*. University Press.

[Milne et al., 2009] Milne, I., Lindner, D., Bayer, M., Husmeier, D., McGuire, G., Marshall, D. F., and Wright, F. (2009). TOPALi v2: a rich graphical interface for evolutionary analyses of multiple alignments on HPC clusters and multi-core desktops. *Bioinformatics*, 25(1):126–127. PMID: 18984599 PMCID: 2638937.

[Miura et al., 2006] Miura, E., Iijima, T., Yuzaki, M., and Watanabe, M. (2006). Distinct expression of cbln family mRNAs in developing and adult mouse brains. *European Journal of Neuroscience*, 24(3):750–760.

[Mugnaini and Morgan, 1987] Mugnaini, E. and Morgan, J. I. (1987). The neuropeptide cerebellin is a marker for two similar neuronal circuits in rat brain. *Proceedings of the National Academy of Sciences*, 84(23):8692–8696.

[Murai et al., 1990] Murai, T., Kodama, H., Naiki, M., Mikami, T., and Izawa, H. (1990). Isolation and characterization of rainbow trout c-reactive protein. *Developmental & Comparative Immunology*, 14(1):49–58.

[Murphy et al., 2008] Murphy, K. P., Travers, P., Walport, M., and Janeway, C. (2008). *Janeway's immuno biology.* Garland Science, New York, 7th ed. / edition.

[Nakamura et al., 2009] Nakamura, O., Wada, Y., Namai, F., Saito, E., Araki, K., Yamamoto, A., and Tsutsui, S. (2009). A novel c1q family member with fucose-binding activity from surfperch, neoditrema ransonnetii (Perciformes, embiotocidae). *Fish & Shellfish Immunology*, 27(6):714–720.

[Nakao et al., 2011] Nakao, M., Tsujikura, M., Ichiki, S., Vo, T. K., and Somamoto, T. (2011). The complement system in teleost fish: Progress of post-homolog-hunting researches. *Developmental & Comparative Immunology*, In Press, Corrected Proof.

[Nayak et al., 2010] Nayak, A., Ferluga, J., Tsolaki, A. G., and Kishore, U. (2010). The non-classical functions of the classical complement pathway recognition subcomponent c1q. *Immunology Letters*, 131(2):139–150.

[Nayak, 2010] Nayak, S. (2010). Probiotics and immunity: A fish perspective. *Fish & Shellfish Immunology*, 29(1):2–14.

[Ndong et al., 2007] Ndong, D., Chen, Y., Lin, Y., Vaseeharan, B., and Chen, J. (2007). The immune response of tilapia oreochromis mossambicus and its susceptibility to streptococcus iniae under stress in low and high temperatures. *Fish & Shellfish Immunology*, 22(6):686–694.

[Nelson, 2006] Nelson, J. S. (2006). *Fishes of the World.* Wiley, 4 edition.

[Nonaka and Kimura, 2006] Nonaka, M. and Kimura, A. (2006). Genomic view of the evolution of the complement system. *Immunogenetics*, 58(9):701–713.

[Palti, 2011] Palti, Y. (2011). Toll-like receptors in bony fish: From genomics to function. *Developmental & Comparative Immunology*, In Press, Corrected Proof.

[Pancer and Cooper, 2006] Pancer, Z. and Cooper, M. D. (2006). The evolution of adaptive immunity. *Annual Review of Immunology*, 24:497–518. PMID: 16551257.

[Pardo et al., 2008] Pardo, B., Fernandez, C., Millan, A., Bouza, C., Vazquez-Lopez, A., Vera, M., Alvarez-Dios, J., Calaza, M., Gomez-Tato, A., Vazquez, M., Cabaleiro, S., Magarinos, B., Lemos, M., Leiro, J., and Martinez, P. (2008). Expressed sequence tags (ESTs) from immune tissues of turbot (Scophthalmus maximus) challenged with pathogens. *BMC Veterinary Research*, 4(1).

[Pasnik and Smith, 2006] Pasnik, D. J. and Smith, S. A. (2006). Immune and histopathologic responses of DNA-vaccinated hybrid striped bass morone saxatilis x m. chrysops after acute mycobacterium marinum infection. *Diseases of Aquatic Organisms*, 73(1):33–41. PMID: 17240750.

[Padassi et al., 2008] Padassi, H., Tacnet-Delorme, P., Garlatti, V., Darnault, C., Ghebrehiwet, B., Gaboriaud, C., Arlaud, G. J., and Frachet, P. (2008). C1q binds phosphatidylserine and likely acts as a Multiligand-Bridging molecule in apoptotic cell recognition. *The Journal of Immunology*, 180(4):2329–2338.

[Peatman et al., 2007] Peatman, E., Baoprasertkul, P., Terhune, J., Xu, P., Nandi, S., Kucuktas, H., Li, P., Wang, S., Somridhivej, B., Dunham, R., and Liu, Z. (2007). Expression analysis of the acute phase response in channel catfish (Ictalurus punctatus) after infection with a gram-negative bacterium. *Developmental & Comparative Immunology*, 31(11):1183–1196.

[Peerschke and Ghebrehiwet, 2007] Peerschke, E. I. B. and Ghebrehiwet, B. (2007). The contribution of gC1qR/p33 in infection and inflammation. *Immunobiology*, 212(4-5):333–342. PMID: 17544818.

[Peerschke et al., 2004] Peerschke, E. I. B., Petrovan, R. J., Ghebrehiwet, B., and Ruf, W. (2004). Tissue factor pathway inhibitor-2 (TFPI-2) recognizes the complement and kininogen binding protein gC1qR/p33 (gC1qR): implications for vascular inflammation. *Thrombosis and Haemostasis*, 92(4):811–819. PMID: 15467913.

[Pennisi, 2005] Pennisi, E. (2005). Why do humans have so few genes? *Science*, 309(5731).

[Pettersen et al., 2004] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004). UCSF ChimeraA visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612.

[Phelps et al., 2009] Phelps, H. A., Runft, D. L., and Neely, M. N. (2009). Adult zebrafish model of streptococcal infection. *Current protocols in microbiology*, CHAPTER:Unit9–1. PMID: 19412913 PMCID: 2714046.

[Pinto et al., 2007] Pinto, M. R., Melillo, D., Giacomelli, S., Sfyroera, G., and Lambris, J. D. (2007). Ancient origin of the complement system: emerging invertebrate models. *Advances in Experimental Medicine and Biology*, 598:372–388. PMID: 17892225.

[Podell and Gaasterland, 2007] Podell, S. and Gaasterland, T. (2007). DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biology*, 8(2):R16–16. PMID: 17274820 PMCID: 1852411.

[Postlethwait et al., 1998] Postlethwait, J. H., Yan, Y., Gates, M. A., Horne, S., Amores, A., Brownlie, A., Donovan, A., Egan, E. S., Force, A., Gong, Z., Goutel, C., Fritz, A., Kelsh, R., Knapik, E., Liao, E., Paw, B., Ransom, D., Singer, A., Thomson, M., Abduljabbar, T. S., Yelick, P., Beier, D., Joly, J., Larhammar, D., Rosa, F., Westerfield, M., Zon, L. I., Johnson, S. L., and Talbot, W. S. (1998). Vertebrate genome evolution and the zebrafish gene map. *Nat Genet*, 18(4):345–349.

[Quackenbush et al., 2001] Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., Parvizi, B., Pertea, G., Sultana, R., and White, J. (2001). The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Research*, 29(1):159–164. PMID: 11125077.

[Quaye, 2008] Quaye, I. K. (2008). Haptoglobin, inflammation and disease. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 102(8):735–742.

[Raida and Buchmann, 2009] Raida, M. K. and Buchmann, K. (2009). Innate immune response in rainbow trout (Oncorhynchus mykiss) against primary and secondary infections with yersinia ruckeri o1. *Developmental & Comparative Immunology*, 33(1):35–45.

[Ravi and Venkatesh, 2008] Ravi, V. and Venkatesh, B. (2008). Rapidly evolving fish genomes and teleost diversity. *Current Opinion in Genetics & Development*, 18(6):544–550.

[Roest Crollius and Weissenbach, 2005] Roest Crollius, H. and Weissenbach, J. (2005). Fish genomics and biology. *Genome Research*, 15(12):1675–1682.

[Ronquist and Huelsenbeck, 2003] Ronquist, F. and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574.

[Rucinski et al., 2009] Rucinski, M., Ziolkowska, A., Szyszka, M., and Malendowicz, L. K. (2009). Precerebellin-related genes and precerebellin 1 peptide in the adrenal gland of the rat: expression pattern, localization, developmental regulation and effects on corticosteroidogenesis. *International Journal of Molecular Medicine*, 23(3):363–371. PMID: 19212655.

[Rykunov et al., 2008] Rykunov, D., Steinberger, E., Madrid-Aliste, C. J., and Fiser, A. (2008). Improved scoring function for comparative modeling using the M4T method. *Journal of Structural and Functional Genomics*, 10(1):95–99.

[Sandelin et al., 2004] Sandelin, A., Bailey, P., Bruce, S., Engstrom, P., Klos, J., Wasserman, W., Ericson, J., and Lenhard, B. (2004). Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics*, 5(1).

[Sastry et al., 1991] Sastry, K., Zahedi, K., Lelias, J. M., Whitehead, A. S., and Ezekowitz, R. A. (1991). Molecular characterization of the mouse mannose-binding proteins. the mannose-binding protein a but not c is an acute phase reactant. *The Journal of Immunology*, 147(2):692–697.

[Schierwater et al., 2009] Schierwater, B., Eitel, M., Jakob, W., Osigus, H., Hadrys, H., Dellaporta, S. L., Kolokotronis, S., and DeSalle, R. (2009). Concatenated analysis sheds light on early metazoan evolution and fuels a modern Urmetazoon hypothesis. *PLoS Biol*, 7(1).

[Schmalstieg and Goldman, 2008] Schmalstieg, Frank C, J. and Goldman, A. S. (2008). Ilya ilich metchnikoff (1845-1915) and paul ehrlich (1854-1915): the centennial of the 1908 nobel prize in physiology or medicine. *Journal of Medical Biography*, 16(2):96–103. PMID: 18463079.

[Schmittgen and Livak, 2008] Schmittgen, T. D. and Livak, K. J. (2008). Analyzing real-time PCR data by the comparative CT method. *Nat. Protocols*, 3(6):1101–1108.

[Shapiro and Scherer, 1998] Shapiro, L. and Scherer, P. E. (1998). The crystal structure of a complement-1q family protein suggests an evolutionary link to tumor necrosis factor. *Current Biology*, 8(6):335–340.

[Sjwall et al., 2007] Sjwall, C., Wetter, J., Bengtsson, T., Askendal, A., Almroth, G., Skogh, T., and Tengvall, P. (2007). Solid-phase classical complement activation by c-reactive protein (CRP) is inhibited by fluid-phase CRP-C1q interaction. *Biochemical and Biophysical Research Communications*, 352(1):251–258.

[Solomon et al., 2007] Solomon, S., Xu, Y., Wang, B., David, M. D., Schubert, P., Kennedy, D., and Schrader, J. W. (2007). Distinct structural features ofCaprin-1 mediate its interaction with G3BP-1 and its induction of phosphorylation of eukaryotic translation initiation factor 2alpha, entry to cytoplasmic stress granules, and selective interaction with a subset of mRNAs. *Mol. Cell. Biol.*, 27(6):2324–2342.

[Stapleton et al., 2004] Stapleton, T., Luchman, A., Johnston, J., Browder, L., Brenner, S., Venkatesh, B., and Jirik, F. R. (2004). Compact intergenic regions

of the pufferfish genome facilitate isolation of gene promoters: characterization of fugu 3'-phosphoadenosine 5'-phosphosulfate synthase 2 (fPapss2) gene promoter function in transgenic xenopus. *FEBS Letters*, 556(1-3):59–63.

[Stevens et al., 2007] Stevens, B., Allen, N. J., Vazquez, L. E., Howell, G. R., Christopherson, K. S., Nouri, N., Micheva, K. D., Mehalow, A. K., Huberman, A. D., Stafford, B., Sher, A., Litke, A. M., Lambris, J. D., Smith, S. J., John, S. W. M., and Barres, B. A. (2007). The classical complement cascade mediates CNS synapse elimination. *Cell*, 131(6):1164–1178.

[Su et al., 2006] Su, Z., Wang, J., Yu, J., Huang, X., and Gu, X. (2006). Evolution of alternative splicing after gene duplication. *Genome Research*, 16(2):182–189. PMID: 16365379 PMCID: 1361713.

[Sullivan et al., 2009] Sullivan, C., Charette, J., Catchen, J., Lage, C. R., Giasson, G., Postlethwait, J. H., Millard, P. J., and Kim, C. H. (2009). The gene history of zebrafish tlr4a and tlr4b is predictive of their divergent functions. *Journal of immunology (Baltimore, Md. : 1950)*, 183(9). PMID: 19812203 PMCID: 2819326.

[Sullivan and Kim, 2008] Sullivan, C. and Kim, C. H. (2008). Zebrafish as a model for infectious disease and immune function. *Fish & Shellfish Immunology*, 25(4):341–350.

[Swain and Nayak, 2009] Swain, P. and Nayak, S. (2009). Role of maternally derived immunity in fish. *Fish & Shellfish Immunology*, 27(2):89–99.

[Tahtouh et al., 2009] Tahtouh, M., Croq, F., Vizioli, J., Sautiere, P., Van Camp, C., Salzet, M., Daha, M. R., Pestel, J., and Lefebvre, C. (2009). Evidence for a novel chemotactic c1q domain-containing factor in the leech nerve cord. *Molecular Immunology*, 46(4):523–531.

[Tauber, 2003] Tauber, A. I. (2003). Metchnikoff and the phagocytosis theory. *Nat Rev Mol Cell Biol*, 4(11):897–901.

[Tom Tang et al., 2005] Tom Tang, Y., Hu, T., Arterburn, M., Boyle, B., Bright, J. M., Palencia, S., Emtage, P. C., and Funk, W. D. (2005). The complete complement of c1q-domain-containing proteins in homo sapiens. *Genomics*, 86(1):100–111.

[Traver et al., 2003] Traver, D., Herbomel, P., Patton, E. E., Murphey, R. D., Yoder, J. A., Litman, G. W., Catic, A., Amemiya, C. T., Zon, L. I., and Trede, N. S. (2003). The zebrafish as a model organism to study development of the immune system. *Advances in Immunology*, 81:253–330. PMID: 14711058.

[Umrath and Silberbauer, 1967] Umrath, K. and Silberbauer, I. (1967). [Effect of reserpine and chlordiazepoxide on bound cerebellin, an excitatory transmitter substance of the cerebellum]. *Zeitschrift Fr Biologie*, 115(6):417–421. PMID: 4386812.

[Urade et al., 1991] Urade, Y., Oberdick, J., Molinar-Rode, R., and Morgan, J. I. (1991). Precerebellin is a cerebellum-specific protein with similarity to the globular domain of complement c1q b chain. *Proceedings of the National Academy of Sciences*, 88(3):1069–1073.

[Vegh et al., 2006] Vegh, Z., Kew, R. R., Gruber, B. L., and Ghebrehiwet, B. (2006). Chemotaxis of human monocyte-derived dendritic cells to complement component c1q is mediated by the receptors gC1qR and cC1qR. *Molecular Immunology*, 43(9):1402–1407.

[Venkatesh et al., 2000] Venkatesh, B., Gilligan, P., and Brenner, S. (2000). Fugu: a compact vertebrate reference genome. *FEBS Letters*, 476(1-2):3–7.

[Visel et al., 2007] Visel, A., Bristow, J., and Pennacchio, L. A. (2007). Enhancer identification through comparative genomics. *Seminars in Cell & Developmental Biology*, 18(1):140–152.

[Wagner, 2008] Wagner, A. (2008). Robustness and evolvability: a paradox resolved. *Proceedings of the Royal Society B: Biological Sciences*, 275(1630):91–100.

[Wall et al., 2000] Wall, L., Christiansen, T., and Orwant, J. (2000). *Programming Perl*. O'Reilly Media, Inc.

[Wei et al., 2007] Wei, P., Smeyne, R. J., Bao, D., Parris, J., and Morgan, J. I. (2007). Mapping of cbln1like immunoreactivity in adult and developing mouse brain and its localization to the endolysosomal compartment of neurons. *European Journal of Neuroscience*, 26(10):2962–2978.

[Westerfield, 2007] Westerfield, M. (2007). *The zebrafish book : a guide for the laboratory use of zebrafish (Danio rerio)*. Univ. of Oregon Press, Eugene OR, 5th ed. edition.

[Wommack et al., 1999] Wommack, K. E., Ravel, J., Hill, R. T., Chun, J., and Colwell, R. R. (1999). Population dynamics of chesapeake bay virioplankton: Total-Community analysis by Pulsed-Field gel electrophoresis. *Applied and Environmental Microbiology*, 65(1):231–240. PMID: 9872784 PMCID: 91007.

[Woolfe et al., 2004] Woolfe, A., Goodson, M., Goode, D. K., Snell, P., McEwen, G. K., Vavouri, T., Smith, S. F., North, P., Callaway, H., Kelly, K., Walter, K., Abnizova, I., Gilks, W., Edwards, Y. J. K., Cooke, J. E., and Elgar, G.

(2004). Highly conserved Non-Coding sequences are associated with vertebrate development. *PLoS Biol*, 3(1).

[Yan et al., 2009] Yan, Q., Zhang, J., Zou, W., Chen, Q., Zhuang, Z., and Wang, X. (2009). Immune response of pseudosciaena crocea to the injection of vibrio alginolyticus. *Chinese Journal of Oceanology and Limnology*, 27(1):85–91.

[Yanai et al., 2005] Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., Lancet, D., and Shmueli, O. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, 21(5):650–659.

[Yoder et al., 2002] Yoder, J. A., Nielsen, M. E., Amemiya, C. T., and Litman, G. W. (2002). Zebrafish as an immunological model system. *Microbes and Infection*, 4(14):1469–1478.

[Yuzaki, 2008] Yuzaki, M. (2008). Cbln and c1q family proteins new transneuronal cytokines. *Cellular and Molecular Life Sciences*, 65(11):1698–1705.

[Yuzaki, 2009] Yuzaki, M. (2009). New (but old) molecules regulating synapse integrity and plasticity: Cbln1 and the [delta]2 glutamate receptor. *Neuroscience*, 162(3):633–643.

[Zaccone, 2009] Zaccone, G. (2009). *Fish Defenses Vol. 1: Immunology*. Science Publishers.

[Zhang et al., 2008a] Zhang, H., Song, L., Li, C., Zhao, J., Wang, H., Qiu, L., Ni, D., and Zhang, Y. (2008a). A novel c1q-domain-containing protein from zhikong scallop chlamys farreri with lipopolysaccharide binding activity. *Fish & Shellfish Immunology*, 25(3):281–289.

[Zhang et al., 2008b] Zhang, Q., Zmasek, C. M., Dishaw, L. J., Mueller, M. G., Ye, Y., Litman, G. W., and Godzik, A. (2008b). Novel genes dramatically alter regulatory network topology in amphioxus. *Genome Biology*, 9(8). PMID: 18680598.

[Zhang et al., 2011] Zhang, Y., Salinas, I., and Oriol Sunyer, J. (2011). Recent findings on the structure and function of teleost IgT. *Fish & Shellfish Immunology*, In Press, Corrected Proof.