ABSTRACT

WCISEL, DUSTIN JAMES. An Evolutionary Perspective of Innate Immune Receptor Diversity. (Under the direction of Dr. Jeffrey A. Yoder).

The zebrafish (*Danio rerio*) is a useful model for studying gene function and evolution and has recently gained momentum as a model for studying the pathogenesis of and immune response to infectious diseases. Zebrafish share many of the hallmark features of mammalian immunity. This includes an adaptive defense which utilizes rearranging immunoglobulin genes that encode antibodies, as well as non-rearranging innate immune receptors such as Toll-like receptors (TLRs) and immunoglobulin (Ig) domain-containing innate immune receptors (IIIRs). Zebrafish also encode fish-specific families of IIIRs, such as novel immune-type receptors (NITRs), diverse immunoglobulin domain-containing proteins (DICPs), polymeric Ig receptor-like (PIGRLs) and novel immunoglobulin-like transcripts (NILTs). Within different fish lineages, these multigene families have undergone recent and rapid gene birth and death events and subsequent sequence diversification. In addition to these species-specific variations, multiple gene-content haplotypes are observed for IIIR gene clusters within a single species. This variation likely plays an important role in the immune response. However, the contribution of gene-content haplotypes to disease susceptibilities remains under-appreciated in most immunological studies.

In order to bridge the evolutionary gap between humans and distantly related vertebrate models such as zebrafish, the genome of the spotted gar (*Lepisosteus oculatus*) was recently sequenced. While both gar and zebrafish are bony-fish (Actinopterygii), gar is an outgroup to the more than 30,000 teleosts (including zebrafish) that experienced an ancient teleost-specific whole genome duplication (TGD) event. During the process of rediploidization, duplicated genomes acquire deleterious mutations and undergo chromosomal rearrangements resulting in a highly rearranged genome which confounds the identification of homologous genes between species. Therefore, gar has great potential to aid in translational studies as its genome more closely reflects the ancient genome organization shared by the ancestors of all vertebrates. By identifying and defining multiple innate
immune gene families in gar, important evolutionary comparisons between zebrafish, gar and human are possible.

Chapter 2 provides a detailed description of the extensive polymorphic and haplotypic variation of the zebrafish DICP gene cluster. DICP sequences also were identified in the genomes of spotted gar and coelacanth (*Latimeria chalumnae*). Evidence is provided demonstrating that DICP genes are physically linked to major histocompatibility complex (MHC complex) genes in zebrafish, gar and coelacanth indicating that this relationship has survived more than 450 million years of evolution and may reflect a functional interaction.

NILTs were previously reported in three teleost species with two genes predicted from the zebrafish reference genome (genome version Zv5). The work described in Chapter 3 identified 104 NILT-related Ig domains in the genomic region of zebrafish chromosome 1 (genome version Zv9) surrounding the two predicted NILTs. While many NILT Ig domains in this region remain unannotated, transcripts belonging to 25 zebrafish NILTs genes were identified. In addition, alternative haplotypes, reflecting significant gene-content variation, were described for the gene cluster in zebrafish.

In Chapter 4, an analyses of select innate immune genes revealed that the gar genome encodes a larger number of TLRs than mammals, overturning expectations that increased numbers of TLR genes in teleosts was a result of the TGD. In addition, the identification of NITRs in the gar genome demonstrated, for the first time, that this family of IIIRs is not restricted to teleost species and therefore, is more ancient than previously predicted.

The combination of sequence orthology and functional analyses has the potential to greatly increase our understanding of the evolution of immune genes and our ability to use zebrafish to model human diseases. Future functional experiments investigating the role of different IIIRs in immune function will undoubtedly reveal many important features of these proteins and will broaden our understanding of their evolution within all vertebrate lineages.
An Evolutionary Perspective of Innate Immune Receptor Diversity

by
Dustin Wcisel

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Functional Genomics

Raleigh, North Carolina

2017

APPROVED BY:

________________________________________  _______________________________________
Jeffrey Yoder                          Matthew Breen
Committee Chair

________________________________________  _______________________________________
Michael Sikes                          Jeffrey Thorne
BIOGRAPHY

Dustin grew up on the sunrise side of Northern Michigan, in a small town named Alpena. The second oldest of five children, Dustin spent a good deal of his childhood fishing and playing sports with his siblings and neighbors. Dustin’s high school years were shaped by a friend group with a common interest in building fast but affordable computers. In 2006, he moved 250 miles away from family to pursue a bachelor's degree in Cellular and Molecular Biology at Grand Valley State University in Grand Rapids, MI. After obtaining his degree, he stayed in the area and worked as a lab manager in a population genetics lab researching invasive aquatic species. While he refined his basic lab skills, exciting and powerful technological advances in genome sequencing were emerging, the use of which would combine Dustin’s molecular skill-set with his love of computers.

To be on the forefront of these emerging technologies, Dustin decided to pursue a degree in Genomic Sciences: Functional Genomics from North Carolina State University in 2012. Dustin chose to join the Yoder lab, as there were many projects afoot which utilized next-generation sequencing technology to generate complex data sets. Defining the characteristics of complex immune gene family in zebrafish was chosen as his dissertation topic. Very quickly, the project expanded to include many additional immune gene families, as well as many other species. Dustin hopes to continue his research on the evolution of genes and genomes with an eye on understanding human disease.

When Dustin is not chasing bugs in his latest programming script to analyze a data set, he enjoys spending time with his fiancée Mo (aka Kristin) and his dog Ollie. On any given night, you’re likely to run into the three at a kickball/softball field, a campsite in the mountains, a fishing boat in a lake, or a brewery enjoying a stout with friends.
ACKNOWLEDGMENTS

I am immensely grateful to my graduate advisor, Dr. Jeff Yoder for not only allotting me the freedom, but also encouraging me to explore my research interests. Dr. Yoder gave me the time, advice and encouragement to develop computer-based research skills to compliment my molecular biology training. I am grateful to my distinguished committee members, Dr. Matthew Breen, Dr. Michael Sikes and Dr. Jeffrey Thorne for their unwavering patience, flexibility and guidance as my dissertation evolved. I would also like to thank the members of the Yoder Lab for their friendship, help, and especially for tolerating my sense of humor. Dr. Yoder’s lab environment was the absolute perfect fit for me and I would not have been nearly as successful anywhere else.

Extremely influential to both who I am as a scientist and as a person, is my undergraduate advisor, Dr. Ryan Thum. He provided advice at a critical time in my life and is a role model of what it means to be a responsible scientist and humanitarian. I would not be who I am today if not for the guidance of Dr. Thum.

Most importantly of all, none of this would have been possible without the constant love and support of my fiancée, Mo (aka Kristin). Mo, I am forever indebted to you for all you have done and all of the sacrifices you have made to allow me to pursue my goals.
**TABLE OF CONTENTS**

**LIST OF TABLES**

<table>
<thead>
<tr>
<th>Chapter or Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Tables</td>
<td>v</td>
</tr>
</tbody>
</table>

**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>Chapter or Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Figures</td>
<td>vi</td>
</tr>
</tbody>
</table>

**CHAPTER 1: INTRODUCTION**

<table>
<thead>
<tr>
<th>Chapter or Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>History of the zebrafish reference genome</td>
<td>1</td>
</tr>
<tr>
<td>Whole genome duplication</td>
<td>2</td>
</tr>
<tr>
<td>Gene-content haplotypes</td>
<td>5</td>
</tr>
<tr>
<td>Outline of dissertation</td>
<td>6</td>
</tr>
<tr>
<td>References</td>
<td>8</td>
</tr>
</tbody>
</table>

**CHAPTER 2: THE IDENTIFICATION OF ADDITION ALY ZEBRAFISH DICP GENES REVEALS HAPLOTYPE VARIATION AND LINKAGE TO MHC CLASS I GENES**

<table>
<thead>
<tr>
<th>Chapter or Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>12</td>
</tr>
<tr>
<td>Methods and materials</td>
<td>13</td>
</tr>
<tr>
<td>Results</td>
<td>14</td>
</tr>
<tr>
<td>Discussion</td>
<td>16</td>
</tr>
<tr>
<td>References</td>
<td>25</td>
</tr>
</tbody>
</table>

**CHAPTER 3: IDENTIFICATION OF NOVEL IMMUNOGLOBULIN-LIKE TRANSCRIPTS AND GENE CONTENT-HAPLOTYPES IN ZEBRAFISH**

<table>
<thead>
<tr>
<th>Chapter or Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>31</td>
</tr>
<tr>
<td>Methods</td>
<td>32</td>
</tr>
<tr>
<td>Results and discussion</td>
<td>34</td>
</tr>
<tr>
<td>Conclusions</td>
<td>38</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>48</td>
</tr>
<tr>
<td>References</td>
<td>49</td>
</tr>
</tbody>
</table>

**CHAPTER 4: SPOTTED GAR AND THE EVOLUTION OF INNATE IMMUNE RECEPTORS**

<table>
<thead>
<tr>
<th>Chapter or Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>53</td>
</tr>
<tr>
<td>Methods</td>
<td>53</td>
</tr>
<tr>
<td>Results and discussion</td>
<td>54</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>68</td>
</tr>
<tr>
<td>Literature Cited</td>
<td>68</td>
</tr>
</tbody>
</table>

**CHAPTER 5: SUMMARY AND FUTURE DIRECTIONS**

<table>
<thead>
<tr>
<th>Chapter or Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diverse immunoglobulin domain-containing proteins (DICPs)</td>
<td>72</td>
</tr>
<tr>
<td>Novel immunoglobulin-like transcripts (NILTs)</td>
<td>73</td>
</tr>
<tr>
<td>Lessons from spotted gar</td>
<td>73</td>
</tr>
<tr>
<td>Implications and future directions</td>
<td>74</td>
</tr>
<tr>
<td>Concluding thoughts</td>
<td>77</td>
</tr>
<tr>
<td>References</td>
<td>78</td>
</tr>
</tbody>
</table>

**APPENDICES**

<table>
<thead>
<tr>
<th>Chapter or Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appendix A: Multigene Families of Immunoglobulin Domain-containing Innate Immune Receptors in Zebrafish: Deciphering the Differences</td>
<td>81</td>
</tr>
<tr>
<td>Appendix B: The Confounding Complexity of Innate Immune Receptors within and Between Teleost Species</td>
<td>82</td>
</tr>
<tr>
<td>Appendix C: Chapter 2 Supplemental Material</td>
<td>93</td>
</tr>
<tr>
<td>Appendix D: Chapter 3 Supplemental Material</td>
<td>104</td>
</tr>
<tr>
<td>Appendix E: Chapter 4 Supplemental Material</td>
<td>132</td>
</tr>
<tr>
<td>Appendix F: Generation of Targeted and Heritable CRISPR-Induced Mutations in Zebrafish</td>
<td>148</td>
</tr>
<tr>
<td>Appendix G: Chapter 5 Supplemental Material</td>
<td>150</td>
</tr>
</tbody>
</table>
LIST OF TABLES

CHAPTER 2:
  Table 1. Primer sequences and PCR cycling parameters for detection and recovery of DICP sequences...........................................................................................................................................................................17

CHAPTER 3:
  Table 1. 5'/3' RACE Primers..........................................................................................................................................................................................36

CHAPTER 4:
  Table 1. Gar Toll-like receptors (TLRs)........................................................................................................................................................................56
LIST OF FIGURES

CHAPTER 1:
Figure 1. The non-teleost Actinopterygii .................................................................3
Figure 2. A simplified schematic of genomic rearrangements that may follow whole genome duplication of four chromosomes .................................................................4

CHAPTER 2:
Figure 1. Overview of the oligonucleotide design employed for amplifying DICP sequences ... 16
Figure 2. DICP gene expression in immune-related tissues ..............................................18
Figure 3. DICP gene expression in myeloid and lymphoid cells ......................................19
Figure 4. DICP gene expression during zebrafish development .......................................20
Figure 5. Phylogenetic comparison of newly identified DICP Ig domains with previously described sequences ..................................................................................................................23
Figure 6. Exon-intron architecture of previously identified DICPs ......................................24
Figure 7. Exon-intron architecture of newly identified DICPs ...........................................25
Figure 8. Alternative haplotypes for the chromosome 3 DICP gene cluster .........................26

CHAPTER 3:
Figure 1. Sequence comparisons of previously reported NILT immunoglobulin domains .........39
Figure 2. Integrated Genomics Viewer display of Zv9 zebrafish chromosome one NILT region 41
Figure 3. Dot plot of alternative NILT haplotypes in zebrafish ..........................................41
Figure 4. Genotyping confirms hap100 and hap101 in zebrafish lines ..................................42
Figure 5. Multidimensional scaling of 104 zebrafish NILT Ig domains ..............................45
Figure 6. Genomic annotation of zebrafish NILT Ig domains and genes .........................46
Figure 7. NILT repertoire by haplotype .............................................................................47

CHAPTER 4:
Figure 1. Typical protein structures of TLRs and IIIRs ......................................................55
Figure 2. Phylogenetic relationships between gar, teleost, mammalian, avian, reptilian, and coelacanth TLR TIR domains .................................................................57
Figure 3. Conserved sequence features of gar NITRs .........................................................61
Figure 4. Genomic organization of NITR genes in spotted gar .........................................63
Figure 5. NITR sequence diversity between and within species .......................................64
Figure 6. Conserved sequence features of gar DICP D1 and D2 domains .........................65
Figure 7. Evolutionary relations of DICP D1 and D2 domains ........................................66
Figure 8. DICP genes of zebrafish, spotted gar, and coelacanth are linked to genes associated with the MHC .................................................................................................................67
Figure 9. Evolutionary context of TLRs, NITRs and DICPs among selected vertebrate lineages .........................................................................................................................68
CHAPTER 1: INTRODUCTION

Zebrafish (*Danio rerio*) gained favor as a vertebrate model as both genetic and experimental methods are easily applied. In the 1970s, George Streisinger at the University of Oregon began rearing zebrafish in a laboratory environment and established several lines of zebrafish, including two homozygous diploid lines (Streisinger et al. 1981). However, it wasn’t until Nobel laureate Christiane Nüsslein-Volhard at the Max-Planck Institute performed a large-scale screening of chemically-induced genomic mutations using zebrafish that the model caught the attention of the National Institute of Health (NIH) (Haffter et al. 1996). In an NIH sponsored workshop in 1999 to discuss vertebrate developmental models, zebrafish was ranked third behind human and mouse models (Rasooly et al. 2003). Zebrafish continued to prove its utility for gene knockdowns (Nasevicius and Ekker 2000) and over-expression studies (Kobayashi et al. 1998) eventually leading to the initiation of the zebrafish genome sequencing project by the Wellcome Trust Sanger Institute in 2001 (Vogel 2000).

History of the zebrafish reference genome

Since the first draft in 2002, the reference genome has gone through many rounds of improvement and is currently in its 11th rendition. The first three versions (Zv1 through Zv3) were based solely on whole-genome shotgun sequencing. Seeking quality to rival human and mouse reference genomes, clone-by-clone based sequencing was employed for versions Zv4 and later (Howe et al. 2013). Since the 10th version, the genome has been curated by the Genome Reference Consortium (GRC), an international collaboration that oversees updates and improvements to human, mouse, and zebrafish reference genomes and the naming scheme was changed to GRCz10 to reflect the change (Schneider and Church 2013).

Tübingen was the laboratory line of zebrafish initially selected for genome sequencing as it had been recently utilized in Dr. Nüsslein-Volhard’s mutagenesis screens (Haffter et al. 1996). This line was procured from a pet shop in Tübingen, Germany and was inbred for several generations.
Prior to whole genome shotgun sequencing, DNA from several individual zebrafish were combined. While this strategy made sense to capture polymorphisms, large haplotypic variations resulted in ambiguous assembled contigs as well as artifactual sequence duplications (Howe et al. 2013). Numerous haplotypes displaying gene content variation were reported as tandem duplications in early versions of the genome. For example, comparing sequenced bacterial artificial chromosomes used in the assembly of Zv4-current zebrafish reference genome led to the identification of many alternative haplotypes for the novel immune-type receptor (NITR) gene family (described in detail Appendix A and Yoder et al. (2004). Brown et al. (2012) estimated the degree of variation found within the Tübingen line was the highest among any zebrafish line examined and exceeded that found in the human population.

With the recent advances in genome sequencing technologies, many genome sequencing projects are realizing genome assemblies seeking a single consensus sequence are insufficient to represent the complexity of a species’ genome. The eleventh zebrafish reference genome, GRCz11 which was released in May of 2017, marked a major update to the reference genome and for the first time displayed alternative haplotypes. The extensive copy number variation found in zebrafish by Brown et al. (2012) and acknowledged by the GRC highlights the importance of considering the genetic background of the biological or medical research specimen. This genetic complexity may explain some differences in intra- and inter-species response. For example, different zebrafish lines respond differently to ethanol exposure during development (Dlugos and Rabin 2003; Loucks and Carvan 2004).

**Whole genome duplication**

Further complicating the use of zebrafish as a vertebrate model is the whole genome duplication (WGD) experienced by all teleost species (TGD) (Amores et al. 1998). Teleost species represent an expansion of ray-finned (Actinopterygii) fishes with more than 30,000 extant species including zebrafish and nine other species with genomes curated by Ensembl (version 90).
(Eschmeyer, Fricke, and van der Laan 2016; Aken et al. 2016). Only ~70 species of ray-finned fish that did not undergo the TGD are alive today (Figure 1). These species include the recently sequenced spotted gar (*Lepisosteus oculatus*), the only Actinopterygii available on Ensembl which did not undergo the TGD and the bowfin (*Amia calva*) with a sequenced genome anticipated by the end of 2017 (Braasch et al. 2016). Their basal position during Actinopterygii diversification and outgroup status of the TGD uniquely qualifies these species to further our understanding and use of teleost species as models of human diseases. For example, it was demonstrated the spotted gar can facilitate the identification of human orthologs in zebrafish (Braasch et al. 2016; Sullivan et al. 2017).

**Figure 1.** The non-teleost Actinopterygii. The Polypteriformes (bichirs and ropefishes) are the sister lineage of all other Actinopterygians. The Acipenseriformes (sturgeons and paddlefishes) are the sister lineage of Neopterygii, and Holostei (bowfin and gars) are the sister lineage of Teleostei (aka teleosts such as zebrafish, stickleback and medaka). Teleost whole genome duplication (TGD) was predicted to have occurred more than 300 MYA (Vandepoele et al. 2004), Figure adapted from (Near et al. 2012). Fish images from https://en.wikipedia.org/wiki/Actinopterygii.

WGDs can result in rapid rearrangement of chromosomes. Figure 2 shows the types and complexity of chromosomal rearrangements that can be expected as time progresses. It has been suggested that the zebrafish genome is the most highly rearranged of any sequenced fish genome, with both higher inter- and intra-chromosomal rearrangement rates (Sémon and Wolfe 2007a). This large degree of rearrangement complicates the identification of orthologous genes between species.
Gene duplication is thought to be one of, if not the, major driving force of evolution (Nei 1983, 14–37). WGDs afford a massive amount of genes the freedom from evolutionary pressures to maintain sequence and function. Duplicated genes resulting from WGD events are termed ohnologs after the seminal book *Evolution after Genome Duplication* by Susumo Ohno (1970). Most ohnologs quickly accumulate deleterious mutations and become pseudogenes (pseudogenization).

Subfunctionalization occurs when partial pseudogenization occurs in both ohnologs resulting in the required joint effort of both genes to retain the ancestral function (Lynch and Force 2000). Rarely, one duplicated copy maintains the ancestral function while the other acquires advantageous mutations resulting in novel functions (neofunctionalization) (Ohno 1970). Therefore, even with clear evidence of sequence homology, caution is warranted when extrapolating function between species, especially between species that have and have not experienced WGD—such as between humans and zebrafish.
Gene-content haplotypes

Rapid and recent diversification of multigene families may result in gene-content haplotypes. That is, genomes from individuals of the same species may encode different numbers and combinations of members of the same multigene family. One example of gene-content haplotypic variation observed in humans occurs within the killer-cell immunoglobulin-like receptor (KIR) gene cluster. KIRs are expressed on the cell surface of natural killer cells (NK cells) and recognize MHC class I (MHCI) molecules that are normally expressed on the surface of all normal cells as a marker of “self” (Stewart et al. 2005). KIRs with opposing signals for activating and inhibiting immune cells exist. When activating KIRs bind their ligand that can influence the NK cell to release cytotoxic granules and induce apoptosis in the target cell: inhibitory KIRs block this signaling pathway (Billadeau and Leibson 2002). Gene content haplotypes of human KIRs and polymorphic variation of MHCI have been implicated in disease susceptibility and resistance. For example, a heightened number of genome-encoded activating KIRs reduces the rate of cytomegalovirus (CMV) infection in kidney transplant recipients (Stern et al. 2008). A more complete list of diseases associated with KIR-MHCI haplotypes are provided in appendix B.

While MHCI haplotypes have been documented in zebrafish (Dirscherl et al. 2014; McConnell et al. 2016), genes orthologous to human KIRs have not been identified in teleost species. Zebrafish novel immune-type receptors (NITRs) have been proposed to be the “functional orthologs” of mammalian KIRs, but an evolutionary origin has not been established (J. A. Yoder et al. 2001). Several additional cell surface receptor families have been identified in teleost species but orthology to human genes remain to be resolved. These include the multigene families of diverse immunoglobulin domain-containing proteins (DICPs) and novel immunoglobulin-like transcripts (NILTs) (Haire et al. 2012; Stet et al. 2005). Collectively, these immunoglobulin domain-containing innate immune receptors are referred to as IIIRs (coined in Appendix A). The species distribution,
evolutionary origins, tissue and cellular expression patterns, function and significance of the sequence variation remain to be fully resolved for teleost IIIRs.

Outline of dissertation

A core part of my dissertation was the accurate identification of orthologous innate immune receptor genes between species. It is widely agreed that two complementary approaches to prove orthology are required: sequence homology and conserved gene synteny. There are many methods and algorithms to assess sequence homology. BLAST and all of its progeny (e.g., BLASTP, BLAT, WU-BLAST) is one such method and was applied liberally (Kent 2002; Mahram and Herbordt 2012; Lopez et al. 2003). Additionally, hidden Markov models (HMM), multiple sequence alignments (MSA), phylogenetic trees and principal component analyses were employed when appropriate [http://hmmer.org/, (Sievers and Higgins 2014; Kumar, Stecher, and Tamura 2016; Pelé et al. 2012)].

In Chapter 2, I explored available teleost genomes for orthologs of zebrafish diverse immunoglobulin domain-containing proteins (DICPs). Indeed, many candidate genes were identified and examined closely. Although many potential DICP sequences were identified in numerous species, we only confidently classified sequences as DICPs if they were found in genomic proximity to MHC class I. I found this was the case for two additional species closely related to zebrafish, but could not identify the linkage in other teleost species. Additionally, the focus of chapter 2 was on the unprecedented amount of sequence variation observed between DICP genes of different lines of zebrafish. Numerous DICP transcripts were identified that did not match the reference genome, highlighting the significance of these differences and emphasizing the need for complementary sequencing of additional zebrafish genomes.

Chapter 3 expands our knowledge of the IIIR gene family, novel immunoglobulin-like transcripts (NILTs). These genes were first identified in common Carp (Cyprinis carpio) and previously, only two genes were identified in zebrafish (Stet et al. 2005). This chapter reveals at least two NILT gene content haplotypes can be found in zebrafish. For the first time, I describe NILTs in a
whole genome context, which will aid in identifying NILTs in other species. Even in the absence of functional data, conserved sequence features of NILT genes suggests a role in immune function. This chapter continues the theme of prevalent gene-content haplotypes at immune loci in zebrafish.

I had the opportunity to assist in annotating select immune genes within the spotted gar genome. Gars are Actinopterygii that did not undergo the TGD and “facilitate teleost to human comparisons”, thereby, increasing our ability to use fish as model species while also bringing an evolutionary important species into the genomic era (Braasch et al. 2016). This resulted in a collaborative manuscript published in Nature Genetics as well as publishing Chapter 4 in a special issue of the Journal of Experimental Zoology: Part B Molecular Evolution (Wcisel et al. 2017). I identified, for the first time, the occurrence of NITRs in a species outside of teleost, radically changing the estimated age of this gene family. Spotted gar NITRs do not appear to have undergone the rapid gene birth/loss and diversification that characterizes zebrafish NITRs. Additionally, I confirmed reports by others that the recently sequenced genome of coelacanth (Latimeria chalumnae), a lobed-fin fish and not an Actinopterygii, contains DICPs and determined that they are encoded in close proximity to MHCI genes -- a defining feature of DICPs established in chapter 2 (Rodriguez-Nunez et al. 2016; Boudinot et al. 2014; Amemiya et al. 2013).

Finally, in Chapter 5, I highlight some of the future directions for which my research has laid the foundation. Even in the absence of functional data, the contribution of each research chapter provides a significant resource for investigating the association of molecular markers with disease susceptibility. These resources are also critical for the development and evaluation of alternative animal models as well as a “road map” for genetic breeding strategies such as those applied in aquaculture and conservation biology (Gjedrem and Robinson 2014; Stone 1991). Nevertheless, future studies seeking to understand the molecular function of these gene families will undoubtedly reveal important insights and contribute directly to our understanding of infectious disease and the evolution of the immune system.
References


CHAPTER 2: THE IDENTIFICATION OF ADDITIONAL ZEBRAFISH DICP GENES REVEALS HAPLOTYPE VARIATION AND LINKAGE TO MHC CLASS I GENES*

Published in *Immunogenetics* (2016) 68:295-312

DOI 10.1007/s00251-016-0901-6

Ivan Rodriguez-Nunez¹, Dustin J. Wcisel¹, Ronda T. Litman², Gary W. Litman²³, Jeffrey A. Yoder¹

1. Department of Molecular Biomedical Sciences and Center for Comparative Medicine and Translational Research North Carolina State University, Raleigh NC
2. Department of Pediatrics University of South Florida Morsani College of Medicine, USF/ACH Children’s Research Institute, St. Petersburg, FL
3. Department of Molecular Genetics All Children’s Hospital Johns Hopkins Medicine St. Petersburg, FL

*Dr. Ivan Rodriguez-Nunez conceived and performed all laboratory experiments and wrote the manuscript. Dustin participated in discussions about and revisions of the manuscript and applied bioinformatic methods to identify DICPs and MHC sequences in species other than zebrafish. Dustin generated the data presented in the section entitled “Linkage of DICP and MHC class I genes in cyprinid fishes” and displayed in Supplemental Figures S11, S12 and S13 (Appendix C).
The identification of additional zebrafish DICP genes reveals haplotype variation and linkage to MHC class I genes

Ivan Rodriguez-Nunez1 · Dustin J. Weisel1 · Ronda T. Litman1 · Gary W. Litman2,3 · Jeffrey A. Yoder4

Received: 27 August 2015 / Accepted: 7 January 2016 / Published online: 22 January 2016
© Springer-Verlag Berlin Heidelberg 2016

Abstract Bony fish encode multiple multi-gene families of membrane receptors that are comprised of immunoglobulin (Ig) domains and are predicted to function in innate immunity. One of these families, the diverse immunoglobulin (Ig) domain-containing protein (DICP) genes, maps to three chromosomal loci in zebrafish. Most DICPs possess one or two Ig ectodomains and include membrane-bound and secreted forms. Membrane-bound DICPs include putative inhibitory and activating receptors. Recombinant DICP Ig domains bind lipids with varying specificity, a characteristic shared with mammalian CD300 and TREM family members. Numerous DICP transcripts amplified from different lines of zebrafish did not match the zebrafish reference genome sequence suggesting polymorphic and haplotype variation. The expression of DICPs in three different lines of zebrafish has been characterized employing PCR-based strategies. Certain DICPs exhibit restricted expression in adult tissues whereas others are expressed ubiquitously. Transcripts of a subset of DICPs can be detected during embryonic development suggesting roles in embryonic immunity or other developmental processes. Transcripts representing 11 previously uncharacterized DICP sequences were identified. The assignment of two of these sequences to an unplaced genomic scaffold resulted in the identification of an alternative DICP haplotype that is linked to a MHC class I Z lineage haplotype on zebrafish chromosome 3. The linkage of DICP and MHC class I genes also is observable in the genomes of the related grass carp (Ctenopharyngodon idellus) and common carp (Cyprinus carpio) suggesting that this is a shared character with the last common Cyprinidae ancestor.

Keywords Innate immunity · Immunoglobulin domain · Haplotype · Polymorphism · Immune receptor

Introduction

Zebrafish (Danio rerio) are a well-recognized animal model in developmental biology, immunity and infection, toxicology, as well as cancer (Konantz et al. 2012; Novos and Figueras 2012; Renshaw and Trede 2012; Sipes et al. 2011; Sullivan and Kim 2008; Veldman and Lin 2008). Investigations in many different zebrafish lines indicate high levels of genetic variation, including copy number variants (CNVs) (Brown et al. 2012). Sequencing of the genome of a single wild-collected zebrafish and comparison to the reference genome revealed 5.2 million single nucleotide polymorphisms and over 1.6 million insertion-deletion variations (Patowary et al. 2013). This extensive genotype variation likely is reflected in phenotypic variation (Louvekis and Carván 2004).
Nearly one third of zebrafish genes shown to be conserved exclusively in the teleost lineage are predicted to encode immune response genes (based on software analyses of protein features) possibly reflecting a likely expansion of immune-related genes in teleost fish (Yang et al. 2013). Immune genes in zebrafish and other fish species are predicted to be under positive selection resulting in high levels of sequence variation (Aparicio et al. 2002; Patowary et al. 2013; Star et al. 2011). In addition to immunoglobulins and T-cell antigen receptors, zebrafish possess multiple gene families of immunoglobulin (Ig)-domain-containing putative innate immune receptors such as the novel immune-type receptors [NITRs (Yoder et al. 2001, 2004, 2008, 2010)], novel immunoglobulin-like transcripts [NILTs (Stet et al. 2005)], leukocyte immune-type receptor [LITRs (Stafford et al. 2006)], polymeric Ig receptor (pIgR)-like proteins [PIGRLs (Kortun et al. 2014)], and diverse immunoglobulin domain-containing proteins [DICPs (Haire et al. 2012)]. These gene families, of which some may be restricted to bony fish, are recently derived, rapidly evolving, and are associated with significant polymorphic and haplotypic variation (Haire et al. 2012; Rodriguez-Nunez et al. 2014; Yoder et al. 2010). The first report of a DICP transcript likely was from an EST project of the common carp (C. carpio) (Sakai et al. 2005) in which the sequence (GenBank AB988477) erroneously was referred to as a NTR. A subsequent report identified similar sequences on zebrafish chromosome 16 and referred to them as “NTR-Wxc” sequences to distinguish them from NITRs (Ohashi et al. 2010); however, only a single zebrafish “NTR-Wxc” (DICP) sequence was included in this report (GenBank XM_001345040). Recently, 27 DICP genes and pseudogenes were described on zebrafish chromosomes 3, 14, and 16 and were recognized to constitute a single derived multigene family constituting three distinct groups (Haire et al. 2012). The DICP family possesses two types of Ig cotodomains, D1 and D2, and individual DICPs are predicted to possess one (D1 or D2), two (D1-D2 orientation), or four (D1-D2-D1-D2 orientation) Ig domains. Multiple membrane-bound DICPs possess cytoplasmic immunoreceptor tyrosine-based inhibition motifs (ITIMs) consistent with inhibitory function. A single DICP (dicp2.1) lacks ITIMs but possesses a charged residue within its transmembrane domain indicating that it potentially could partner with an activating adaptor protein (e.g., Dap12, Fcγ, etc.). Membrane-bound DICPs lacking these characteristic peptide motifs and secreted DICPs also were identified. Polymorphisms and alternative mRNA processing were shown to contribute to DICP diversity. Recombinant DICP Ig domains bind phospholipids, a property shared with select Ig domains of the mammalian CD300 and TREM receptor families (Cannon et al. 2012; Haire et al. 2012). A specific functional role for DICPs as of yet has not been defined; however, the overall similarities in their structure and ligand recognition to CD300 and TREM proteins suggests that DICPs have a role in mediating innate immunity.

In order to better understand the transcriptional regulation and sequence variation of the DICP family, inter-individual variation of DICP cDNA amplicons within and among three lines of zebrafish (AB, TU, and EKW) have been characterized. DICP expression was evaluated from multiple tissues of individual zebrafish, including lymphoid and myeloid cell populations and at different stages of development. DICP amplicons were sequenced to determine polymorphisms and allelic variation between and within lines. Certain DICPs are shown to display restricted tissue expression, whereas others are expressed ubiquitously. Transcripts of several DICPs were detected during embryonic development. Additional DICP genes representing an alternative chromosome 3 haplotype that is linked to a MHC class I gene haplotype were described. DICP and MHC class I sequences were also found to be linked in the genomes of the related grass carp (Ctenopharyngodon idella) and common carp. Collectively, these findings highlight the sequence complexity and dynamic nature of the DICP family and suggest that DICP and MHC genes may be linked in all cyprinid fishes.

Materials and methods

Zebrafish

All experiments involving live zebrafish were performed in accordance with relevant institutional and national guidelines and regulations and were approved by the North Carolina State University Institutional Animal Care and Use Committee. TU and AB zebrafish were acquired from the Zebrafish International Resource Center (http://zebrafish.org). EKW zebrafish were purchased from EkkWhl Waterlife Resources (Ruskin, FL). Adult and embryonic zebrafish were maintained and euthanized as described (Jims et al. 2009).

RNA isolation and reverse transcription-PCR for transcript detection

Tissues were dissected from three individual TU, AB, and EKW zebrafish. Lymphoid and myeloid cell populations were isolated from pooled kidneys of five EKW zebrafish as described (Traver et al. 2003; Traver 2004). In brief, kidneys from adult zebrafish were dissected and homogenized with a 40-μm nylon-mesh filter in ice-cold PBS supplemented with 5% FBS. Propidium iodide was added to a concentration of 1 μg/ml. Myeloid and lymphoid cells were isolated from this single-cell suspension by sorting based on propidium iodide exclusion, forward scatter, and side scatter with a BD FACs Ara II SORP flow cytometer (Beckton Dickinson).
populations were sorted twice to optimize cell purity. Zebrafish embryos were collected by natural matings, maintained at 28 °C as described (Westerfield 2007), and ten embryos of each line were pooled at various developmental stages for RNA isolation. Total RNA was isolated using TRIzol (Invitrogen, Carlsbad, CA). Concentration and purity of RNAs were determined using a Nanodrop.

cDNAs were synthesized from total RNA (1.5 μg for tissues, 0.5 μg for lymphoid and myeloid cells, and 2.0 μg for embryos) using oligo dT primers and SuperScript III Reverse Transcriptase (Life Technologies, Carlsbad, CA). PCR primer pairs were designed by Primer 3 software (Untergasser et al. 2012) to amplify the Ig domains of several related members of the DICP family and to span at least one intron (Fig. 1 and Table 1). Relative gene expression levels were determined by PCR using Titanium™ Taq DNA polymerase (Clontech, Mountain View, CA); annealing temperatures, extension times, and number of cycles for each primer pair are listed in Table 1. Despite employing various cycling parameters, attempts to detect dcep1.3-dcep1.3-6 amplicons (primer pair ATGGCCTGATAGAGTCTCCGTCTTCTTGCG and GGAATAATCTCCCATGAAATTCTG) and dcep2.2 (primer pair ATGCCTGGGACTGATCCCCATTTCTCGC and GAATGCACGACTGCTGGTGGCTACGT) with the individual zebrafish in this study were unsuccessful (data not shown). A β-actin primer pair (Yoder et al. 2010) was used as a standard reference while a myeloperoxidase primer pair (MPO) (Yoder et al. 2010) and a T-cell receptor alpha primer pair (TCRα) (Wittumert et al. 2011) served as positive controls for lymphoid and lymphoid cells, respectively. Amplicons were cloned into the pGEM®-T-Easy plasmid (Promega) and sequenced.

Rapid amplification of cDNA ends (RACE) for amplification of DICP transcripts

A rapid amplification of cDNA ends (RACE) strategy was used to amplify and clone the 3’ ends of new DICP transcripts identified in this study. Total RNA from the kidney of TU zebrafish 1 and EK W zebrafish 2, the liver of EK W zebrafish 1, and AB zebrafish embryos at 36 h post-fertilization (hpf) were reverse transcribed using the GeneRacer™ oligo dT primer and SuperScript III Reverse Transcriptase supplied with the GeneRacer™ kit, and amplification strategies were conducted as recommended by the manufacturer (Invitrogen). An initial “touchdown” PCR (denaturing at 95 °C for 30 s, touchdown annealing at 70 to 68 °C for 30 s (during which the annealing temperature is lowered by 0.5 °C per cycle), and extension at 72 °C for 90 s (5 cycles); immediately followed by denaturing at 95 °C for 30 s, annealing/extension at 68 °C for 90 s (25 cycles)] was performed with the DICP.1/2/9/11/16/19 and DICP.7/8/17/22 forward primers (Table 1) in combination with the GeneRacer™ 3’ primer (Invitrogen). Subsequently, nested PCR products were performed with the PCR products from the touchdown PCR along with gene specific nested primers (Table 1) and the GeneRacer™ 3’ nested primer (Invitrogen) with the cycling parameters listed in Table 1. PCR products were cloned into pGEM®-T Easy (Promega) and sequenced. Attempts to amplify the 3’ ends of dcep1.26, dcep1.27, dcep1.28, and dcep1.29 transcripts were unsuccessful (data not shown).

Reverse transcription-PCR amplification for sequence analyses

In order to define DICP sequence variation between individual zebrafish, RT-PCR was performed from zebrafish kidney cDNA using the high-fidelity proofreading KAPA HiFi DNA Polymerase (Kapa Biosystems) and the DICP.1,7/8/17/22, DICP.2,1, and DICP.3.1 primer pairs. Primer sequences and cycling parameters are listed in Table 1 and Fig. 1. Amplicons were cloned into the pGEM®-T Easy plasmid and sequenced.

Haplotypes analyses

To investigate a predicted alternative DICP haplotype, PCR was performed using genomic DNA from adult zebrafish, the DICP.1 and DICP.2.2 primer pairs, Titanium Taq DNA polymerase, and the cycling parameters described in Table 1. Genomic DNA was obtained from fins of the adult zebrafish described above using a modified HotSHOT protocol (Mecker et al. 2007). The linkage of MHC class I Z lineage genes with this DICP haplotype was confirmed by genomic PCR using the MHC class I primers and cycling conditions described previously (Dirschier and Yoder 2014). Genomic DNA from zebrafish with defined MHC class I Z gene haplotypes were kindly provided by Hayley Dirschier (Dirschier and Yoder 2014). Amplicons were cloned into the pGEM®-T-Easy plasmid and sequenced.

Sequence analyses

The sequences obtained from the DICP transcripts were translated in silico and predicted protein domains identified by SMART software (Letunic et al. 2012). The nucleotide and amino acid (aa) sequences encoded by the DICP transcripts were used as queries for BLAST searches of the zebrafish reference genome (Howe et al. 2013), the nucleotide collection, the high throughput genomic sequences (HTGS), and the non-redundant protein sequences from the NCBI. Sequence alignments were generated using ClustalW2 (Larkin et al. 2007). Phylogenetic trees were constructed with the Neighbor-Joining method (Saitou and Nei 1987) and 1000 bootstrap replicates using MEGA5 (Tamura et al. 2011).
Fig. 1. Overview of the oligonucleotide primer design employed for amplifying DICP sequences. Primer pairs are listed on the left. Genes targeted by each primer pair and the overall genomic organization of these genes are listed to the right of each primer pair. Families of DICPs are defined by a number that denotes the DICP cluster (e.g., DICP1 cluster on chromosome 3) and gene names include a second number that denotes the order in which genes were identified (e.g., dicp1.1). Gray rectangles represent exons and black arrowheads approximate the relative location of each primer. Protein domains associated with each exon are indicated above the genomic organization (L, peptide leader sequence; D1, Ig domain; D2, Ig domain; LC, low complexity regions; TM, transmembrane domain; Cyt, cytoplasmic tail). Primer sequences are listed in Table 1.

Data access

All new DICP sequences reported here are provided in Online Resource 1 and have been deposited in the GenBank database under accession numbers KT858285–KT858478.

Results

DICP transcript detection and nomenclature

Twenty-seven DICP genes have been identified from the zebrafish reference genome (version Zv8) as well as from individual genomic (BAC) clones and can be placed into three groups, (DICP1, DICP2, and DICP3) based on sequence similarity and chromosomal location (chromosomes 3, 14, and 16, respectively) (Haire et al. 2012). In order to define the normal expression of various DICP genes, seven primer pairs (DICP1.1/29/11/16/19, DICP1.7/8/17/22, DICP1.22, DICP2.1, DICP3.2/3, and DICP3.6) were employed to amplify a range of different DICP transcripts (Fig. 1 and Table 1). For example, the primer pair DICP1.1/29/11/16/19 was designed to amplify transcripts encoded by the dicp1.1, dicp1.2, dicp1.8, dicp1.11, dicp1.16, and dicp1.19 genes. Although the DICP1.7/8/17/22 primer pair was designed to amplify transcripts of the previously described dicp1.7, dicp1.8, and dicp1.17 genes, it also amplified transcripts of dicp1.22, a previously uncharacterized DICP gene (see below). Subsequently, a DICP1.22 primer pair was designed to specifically amplify a 545 base pair (bp) amplicon of dicp1.22 (Fig. 1 and Table 1). RT-PCR was employed to evaluate DICP expression in various immune-related tissues from nine individual zebrafish from the TU, AB, and Ekw lines (Fig. 2), lymphoid and myeloid cells from the Ekw line (Fig. 3), and from various embryonic stages of development from the TU, AB, and Ekw lines (Fig. 4). Variable expression patterns were observed between zebrafish lines as well as individuals of the same line. Amplicons were cloned and sequenced to verify that they represent DICPs. In order to distinguish among transcript variants of the same gene, sequence identity numbers are included as superscripts after each gene symbol (e.g., transcript variant 5571 for dicp1.1 is shown as dicp1.1<sup>5571</sup>). Details of all new DICP sequences identified in this study are provided in Online Resource 1.

DICP expression in adult zebrafish tissues

Although the sequence of several DICP amplicons recovered from adult zebrafish tissues (Fig. 2) correspond with DICP sequences that were described previously (Haire et al. 2012), some primer pairs produced different size amplicons in a zebrafish line-dependent manner. For example, the DICP1.1/29/11/16/19 primer pair generated amplicons of ~700–800 bp from tissues from AB and Ekw zebrafish, but TU zebrafish gave rise to amplicons of ~950–1050 bp. Sequencing of the smaller amplicons from AB and Ekw fish revealed different dicp1.1 transcripts from AB fish (dicp1.1<sup>5571</sup> and dicp1.1<sup>5574</sup>) and Ekw fish (dicp1.1<sup>5564</sup>, dicp1.1<sup>5577</sup>, and dicp1.1<sup>5578</sup>). Sequencing of the larger amplicons from the TU individuals revealed new DICP sequences, dicp1.23, dicp1.24, dicp1.25, and dicp1.30, which are discussed below. Similar observations were made for amplicons generated by the DICP1.7/8/17/22 primer pair. Amplicons from AB and Ekw fish were ~700–800 bp and amplicons from TU fish were ~1000 bp. Sequencing of the smaller AB and Ekw amplicons revealed dicp1.7 and dicp1.8 transcripts (dicp1.2<sup>9064</sup>, dicp1.8<sup>9064</sup>, and dicp1.8<sup>9065</sup>). Sequencing of the larger amplicons revealed a new DICP, dicp1.22 (dicp1.22<sup>5579</sup>, dicp1.22<sup>5577</sup>, dicp1.22<sup>5580</sup>, dicp1.22<sup>5583</sup>, dicp1.22<sup>5586</sup>, and dicp1.22<sup>5587</sup>). A primer pair designed to amplify only dicp1.22 subsequently
### Table 1  Primer sequences and PCR cycling parameters for detection and recovery of DICP sequences

<table>
<thead>
<tr>
<th>Primer pair</th>
<th>Forward primer Reverse primer</th>
<th>Titanium Taq DNA polymerase</th>
<th>KAPA Hif DNA polymerase</th>
<th>Number of cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>DICP1.1/2/9/11/16/19</td>
<td>ATGCGCTGAGTGAGAGGTGCCTCTGGTTTGC</td>
<td>70 30 NA NA</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>DICP1.7/8/17/22</td>
<td>ATGCGCTGATGAGGTGCCTCTGGTTTGC</td>
<td>68 30 NA NA</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>DICP1.22</td>
<td>CTGCAAGACTGAGAAACGCTGAGG</td>
<td>70 30 NA NA</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>DICP2.1</td>
<td>ATGCGCTGAGTGAGAGGTGCCTCTGGTTTGC</td>
<td>68 30 NA NA</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>DICP3.1</td>
<td>AAGGTAGCTGAGAAGGAGGTGCCTCTGGTTTGC</td>
<td>68 30 NA NA</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>DICP3.2/3</td>
<td>AAGGTAGCTGAGAAGGAGGTGCCTCTGGTTTGC</td>
<td>68 30 NA NA</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>DICP3.6</td>
<td>AAGGTAGCTGAGAAGGAGGTGCCTCTGGTTTGC</td>
<td>TD 70-60 30 NA NA</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>DICP1.1 (nested only)</td>
<td>GATCGTACTGAGAAGGAGGTGCCTCTGGTTTGC</td>
<td>67 30 NA NA</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>DICP1.22 nested-1</td>
<td>GATCGTACTGAGAAGGAGGTGCCTCTGGTTTGC</td>
<td>65 60 NA NA</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>DICP1.23 nested-1</td>
<td>GACGTACTGAGAAGGAGGTGCCTCTGGTTTGC</td>
<td>65 60 NA NA</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>DICP1.24 nested-1</td>
<td>GACGTACTGAGAAGGAGGTGCCTCTGGTTTGC</td>
<td>65 60 NA NA</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>DICP1.25 nested-1</td>
<td>GACGTACTGAGAAGGAGGTGCCTCTGGTTTGC</td>
<td>65 60 NA NA</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>DICP1.26 nested-1</td>
<td>GACGTACTGAGAAGGAGGTGCCTCTGGTTTGC</td>
<td>65 60 NA NA</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>DICP1.30 nested-1</td>
<td>GACGTACTGAGAAGGAGGTGCCTCTGGTTTGC</td>
<td>65 60 NA NA</td>
<td>25</td>
<td></td>
</tr>
</tbody>
</table>

*NA not applicable  
*TD Touch-down PCR during which the annealing temperature is lowered from 70 to 60 °C by 0.5 °C per cycle for 20 cycles and then an additional 15 cycles are completed with a 60 °C annealing temperature.
Fig. 2 DICP gene expression in immune-related tissues. DICP expression was evaluated using primer pairs listed in Table 1 and tissues from nine individual adult zebrafish of the TU, AB, and EKW lines. RT-PCR amplicons shown were generated with Titanium Taq DNA polymerase and yellow rectangles indicate those products that were cloned and sequenced to confirm their identity. Orange rectangles indicate amplicons that subsequently were generated with a proof-reading DNA polymerase (KAPA HIF) for evaluation of sequence variation. The size and identity of recovered amplicons is listed on the right of the gel image with red text indicating nonfunctional transcripts. β-actin expression was used as a reference for cDNA quantity and quality.
generated amplicons from TU tissues, but it failed to amplify detectable amplicons from the AB and EKWs individuals in this study (Fig. 2). The DICP1.7/8/17/22 primer pair also generated multiple *dcep1.17* amplicons from the AB, EKWs, and TU individuals, which were not visualized but could be identified by cloning and sequencing (*dcep1.1p5*06, *dcep1.1p7*25, *dcep1.1p7*35, *dcep1.1p7*45, *dcep1.1p7*55, *dcep1.1p7*65, *dcep1.1p7*75, *dcep1.1p7*85, and *dcep1.1p7*95, as well as amplicons reflecting two new genes, *dcep1.27* and *dcep1.28* (discussed below)).

Comparable overall expression of the putative derivating DICP, *dcep2.1* (Haire et al. 2012), is observed in multiple tissues from adult TU, AB, and EKWs individuals. Although some minor variation in patterns of *dcep2.1* expression was observed between individual zebrafish, expression was consistently highest in kidney and spleen (Fig. 2). The DICP3.1 primer pair generated multiple D1-D2 *dcep3.1* amplicons from nearly all examined tissues from all zebrafish lines (Fig. 2). Some of the recovered sequences were more similar to the *dcep3.1b* allele previously identified in BAC CH13-346H11 (GenBank FP92011), whereas the remaining *dcep3.1* sequences were more similar to the *dcep3.1b* allele from genomic scaffold 1852 (GenBank NW_001837462.2) of the zebrafish Zv8 reference genome (Haire et al. 2012). Although the DICP3.2/3 primer pair generated amplicons from all tissues and fish that were examined, the only functional transcripts that were recovered were from *dcep3.3*. The transcript recovered from the TU fish (*dcep3.3b*) was most similar to *dcep3.1b* encoded in genomic scaffold 152 (Haire et al. 2012). In contrast, the *dcep3.3* sequences recovered from the AB and EKWs fish (*dcep3.3b*, *dcep3.3b*, and *dcep3.3b*30) were most similar to the *dcep3.1b* allele identified in BAC CH13-328B17 (GenBank FP90562). The DICP3.6 primer pair produced amplicons from only six of the nine individuals examined: one of three TU fish, two of three AB fish, and three of three EKWs fish (Fig. 2). Sequencing of the amplicons revealed only one functional transcript, which was from an AB fish (*dcep3.6b*). The DICP3.6 primer pair also amplified a *dcep3.3* sequence from the spleen of a TU fish (*dcep3.3b*). Sequencing of additional amplicons from multiple DICP3 primer pairs revealed additional new, but non-functional, DICP gene sequences (*dcep3.7* and *dcep3.8*), which are discussed below.

**DICP expression in zebrafish lymphoid and myeloid cells**

DICP transcripts were recovered in varying relative abundance from lymphoid and myeloid cells isolated from adult EKWs zebrafish (Fig. 3). The DICP1.1/2/9/11/16/19 primer pair amplified functional transcripts of DICPs that were identified previously (Haire et al. 2012) such as *dcep1.1* transcripts that were recovered from both myeloid and lymphoid cells (*dcep1.1p5*45, *dcep1.1p5*55, and *dcep1.1p5*60) but were expressed at higher levels in the lymphocyte population (Fig. 3). This same primer pair revealed a *dcep1.16* functional transcript from myeloid cells (*dcep1.16p4*25). Using the DICP1.7/8/17/22 primer pair, multiple amplicons of different sizes were obtained from lymphoid and myeloid mRNA (Fig. 3), including transcripts of *dcep1.8* and *dcep1.17* from lymphoid cells (*dcep1.8*46, *dcep1.8*50, and *dcep1.17*305, and *dcep1.17*305) and *dcep1.17* transcripts from myeloid cells (*dcep1.17p6*65, *dcep1.17p5*65, and *dcep1.17p5*66). Sequencing of the leukocyte amplicons generated by the DICP1.7/8/17/22 primer pair revealed one new DICP gene sequence (*dcep1.29*), which is discussed below. The DICP1.22 primer pair produced no detectable amplicons from leukocytes of EKWs individuals (data not...
Fig. 4 DICP gene expression during zebrafish development. RT-PCR was employed to detect DICP transcripts at different developmental stages from TU, AB, and EKW zebrafish lines. Ten embryos were pooled for each cDNA template. RT-PCR was performed with Titanium Taq DNA polymerase. Yellow rectangles indicate amplicons that were cloned and sequenced to confirm their identity. The size and identity of recovered amplicons is listed on the right of the gel images; red text indicates non-functional transcripts. β-actin expression was used as a reference for cDNA quantity and quality.

shown), which may reflect the absence of this gene from the genomes of the fish from which leukocytes were isolated (discussed below).

The expression of the putative activating DICP, dicp2.1, was detected in both leukocyte lineages. The DICP2.1 primer pair generated one predominant band of ~750 bp from lymphoid cell cDNA and several bands that range from 400 to 1500 bp from the myeloid cell cDNA (Fig. 3). These amplicons include functional dicp2.1 transcripts from lymphoid (dicp2.1539 and dicp2.1566) and myeloid cells (dicp2.1530 and dicp2.1590), however, both lymphocyte transcripts possess a deletion within the D1 or D2 domain.
Although additional amplicons were not identified from the myeloid cell cDNA, the ~400 bp amplicon might correspond to a functional dicp3.1 transcript encoding a single D2 lg domain that was identified from AB zebrafish kidneys (dicp3.1.1g09). Similarly, the ~1200 bp amplicon might correspond to non-functional dicp3.1 transcripts (described below).

The DICP3.1 primer pair generated amplicons from both myeloid (dicp3.1.1436 and dicp3.1.1446) and lymphoid (dicp3.1.1465) cells (Fig. 5) which also had been recovered from zebrafish kidneys, as well as amplicons representing new dicp3.1 transcripts from myeloid cells (dicp3.1.1125/1272 and dicp3.1.1272) and lymphoid cells (dicp3.1.1275, dicp3.1.1275, and dicp3.1.1275) Transcripts from myeloid cells were more similar to the dicp3.1.1437 allele, except for dicp3.1.1441, which was more similar to the dicp3.1.1451 allele (Haire et al. 2012).

All lymphoid dicp3.1 transcripts were more similar to the dicp3.1.1451 allele. The DICP3.2/3 primer pair recovered functional dicp3.1 transcripts (dicp3.1.1328, dicp3.1.1334, dicp3.1.1334, dicp3.1.1334, dicp3.1.1334, dicp3.1.1334, dicp3.1.1334, and dicp3.1.1334) from myeloid cells and one functional dicp3.3 transcript from lymphoid cells that included a deletion within the D1 domain (dicp3.3.9257). The DICP3.6 primer pair produced only non-functional dicp3.6 transcripts from lymphocytes. Sequencing of the leukocyte amplicons generated by the DICP3 primer pairs revealed a functional dicp3.7 transcript and additional non-functional dicp3.7 and dicp3.8 transcripts (see below).

**Variable DICP expression during zebrafish embryo development**

Diverse expression patterns of DICP1 genes are revealed at different stages of embryonic development (Fig. 4). The DICP1.1/2/9/11/16/19 primer pair generated amplicons exhibiting the same relative pattern of expression from TU, AB, and EKW embryos; however, differences in amplicon length are evident between different developmental stages and genetic backgrounds. The earliest developmental stage at which transcripts can be detected is 12 hpf, with amplicon lengths ranging from 407 to 1020 bp. By 36 hpf, 48 hpf, some of these amplicons were not detectable but ~1000 bp amplicons were detected in the three zebrafish lines. Sequence analysis of these amplicons detected previously described DICP transcripts (Haire et al. 2012), including dicp1.9 and dicp1.19 (dicp1.9.1425, dicp1.9.1425, dicp1.9.1425, and dicp1.9.1425), as well as transcripts of new DICP genes (dicp1.24, dicp1.25, and dicp1.26; see below). Amplicons generated with the DICP1.7/8/17/22 primer pair displayed similar relative lengths and patterns between TU and EKW embryos (Fig. 4); functional dicp1.8 transcripts can be detected at 6 dpf (dicp1.8.503, dicp1.8.503, and dicp1.8.503). Functional transcripts of dicp1.17 also were recovered from 6 dpf TU (dicp1.17.15440 and dicp1.17.15440) and EKW embryos (dicp1.17.15440). The most abundant functional transcripts from AB embryos were shown to be dicp1.17 at 24 hpf (dicp1.17.15440 and dicp1.17.15440), including one dicp1.17 transcript that lacks a transmembrane domain (dicp1.17.15440).

The DICP1.22 primer pair generated two amplicons from embryos of all three zebrafish lines (Fig. 4). The shorter amplicon corresponds to a functional transcript of dicp1.22 (dicp1.22.1557, dicp1.22.1557, and dicp1.22.1557); the larger amplicon, which likely corresponds to a non-functional dicp1.22 transcript, was not sequenced (see below). The dicp1.22 transcripts in embryos of the three zebrafish lines contrast with the expression observed in the adult tissues where transcripts were detected only in the TU line, negating the possibility that dicp1.22 or a specific allele of this gene is present only in the TU line. The only DICP with definitive maternal transcripts in the one-cell embryo stage (0–1 hpf) was dicp1.22 from the AB line.

Expression of the putative activating receptor, dicp2.1, appears to be absent or at very low levels during embryonic development with amplicons first being identifiable from 6 dpf TU and AB embryos (Fig. 4). Sequencing of these amplicons revealed several functional dicp2.1 transcripts (dicp2.1.1533, dicp2.1.1533, dicp2.1.1533, dicp2.1.1533, dicp2.1.1533, dicp2.1.1533, dicp2.1.1533, and dicp2.1.1533). Although dicp2.1 amplicons were not visible in the gel from EKW embryos, multiple functional dicp2.1 amplicons were recovered from 6 dpf EKW embryos (dicp2.1.1533, dicp2.1.1533, dicp2.1.1533, dicp2.1.1533, dicp2.1.1533, dicp2.1.1533, and dicp2.1.1533), including two transcripts with a deletion in the D2 domain (dicp2.1.1533 and dicp2.1.1533).

Transcripts of DICP3 genes could be detected by 6 hpf with expression maintained throughout embryonic development (with the exception of dicp3.6.1, which was not detected in AB and EKW embryos). The DICP1.1 primer pair detected multiple transcripts from the three genetic backgrounds (dicp1.1.1508, dicp1.1.1508, and dicp1.1.1508), larger transcripts with additional coding sequence between the D1 and D2 domains (dicp3.1.1508 and dicp3.1.1508) also were identified. Amplicons displaying highly similar lengths and patterns were recovered from the AB and EKW embryos with the DICP3.2/3 primer pair, including functional dicp3.3 transcripts (dicp3.3.1508 and dicp3.3.1508). A dicp3.3 transcript also was recovered from TU embryos that lacked the D1 domain (dicp3.3.1508). The DICP6.6 primer pair revealed only non-functional dicp3.6 transcripts from TU embryos.

**New DICP sequences**

DICP transcripts possessing D1 domains that were <90 % identical to any previously described DICP were designated as a new gene, such as the dicp1.22 transcripts that were recovered from kidney cDNA of TU zebrafish (Fig. 2) and described above. In addition, the DICP1.1/2/9/11/16/19

---

This text is an excerpt from a scientific publication discussing the expression and transcription patterns of specific genes in zebrafish embryos. The focus is on the identification and characterization of DICP genes, which are involved in immune response and development. The text describes the use of PCR and sequencing to identify and analyze these genes at different stages of embryonic development.
primer pair generated amplicons from the TU zebrafish that were larger than the amplicons detected in the AB and EKWFish (Fig. 2). Although the shorter AB and EKWFish amplicons represent dicp1.1 transcripts (see above), sequence of the larger amplicons from a TU zebrafish reveal two new DICPs, dicp1.23 and dicp1.25 (dicp1.23^657, dicp1.23^568, and dicp1.25^551). Four additional DICP amplicon sequences were identified, although they cannot be observed in Fig. 2. The DICP.1.2/3/5/11/16/19 primer pair yielded dicp1.24 and dicp1.30 amplicons (dicp1.24^512 and dicp1.30^538) from EKWFish liver and the DICP.1.7/8/17/22 primer pair yielded dicp1.27 and dicp1.28 amplicons (dicp1.27^590 and dicp1.28^599) from EKWFish kidney.

New DICPs also were recovered from zebrafish leukocytes (Fig. 3) and embryos (Fig. 4). The DICP.1.7/8/17/22 primer pair from lymphocyte cDNA yielded dicp1.29, which encodes a single D1 Ig domain transcript (dicp1.29^501). The DICP.1.1 and DICP.2/3 primer pairs amplified non-functional transcripts of two new genes, dicp3.7 and dicp3.8 from kidney cDNA (Online Resource 1); a functional dicp3.7 transcript was recovered from myeloid cell cDNA that encodes two Ig (D1-D2) domains (dicp3.7^525). Functional transcripts of dicp1.24 and dicp1.25 as well as an additional gene dicp1.26 (dicp1.24^616, dicp1.24^567, dicp1.24^558, dicp1.24^550, dicp1.24^540, dicp1.24^537, dicp1.24^547, dicp1.25^547, dicp1.25^538, and dicp1.26^545) were recovered from embryo cDNA with the DICP.1.2/9/11/16/19 primer pair. The dicp1.26 amplicon encodes two Ig domains (D1-D2). Phylogenetic analyses of all predicted DICP Ig domains are consistent with the classification of these sequences as new DICP genes (Fig. 5).

DICP sequence analyses

In order to investigate the inter-individual sequence variation of select DICP genes, a high-fidelity DNA polymerase was used for RT-PCR with the same individual zebrafish evaluated in Fig. 2. The cloning and sequencing of multiple DICP 1.7/8/17/22, DICP 1.1, and DICP 3.1 amplicons from each individual revealed differences in sequences and splicing (Figs. 6 and 7; Online Resource 2—Table S2; Online Resource 3—Sequence Variation and Figs. S1–S6). In silico translation of the amplicons revealed that all of the amplified transcripts encode one or two Ig domains. A phylogenetic comparison of these DICP Ig domains with all other DICP Ig domains demonstrated that the majority of these Ig domains grouped with the D1 and D2 Ig domains of Dicp1.7, Dicp1.8, Dicp1.17, Dicp1.22, Dicp3.1, and Dicp3.3, as expected (Online Resource 3—Fig. S7).

Additional sequencing of several DICP transcripts with a non-high-fidelity DNA polymerase also revealed evidence of alternative splicing and allelic variation (Figs. 6 and 7 and Online Resource 3). The exon-intron organization of these DICPs was predicted based on the DICP genes described in the Zv9 zebrafish reference genome. Numerous transcripts likely reflect alternative mRNA splicing events, including exon skipping, intron retention, and alternative 3’ and 5’ splice sites, although these sequences may also reflect haplotypic variants that have gained or lost sequences (as compared to the reference genome). Insertions or deletions inside of an exon also were identified that might indicate either an alternative splicing inside of the exon or these insertions/deletions already were part of the gene. A possible instance of exon shuffling or recombination was identified in the dicp1.8 transcript, which possesses two exons from the dicp1.17 gene (Fig. 6 and Online Resource 3—DICP exon-intron architecture). Transcripts from multiple DICP genes contain a predicted premature termination codon (PTC). These transcripts likely correspond either to pre-messenger RNAs (pre-mRNA) that possess a PTC by intron retention or mRNAs reflecting alternative splicing that results in a frameshift and a PTC. A detailed description of observed polymorphisms, alternative splicing, and PTCs are presented in Online Resource 3.

Structural features of new DICPs

By overlapping cDNA sequences obtained by RT-PCR and 3’ RACE, full-length sequences were predicted for Dicp1.22, Dicp1.23, Dicp1.24, Dicp1.25, and Dicp1.30 (Online Resource 3—Figs. S8–S9). Dicp1.22 possesses two Ig domains, a transmembrane domain, and a cytoplasmic tail that lacks any known signaling motifs. Dicp1.23 is predicted to be a secreted protein. Although the dicp1.23 transcripts possess an exon that could encode a transmembrane domain, a frameshift 5’ of this exon results in the use of an alternative reading frame and no transmembrane domain (Online Resource 3—DICP exon-intron architecture). Dicp1.24 possesses a single D1 domain and three cytoplasmic ITIMs and an ITIM-like sequence (ritm). Dicp1.25 possesses D1 and D2 etodoms as well as three cytoplasmic ITIMs. Dicp1.26 possesses two Ig domains and likely a transmembrane domain; however, 3’ sequences were not recovered. Dicp1.27, Dicp1.28, Dicp1.29, and Dicp1.30 each possess a D1 domain and a transmembrane domain; however, 3’ sequences were recovered only for dicp1.30 which encodes two ITIMs (Online Resource 3—Figs. S8–S9).

Linkage of DICP and MHC class I Z haplotypes on zebrafish chromosome 3

In order to identify genomic sequences that encode the newly identified DICPs, the Ig domains of the 11 new DICPs were used as queries for BLAST searches of the zebrafish reference genome (GRCz10; Howe et al. 2013), the nucleotide collection (nt), the unfinished high throughput genome sequence (HTGS), and the non-redundant protein sequence (nr) databases. Only dicp1.22 and dicp1.23 produced significant
matches, Specifically, genomic sequences matching dicpl.22 and dicpl.23 transcripts were identified in unplaced genomic scaffold NA310 (GenBank NW_003336703.1) (Fig. 8a). Scaffold NA310 also encodes the mhelzka (GenBank NM_001302245), cedc134 (coiled-coil domain-containing protein 134-like; GenBank XM_003201673), and gimap6 (GTPase IMAP family member 4-like; GenBank XM_001920324) genes which are predicted to represent an alternate haplotype for the mhelzja (GenBank NM_001109718), cedc134 (coiled-coil domain-containing protein 134; GenBank XM_003198004), and gimap6 (GTPase IMAP family member 8; GenBank XM_001919280) genes on chromosome 3 (Online Resource 3—Fig. S10) (Dinscherl and Yoder 2014). Based on the linkage of dicpl.22 and dicpl.23 to mhelzka, cedc134, and gimap6 and the linkage of dicpl.1—dicpl.21 to mhelzja, cedc134, and gimap6 (Fig. 8a and Online Resource 3—Fig. S10), these sequences were predicted to represent alternate haplotypes with differing gene content for both the DICP gene cluster and the MHC class I Z gene cluster on chromosome 3.

Genomic PCR analyses were employed with individual zebrafish predicted to be homozygous for the different haplotypes (dicpl.1 and dicpl.22) depicted in Fig. 8a and shown to express either dicpl.1 or dicpl.22 transcripts (Fig. 2). Genomic amplicons for dicpl.1 were identified in zebrafish
Fig. 6 Exon-intron architecture of previously identified DICPs. The exon-intron organization of transcript variants encoding previously described DICPs were compared with the DICP genes present in the reference genome. Sequence identifier numbers or GenBank accession numbers are listed to the right of each transcript. Red text indicates predicted non-functional transcripts. Details are provided in Online Resource 3—DICP exon-intron architecture.

that express dicp1.1 and genomic amplicons for dicp1.22 were identified in zebrafish that express dicp1.22 (Figs. 2 and 8b). Individuals encoding dicp1.1 (and not dicp1.22) also encode mhc1zja (and not mhc1zka) and individuals encoding...
diclp1.22 (and not diclp1.1) also encode mhcZ2ka (and not mhcZ1za) (Fig. 8a). These observations support the hypothesis that diclp1.1 and diclp1.22 represent different DICP haplotypes that are tightly linked to different MHC class I Z haplotypes and that these individual zebrafish are homozygous for a single haplotype. Genomic PCR analyses with DNA from zebrafish that are heterozygous for the two MHC class I Z gene haplotypes depicted in Fig. 8a (Dirschler and Yoder 2014) indicate that both diclp1.1 and diclp1.22 haplotypes are present (Fig. 8c).

**Linkage of DICP and MHC class I genes in cyprinid fishes**

Data mining available genomic sequence databases in 2012 identified definitive DICP sequences in zebrafish and common carp (Haire et al. 2012) which are both members of the Cyprinidae family. Although DICP-like sequences were identified in Atlantic salmon (Salmo salar), Japanese pufferfish (Takifugu rubripes), green spotted pufferfish (Tetraodon nigroviridis), and Nile tilapia (Oreochromis niloticus) which belong to the Salmonidae, Tetraodontidae, and Cichlidae families (Haire et al. 2012), it is unknown if they share an evolutionary origin.

In order to investigate if DICP genes could be identified in other species of fish, DICP Ig domains from zebrafish and common carp were used as queries for tBLASTn searches of the teleost genome sequence databases currently available in Ensembl (v82). Definitive DICP sequences were not identifiable in any current genome assemblies of Amazon molly (Poecilia formosa), cave fish (Astyanax mexicanus), cod (Gadus morhua), Japanese pufferfish, medaka (Oryzias latipes), platyfish (Xiphophorus maculatus), stickleback (Gasterosteus aculeatus), green spotted pufferfish, or Nile tilapia. As none of these fish are in the Cyprinidae family, it is possible that definitive DICPs may be restricted to cyprinid species. In order to determine if DICP and MHC genes are linked in other cyprinids, the recently reported genomes of the grass carp (Wang et al. 2015) and common carp (Xu et al. 2014) were searched and definitive evidence for their linkage was established. Conserved synteny was observed between zebrafish chromosome 3, grass carp scaffold C10000243, and common carp linkage Group 38 (Online Resource 3—Fig. S11) demonstrating that DICPs are present and linked to MHC class I Z sequences in three different subfamilies of cyprinids. Assignment of carp sequences to the DICP family and the MHC class I Z lineage is supported by phylogenetic analyses (Online Resource 3—Fig. S12). As annotation has not been performed on the carp gene models, the predicted protein sequences referred to in Online Resource 3—Figs. S11 and S12 are provided in Online Resource 3—Fig. S13 and the BLAST results from using each carp sequence as a query to search the NCBI database of non-redundant (nr) zebrafish proteins (queried November 2015) are provided in Online Resource 3—Table S4.

**Discussion**

The zebrafish DICPs exhibit similarities in structure and ligand recognition to members of the mammalian CD300,
TREM, and FcR-like (FCRL) families (Haire et al. 2012). Certain CD300 and TREM receptors bind specific subsets of phospholipids which may reflect roles in differentiating pathogens, mediating phagocytosis of apoptotic cells and/or recognizing activated lymphocytes (Borrego 2013, Cannon et al. 2012, Pelham and Agrawal 2014). Although recombinant forms of certain DICPs also bind specific subsets of phospholipids, the overall role of DICPs in immune function remains unknown. In order to better characterize the DICPs, we provide a detailed examination of the expression patterns and sequence diversity of the DICP gene family between individual zebrafish from different genetic backgrounds. The comparison of nearly 200 new DICP transcript sequences identified 11 new DICP gene sequences, revealed extensive
polymorphic and haploidy variation between DICPs of individual zebrafish, and documented transcripts that likely reflect alternative mRNA splicing that, in many cases, resulted in presumably non-functional transcripts.

Each primer pair employed to detect DICP expression yielded distinctly different patterns from adult tissues and leukocytes. Although DICPs are differentially expressed in immune tissues, a nearly universal feature of the DICP genes is their expression in the hematopoietic kidney (Fig. 2). The expression of DICPs in several immune-related tissues and their heterogeneous expression within these tissues is reminiscent of the expression patterns of CD300 in mammals (Clark et al. 2009). DICPs are expressed in both myeloid and lymphoid cell lineages with certain DICPs being expressed at higher levels in one lineage (Fig. 3). By way of comparison, expression of mammalian CD300 genes is most abundant in monocytes, but certain CD300 transcripts also are expressed at high levels in T or NK cells (Clark et al. 2009).

Expression of DICPs was evaluated at different developmental stages and transcripts of multiple DICPs (e.g., dicp1.22, dicp3.1, dicp3.3) were detected as early as 6 hpf, whereas transcripts of other DICPs (e.g., dicp2.1) were not detected until well after embryogenesis (Fig. 4). The embryonic expression of certain DICPs raises the possibilities that they may play important and early roles in innate immunity within the developing embryo or that they may play specific roles in embryonic development.

In-depth sequence analyses of the DICP amplicons generated from different tissues, leukocytes, and embryos from different zebrafish lines reveal a large number of transcript variants that reflect the loss of exons or gain of predicted introns which likely represent alternative splice variants, allelic variants, or undefined DICP genes. A number of transcript variants differ in the length and sequence of the stalk domains, which are located either between the Ig domains or between an Ig domain and the transmembrane domain. In contrast, the Ig domains of transcript variants display high sequence conservation, perhaps reflecting a role for different stalk domains to act as flexible spacers for ligand binding. Stalk domains in some immune receptors have been implicated in ligand binding and intracellular signaling suggesting that the observed differences may directly influence function (Berry et al. 2013; Hartmann et al. 2012; Moody et al. 2001; Xu et al. 2006). Alternative splicing, which can lead to the loss of an Ig domain (e.g., dicp2.1 and dicp3.1), likely contributes to the complexity of the immune receptor repertoire (Maeyer and Imami 2011).

Multiple DICP transcript variants introduce PTCs (Online Resource 1) that may (1) correspond to pre-mRNAs which have not been fully processed, (2) encode truncated, secreted DICPs, or (3) be targeted for elimination through the nonsense-mediated mRNA decay (NMD) pathway (Kervestin and Jacobson 2012). In addition to RNA surveillance, the NMD pathway employs mRNAs containing PTCs to control other cellular functions (Hamid and Makeyev 2014), including the regulation of gene expression during hematopoiesis (Frishmeyer-Guerrerio et al. 2011; Penentel et al. 2014; Wong et al. 2013) as well as physiological responses to bacterial infection (Gloggnitzer et al. 2014, Kalyna et al. 2012). Certain DICPs exhibit a higher number of transcripts containing PTCs than functional transcripts, such as dicp1.23, dicp3.2, dicp3.6, and dicp3.7, in contrast to other DICPs, such as dicp1.8, dicp1.17, dicp2.1, and dicp3.1, which present higher numbers of apparently functional transcripts (Online Resource 3—Fig. S14). Most DICP transcripts encoding PTCs were identified from lymphoid cells although a few were recovered from myeloid cells. The source of these differences is unknown and it is recognized that the efficiency of each PCR may influence these observations.

Sequence analyses also identified several DICP transcripts encoding D1 domains that were less than 90 % identical to previously described DICP D1 domains (Haire et al. 2012). Included in this group are 11 new DICP genes, dicp1.22, dicp1.30, dicp3.2, dicp3.7, and dicp3.8. These new genes encode inhibitory receptors (dicp1.24, dicp1.25, and dicp3.30), secreted proteins (dicp1.23 and a dicp1.17 splicing isoform), and proteins with ambiguous function (dicp1.22). Only two of these genes, dicp1.22 and dicp1.23, could be mapped to the reference genome (GRCh38), where they are encoded in scaffold NA310, which also encodes one MHC class I gene cluster haplotype (Dierschel and Yoder 2014). Scaffold NA310 likely represents an alternative haplotype for both MHC class I genes and DICP genes on chromosome 3. The identification of transcripts for dicp1.24–dicp3.30 without representation in the genomic databases suggests that they may be present on the haplotype which encodes dicp1.22 and dicp1.23 or that additional DICP haplotypes remain to be identified for this locus on chromosome 3.

Definitive DICPs have been identified in three subfamilies of the Cyprinidae family: Rasborinae (zebrafish), Leuciscinae (grass carp), and Cyprininae (common carp). The relationship of DICP sequences to DICP-like sequences in non-cyprinid teleosts remains to be elucidated. Although it is possible that DICPs may be restricted to cyprinids, DICPs more likely share an ancient origin with DICP-like sequences from other teleost lineages. The identification of DICPs and their linkage to MHC class I sequences in multiple cyprinid lineages indicates that the DICPs and their linkage to MHC class I has been conserved for at least 70 million years which is the estimated divergence time of the Cyprinidae family (Broughton et al. 2013). The linkage of Ig domain-containing innate immune receptor genes (such as those encoding DICPs) to MHC genes supports the model that similar genes were linked in the ancient “proto-MHC” (or “Ur-MHC”) (Kasahara 1999; Abi-Rached et al. 1999). This model places the primordial MHC
genes in the same chromosomal region as Ig domain-containing receptors, which would be precursors to modern mammalian receptors encoded at the leukocyte receptor complex (LRC) (Flajnik and Kasahara 2010). The proposal that the precursors of the mammalian LRC and MHC were once genetically linked is supported by their demonstrated functional relationship—many receptors encoded at the LRC bind MHC class I proteins as markers of “self” (Vivier et al. 2008). The linkage of DICP and MHC genes in cyprinids may have originated from the proto-MHC; however, it is not known if they are linked functionally. In support of an ancient linkage, DICP-like sequences from medaka (reported as “NTR-Wac” sequences) are linked to MHC class II sequences (Ohashi et al. 2010). The significance of this linkage remains to be elucidated and further studies are required to investigate the DICP-like sequences outside of cyprinid fishes.

It is likely that additional haplotypes of the DICP3 cluster will be identified on chromosome 16 as (1) genes encoding dcp3.7 and dcp3.8 are absent from the current reference genome; (2) the dcp3.6 gene is absent from earlier versions of the reference genome (Zv8) but has been identified in two BAC clones (Haire et al. 2012), and (3) dcp3.6 amplicons only have been detected in certain individual zebrafish (Fig. 2) and only in embryos of the TU line (Fig. 4). The possibility of multiple DICP haplotypes is supported by the high frequency of CNVs observed at all three DICP gene clusters (Rodriguez-Nunez et al. 2014). CNVs provide an important evolutionary strategy for chromosomal rearrangement as well as the creation of novel loci and occurrence in immune-related gene clusters, such as the mammalian MHC and KIR loci (Dirscharl and Yoder 2014; Traheme et al. 2010). As different haplotypes of immune-related genes have been correlated with resistance or susceptibility to infections and autoimmune disease (Jackson et al. 2007; Olson and Holmdahl 2012; Pelak et al. 2011), it will be of interest to determine if different DICP haplotypes influence individual fitness of zebrafish.

Acknowledgment We thank Hayley Dirscharl (NC State University) for helpful discussions and zebrafish genomic DNA samples, John Whitesides and Pati McDermott (Duke Human Vaccine Institute, Duke University School of Medicine) for assistance with cell sorting, Dana Skupka (All Children’s Hospital) for assistance with DNA sequencing, and Barbara Pryor for editorial assistance. The authors are supported by the National Institutes of Health (R01 AI057559 to G.W.L. and J.A.Y. and R01 AE23377 to G.W.L.) and the National Evolutionary Synthesis Center (NESCent), NSF #EF-0965606 (D.J.W.).

References


CHAPTER 3: IDENTIFICATION OF NOVEL IMMUNOGLOBULIN-LIKE TRANSCRIPTS AND GENE CONTENT-HAPLOTYPES IN ZEBRAFISH

Dustin J. Wcisel\textsuperscript{1,2}, Alex Dornburg\textsuperscript{3}, Sean McConnell\textsuperscript{4,5}, Jill L. O. de Jong\textsuperscript{4,5}, Gary W. Litman\textsuperscript{6}, Jeffrey A. Yoder\textsuperscript{1,2,7,8}

1. Genomic Sciences Graduate Program, North Carolina State University, Raleigh, NC 27607
2. Department of Molecular Biomedical Sciences, College of Veterinary Medicine, North Carolina State University, Raleigh, NC 27607
3. North Carolina Museum of Natural Sciences
4. Section of Hematology-Oncology and Stem Cell Transplant, Department of Pediatrics, The University of Chicago, Chicago, IL 60637
5. Center for Research Informatics, The University of Chicago, Chicago, IL 60637
6. Department of Pediatrics, University of South Florida Morsani College of Medicine, St. Petersburg, Florida, USA
7. Center for Human Health and the Environment, North Carolina State University, Raleigh, NC 27607, USA
8. Comparative Medicine Institute, North Carolina State University, Raleigh, NC 27607
**Introduction**

Different species often encode significantly different repertoires of immune receptors which are essential for differentiating between self and non-self. The most significant differences are observed at key points of divergence in speciation and it has been suggested that the evolution of the immune system can influence the rate of speciation (Malmstrøm et al. 2016). For example, the split between jawless and jawed vertebrates is characterized by completely different rearranging immune genes, the hallmark of adaptive immunity: jawless vertebrates use variable lymphocyte receptor genes whereas jawed vertebrates employ immunoglobulin (Ig) genes (Flajnik and Du Pasquier 2004). Differences in the repertoire of non-rearranging Ig domain-containing innate immune receptor (IIIR) genes are also observed between vertebrate lineages. For example, human and teleost both encode a polymeric immunoglobulin receptor (pIgR) which facilitates the transport of antibodies across epithelial boundaries (Kaetzel et al. 1991; Parra et al. 2016); however, in the human genome, the PIGR gene is flanked by genes that encode two different IIIRs that also bind antibodies (FCAMR and FCMR) whereas the zebrafish (*Danio rerio*) pigr locus has experienced a tandem expansion of 24 additional PIGR-like (PIGRL) genes of unknown function (Kortum et al. 2014). Mammals encode another IIIR gene family, the CD300s (CMRF-35 in mouse), that bind phospholipids (Borrego 2013). Although cartilaginous fish (e.g. sharks and skates) do not encode definitive orthologs of CD300s, they do encode modular domain immunoglobulin-like receptors (MDIRs) that share sequence homology with CD300s (Cannon et al. 2006).

Highly similar Ig superfamily (IgSF) genes are classified into multigene families based on sequence homology and genomic location. For example, seven CD300 genes are encoded in tandem on human chromosome 17 (Borrego 2013). These genes share many features with other IIIRs, including the same general protein structure: an ectodomain consisting of one or more Ig domains utilized for ligand specificity, a transmembrane domain and a cytoplasmic tail. The cytoplasmic tail may include conserved motifs such as immunoreceptor tyrosine-based activation/inhibition motif
(ITAM/ITIM) or may associate with ITAM-containing adaptor proteins such as DAP10 or DAP12 through a charged residue in the transmembrane domain. When the ectodomain binds ligand, ITAM/ITIMs recruit cytoplasmic phosphatases or kinases to activate or inhibit cellular response (Billadeau and Leibson 2002).

Orthologous IIIR gene families often exhibit high sequence divergence. For example, novel immunoglobulin-like transcripts (NILTs), first identified in carp (Cyprinus carpio), are only ~30-60% identical to NILTs identified in Atlantic salmon (Salmo salar) (Stet et al. 2005; Ostergaard et al. 2010). This level of sequence difference is comparable to levels exhibited between different gene families within a species, confounding the accurate identification of orthologous gene families (Rodríguez-Nunez et al. 2014). Similarly, MDIRs are ~30-60% identical to teleost NILTs; however, Ig domains in both gene families possess conserved cysteines, spaced 6-7 residues apart (CX$_{6-7}$C), that are predicted to influence the Ig domain structure through a novel disulfide bridge (Cannon et al. 2006; Stet et al. 2005). Unique to some NILT Ig domains is a CX$_3$C motif rather than CX$_{6-7}$C (Kock and Fischer 2008). Despite the sequence differences, conservation of this CX$_{3,6-7}$C motif suggests a common evolutionary origin of PIGR, PIGRL, MDIR, CD300, and NILT genes. Defining the evolutionary origins of these receptors has the potential to guide functional studies of these gene families.

In this report, we take advantage of the zebrafish model and the recently sequenced genome of Atlantic salmon (Salmo salar) to expand and annotate the NILT gene family in zebrafish. We identify and describe 25 NILT transcripts from zebrafish that guide the annotation of the NILT locus on chromosome 1. We demonstrate that the zebrafish genome displays intraspecific variation in the number and combination of NILTs (e.g., gene-content haplotype). A single activating NILT was recovered, as well as five inhibitory, and 18 ambiguous or secreted NILTs. Three NILTs contained a novel putative signaling motif with the sequence TVYAT.
Methods

Zebrafish

All experiments involving live zebrafish were performed in accordance with relevant institutional and national guidelines and regulations and were approved by the North Carolina State University Institutional Animal Care and Use Committee. Wild-type zebrafish lines used included NHGRI-1, a line derived from the Tubingen line, as well as the Tubingen line (TU) itself (LaFave et al. 2014; Haffter et al. 1996). NHGRI-1 fish were a gift from Shawn Burgess, (National Human Genome Research Institute) and TU fish were a gift from John Rawls (Duke University). AB zebrafish were procured through the Zebrafish International Resource Center (ZIRC). LSB and HSB were gifts from John Godwin, (NC State University) and were only three generations derived from wild-caught populations (Wong et al. 2012). EKW zebrafish were purchased from EkkWIll Waterlife Resources (Ruskin, FL) and CG2 fish were a gift from Sergei Revskoy, (The University of Chicago). AB3 is an inbred line of the highly polymorphic AB zebrafish line (Chakrabarti et al. 1983). CG1 and CG2 are separate, homozygous diploid clonal golden lines of zebrafish (Mizgirev and Revskoy 2014).

Identification of NILTs in Zebrafish

Two NILT genes were initially predicted by Stet et al (2005) on chromosome one in version 5 (Zv5) of the zebrafish genome. Additionally, Stet et al (2005) noted the presence of many similar sequences in the area surrounding these predicted genes. We searched the six frame translation of zebrafish chromosome 1 (Zv9) using HMMER3 (Finn, Clements, and Eddy 2011) profiles generated from Clustal Omega (Sievers and Higgins 2014) alignments of either the D1 or D2 Ig domains of previously described NILTs (Stet et al 2005, Kock et al 2008, Østergaard et al 2009, Østergaard et al 2010). Within zebrafish chromosome 1, scaffold 100 (NW_003040340.2) was found to contain 29 NILT-related Ig domains and adjacent scaffold 101 (NW_001878090.3) was found to contain 75
NILT-related Ig domains. Many of these Ig domains were not encoded by known or predicted transcripts in online databases (e.g., Ensembl and NCBI) (Aken et al. 2016; Pruitt et al. 2016).

In order to annotate genes encoding these 104 Ig domains, our previously reported RNAseq data was reassessed (NCBI accession number SRP057116) (Dirscherl and Yoder 2015). Briefly, RNA from kidney, intestine, gill and spleen of a single CG2 zebrafish was extracted, pooled in equal amounts and sequenced on a HiSeq 2000 (Illumina) platform. The raw RNAseq data was cleaned using Trimmomatic v0.32, mapped to the Zv9 reference genome and assembled into transcripts using TopHat2 v2.1.1 software (Kim et al. 2013; Ghosh and Chan 2016; Bolger, Lohse, and Usadel 2014). Transcripts overlapping the genomic Ig domains identified by the above methods were considered candidate NILTs. Sequences encoding proteins with at least one Ig domain with a CX$_3$C, CX$_6$C, or CX$_7$C motif were classified as a NILT family member.

The 104 NILT Ig domains, CG2 RNAseq, and genomic scaffolds from three zebrafish genomes (CG1, CG2 and AB3) were visualized using Integrative Genomics Viewer (IGV 2.4.1) (Robinson et al. 2011). The CG2 genome was described previously (McConnell et al. 2016) and genomic sequences from CG1 and AB3 lines are maintained by Sean McConnell and Jill L.O. de Jong, (The University of Chicago). Scaffolds from these genome assemblies were aligned to the Zv9 reference genome using BWA with a seed length of 2000 basepairs (Li and Durbin 2009). Dotplots to visualize haplotypic differences were created using PipMaker (http://pipmaker.bx.psu.edu/pipmaker/).

**Confirmation of NGS transcripts and RACE**

Efforts were made to either confirm the computationally assembled CG2 transcriptome, or sequence transcripts with Ig domains not found in the assembled transcriptome through PCR and Sanger sequencing. Primers to amplify the 5' end or the entire open reading frame (5'-3') are listed in Table 1. The 5' gene specific primers (5GSP1 and 5GSP2) were designed against the nucleotide sequences of Ig domains to retrieve the 5' end of cDNA. The 3' primers (3GSP1 and 3GSP2) were designed to
sequence the entire open reading frame. Equal amounts of RNA isolated from kidney, intestine, gills and spleen from either an EKW or CG2 line were pooled and 50 nanograms were reverse transcribed using GeneRacer 5′/3′ RACE (Invitrogen). This cDNA was diluted 1:5 for use as template for polymerase chain reactions (PCR) with Titanium Taq DNA polymerase (Clontech). Additionally, Ig domains found in the genome but absent from RNAseq transcripts were used to design primers for 5′/3′ RACE strategies. Resulting amplicons were cloned into the pGEM-T Easy plasmid (Promega) and sequenced. Sequences are provided in Appendix D.

**Table 1. 5′/3′ RACE Primers.** The RACE strategy included a primary and secondary PCR step, which utilized two gene specific primers (GSP1 and GSP2, respectively) in combination with universal priming sites added to the 5′ and 3′ ends of cDNA during reverse transcription. Ig domains targeted are named after the starting nucleotide location within each respective scaffold.
**Haplotype genotyping**

A PCR based genotyping approach was used to confirm the presence of alternative haplotypes in individual wild-type zebrafish. PCR primers were designed to amplify a 300 bp fragment of genomic DNA unique to scaffold 101 (Forward: TCCATATAATGCTAAATATGAGGA, Reverse: CTCTGTGTATTGCAGTAATTGA). Separate primers were designed to amplify a 200 bp fragment of scaffold 100 (Forward: AGGAAAATGCTCAATAGGTTAT, Reverse: GTCATCAAATTITCCCAGTCTT). PCRs were executed individually and amplicon sequences confirmed. Amplicons were combined prior to agarose gel electrophoresis.

**Sequence Analyses**

Alignments of Ig domains were performed using the default settings of Clustal Omega v1.2.1 (Sievers and Higgins 2014). Boxshade plots were made using version 3.21. Phylogenetic trees were created with MEGA7 (Kumar et al 2016) using neighbor-joining (Saitou and Nei 1987), distance correction by JTT (Jones, Taylor, and Thornton 1992) and 2000 bootstrap replicates (Felsenstein 1985). Pairwise distances between 104 NILT Ig-domains (from zebrafish scaffolds 100 and 101) were calculated by Clustal Omega v1.2.1 and subjected to multiple dimensional scaling used the R package “bios2mds” (Pelé et al. 2012). Clustering of similar Ig domains was performed using the “k-means” function of this software package. After 1000 iterations, silhouette scoring determined four groups was the best fit for the data.

The genomic context of six Atlantic salmon NILT genes was previously reported from a single BAC (Ostergaard et al. 2010). As a reference genome of Atlantic salmon has now been released (Lien et al. 2016), the methods outlined above were employed to map this BAC to chromosome 3 of Atlantic salmon. BLASTP was used to search the Atlantic salmon genome for genes orthologous to those flanking the zebrafish NILT region (*GIMAP8*, *NXNL* and *COLGALT1*).
Results and discussion

All reported NILT sequences contain one or two Ig ectodomains (termed D1 and D2), are typically membrane bound and possess cytoplasmic ITIMs or a cytoplasmic ITAM (Fig 1a) (Stet et al. 2005; Østergaard et al. 2010; Østergaard et al. 2009; Kock and Fischer 2008). A phylogenetic analysis of published NILT Ig domains reveals distinct clusters for D1 and D2 domains. The D1 and D2 clades become further resolved when salmonid or cyprinid families are considered (Fig. 1a). This lineage-specific clustering is indicative of recent gene birth and death events followed by rapid diversification (Nei and Rooney 2005). NILT Ig domains contain a pair of cysteine residues spaced 55-65 residues apart, a defining characteristic of V-type Ig domains (Fig. 1b) (Harpaz and Chothia 1994). NILT Ig domains also possess two additional cysteines which are separated by 3, 6, or 7 residues (CX$_{3,6,7}$C) and predicted to form a second disulfide bridge (Stet et al. 2005). Interestingly, certain CX$_{6}$C-containing Ig domains are more similar to CX$_{7}$C Ig domain than they are to other CX$_{6}$C Ig domains (Fig. 1b): therefore, sequence features other than the CX$_{3,6,7}$C motif must be driving the phylogenetic groups.
**Figure 1.** Sequence comparisons of previously reported NILT Ig domains. A. Phylogenetic tree of Ig domains from published NILTs reveal distinct D1 and D2 clades. Genbank numbers are shown next to gene name and species identifier (SASA; Salmo salar (Atlantic salmon), ONMY; Oncorhynchus mykiss (rainbow trout), CYCA; Cyprinus carpio L. (common carp), DARE; Danio rerio (zebrafish). Immunoglobulin domains (D1 and D2) used in the alignment are indicated with black circles, whereas gray Ig domains were utilized in another part of the tree. The transparent gray circle represents a D2 domain which can be removed by alternative mRNA splicing. TM = transmembrane domain represented by a rectangle. ITIMs are represented by forward slashes whereas ITAMs are represented by a curved line within the cytoplasmic tail (CYT). D1 and D2 domains from pIgR of Atlantic salmon and zebrafish served as outgroups. Protein alignment of NILT (B) D1 and (C) D2 domains. Highly conserved key residues are numbered based on the IMGT system (Lefranc et al. 1999). Black shading indicates residues which are identical in >65% of sequences. Gray shading indicates similar residues conserved across sequences. Dashes indicate gaps in the alignment. Paired cysteines are marked with arrows to indicate predicted disulfide bridges. Triangles represent a conserved complementary determining region (CDR1) also present in PIGR Ig domains (Bakos et al. 1993).
The genomic region surrounding two previously described NILTs on zebrafish chromosome 1 (Stet et al. 2005) was subjected to six-frame translation and the resulting predicted amino acid sequences were searched for the presence of Ig domains. A total of 104 Ig domains with sequence similarities to known NILTs were identified on scaffolds 100 and 101 (Fig. 2a). While rich in Ig domains, this genomic region remains poorly annotated (Fig. 2d). In order to determine which Ig domains are expressed and the organization of individual NILT genes, RNAseq data from immune-related tissues of a homozygous diploid, clonal line of zebrafish (CG2) (Dirscherl and Yoder 2015) was assembled using the reference genome (Zv9). CG2 transcripts encoding NILTs mapped exclusively to scaffold 101 and not scaffold 100 (Fig. 2e). This led to the hypothesis that scaffold 100 and scaffold 101 represent alternative haplotypes with dramatic gene-content variation rather than adjacent scaffolds. Draft genome assemblies of inbred AB3 zebrafish and two homozygous diploid, clonal zebrafish lines, CG1 and CG2 were also mapped onto the Zv9 reference genome (Fig. 2b). CG2 sequences mapped strongly to scaffold 101, whereas AB3 sequences mapped strongly to scaffold 100, supporting the existence of alternative haplotypes. CG1 sequences did not map strongly to either scaffold 100 or 101, suggesting the presence of a third, undefined haplotype (Fig. 2b).

NILT haplotypes

In order to further characterize the haplotypes predicted by scaffolds 100 and 101 (hereafter referred to as hap100 and hap101), genomic alignments were compared. The sequence of hap100 and hap101 differ significantly (Fig. 3a) which is likely why they were predicted to be adjacent scaffolds. However, CG2 scaffolds show significant sequence similarity to hap101, strengthening our conclusion that homozygous diploid CG2 zebrafish encode hap101 and not hap100 (Fig. 3b).
Figure 2. Integrated Genomics Viewer display of Zv9 zebrafish chromosome one NILT region. A. Scaffolds 100 and 101 are found on chr1:58,185,634-58,387,896 and chr1:58,387,970-59,421,814 within the Zv9 reference genome. B. De novo assembled genomic sequences from CG1, CG2 and AB3 lines of zebrafish were mapped onto the Zv9 reference genome. C. Positions of NILT Ig domains, color-coded according to groups defined in Figure 5. D. Gene annotations from Ensembl (v84). E. Assembled RNAseq transcripts from CG2 zebrafish.

Figure 3. Dot plot of alternative NILT haplotypes in zebrafish. A. Genomic alignment of scaffold 100 vs the first 300,000 basepairs of scaffold 101. B. CG2 genomic scaffolds aligned to Zv9 scaffold 101. Reverse complements of CG2 scaffolds were used when appropriate. Plots generated with PipMaker (http://pipmaker.bx.psu.edu/pipmaker/).
A PCR genotyping strategy was utilized to demonstrate hap100 and hap101 are inherited independently. Wild-type zebrafish were genotyped to identify homozygous hap100/100, homozygous hap101/101 or heterozygous hap100/101 individuals (Fig. 4a). Heterozygous male and female zebrafish were identified from a population of EKW zebrafish and mated. The genotypes of their offspring were in the approximate ratio of 1:2:1 (hap100/100:hap100/101:hap101/101) which would be expected for Mendelian inheritance of a single locus (Fig. 4b). This provides strong support that hap100 and hap101 represent alternate haplotypes of the NILT locus on zebrafish chromosome one.

**Figure 4.** Genotyping confirms hap100 and hap101 haplotypes in individual zebrafish. A. Two individuals were genotyped from six wild-type zebrafish lines (NHGRI-1, TU, AB, LSB, HSB and EKW) and one homozygous diploid, clonal line (CG2). B. The offspring from two heterozygous EKW zebrafish were genotyped revealing nine heterozygous hap100/101 offspring, four homozygous hap101/101 and three homozygous hap100/100 individuals. These numbers approximate the expected ratio of 1:2:1 for simple Mendelian inheritance of a single locus.
**Analysis of Ig domains**

The D1 and D2 Ig domains of other IIIR families in zebrafish are often easily resolved through phylogenetic analyses. In the case of the NITR, DICP and PIGRL gene families, phylogenetic trees clearly define two distinct Ig groups (D1 and D2) for each of the gene families (Wcisel and Yoder 2016). While this D1/D2 phylogenetic structure was observed for previously reported NILTs (Fig. 1), a phylogenetic tree of all 104 zebrafish NILT-like Ig domains identified here does not form clear groups (data not shown). An alternative approach using multidimensional scaling was employed to identify subgroups of NILT Ig domains (Fig. 5). Clustering methods identified four subgroups of Ig domains within this gene cluster on zebrafish chromosome 1 (Fig. 5a). The first three principal components identified members of each group (Fig. 5b-c). These assignments (provided in supplemental data), were then used to classify the different types of Ig domains in zebrafish NILTs (Figs. 6 and 7).

**Genomic annotation of zebrafish NILTs**

A detailed genome organization overview of the zebrafish NILT loci reveals that both hap100 and hap101 are flanked by the nxnl and colgalt1 genes. The gimap8 gene is also present within both haplotypes (Fig. 6). We used these three genes to search the Atlantic salmon genome and found that COLGALT2 is in close proximity to the six Atlantic salmon NILTs present on chromosome 3. This appears to be the only identifiable gene in close proximity to the NILT cluster in both zebrafish and Atlantic salmon. The zebrafish NILT region borders a ~400kbp region that is rich in NACHT-domain and leucine-rich-repeat-containing (NLR) proteins and zinc finger proteins. This region has been documented elsewhere, and is believed to have undergone many rearrangements and exon shuffling events (Howe et al. 2016). It is possible that zebrafish NILTs underwent many chromosomal rearrangements with this NLR cluster, which would explain the relative lack of conserved synteny between zebrafish and Atlantic salmon NILTs.
Transcripts for 25 zebrafish NILT genes were identified from the computationally assembled transcriptome and RACE sequencing. When possible, computationally predicted transcripts were confirmed by PCR and Sanger sequencing (Fig. 7). Collectively, these sequences account for 23 of 29 Ig domains on hap100 and 38 of 75 Ig domains on hap101. The 5' RACE strategy failed to amplify transcripts belonging to ‘orphaned’ Ig domains (Table 1). Although these sequences may reflect pseudogenes, it is possible that transcription of these sequences requires stimulation (e.g., infection). Indeed, Stet et. al (2005) noted that the extensive manipulations to culture peripheral blood leukocytes, initially void of NILT transcripts, induced expression of NILT1 in common carp. Seven zebrafish NILT transcripts were confirmed by PCR and Sanger sequencing (Fig. 7). The zebrafish NILT repertoire includes activating, inhibitory, secreted, and ambiguous forms of NILTs (Fig. 7). Additionally, we noted cytoplasmic tails of three NILTs encoded a TVYAT motif of unknown function. The predicted amino acid sequences of all previously described NILTs contain one or two Ig domains. Here, we report that the zebrafish genome encodes NILTs containing up to six predicted Ig domains.
Figure 5. Multidimensional scaling of 104 zebrafish NILT Ig domains. A. Average silhouette scores (highly confident group assignment have a score near 1, poorly assigned elements have a score near -1) of 1000 iterations are plotted against number of groups (Index) assessed. The optimum number of groups in the dataset was found to be four using the “k-means” cluster function of the software package. B. 3D scatterplot of principal components (PC1, PC2 and PC3). C. 2D scatterplots of PC1, PC2 and PC3. Ig domains were segregated into four groups based on the first three principal components (PC1-PC3) and color-coded accordingly. Group 1: blue, 2: green, 3: red, and 4: black. Genomic positions of individual Ig domains within each group is depicted in Figures 2 and 6.
Figure 6. Genomic annotation of zebrafish NILT Ig domains and genes. Ig domains, named for their nucleotide starting positions are represented by color coded (as in Figure 5) rectangles. White pentagons reflect gene annotations performed by Gnomon (Souvorov et al. 2010). Yellow pentagons represent genes annotated from the CG2 RNAseq data. Green pentagons represent genes confirmed through RACE, PCR and Sanger sequencing. Pentagons are placed below the Ig domains they encode. Black pentagons represent non-NILT genes. A large number of NACHT-domain and leucine-rich-repeat-containing (NLR) sequences are found immediately downstream of colgalt1 gene on scaffold 101. Stet et. al. previously reported nilt1 and nilt2 in the Zv5 reference genome. nilt1 was not predicted in the Zv9 genome, however the location of nilt2 (LOC100537984) is shown.
Figure 7. Repertoire of zebrafish NILTs. A. hap100 encodes five membrane bound receptors, two of which encode an inhibitory motif. Three presumed secreted NILTs are encoded. B. hap101 encodes a single activating NILT and three inhibitory NILTs. Full length 3’ RACE sequences were gathered for six genes (Appendix D).
Conclusions

The results discussed here lay the foundation for studies which seek to explore the origins and evolution of vertebrate innate immunity. The intraspecific gene content variation of the NILT gene family may contribute to inter-individual differences in immune health and confound immunological studies employing zebrafish as a model. Nevertheless, these differences are more representative of the diversity within the human population. Gene-content haplotypes, reminiscent of those identified here, have been associated with disease susceptibilities. For example, the human KIR gene family exhibits gene-content haplotypes and mothers with certain combinations of KIRs (and their MHC ligands) are predisposed to preeclampsia (Hiby et al. 2004). Therefore, the contribution of NILT gene-content haplotypes to population health, as well as individual health, is of significant interest.

Functional studies will likely reveal important features of these genes. The conservation of NILTs between cyprinid and salmonid species boundaries suggest they have maintained an important immunological role. It will be of great interest to investigate NILTs in more diverse teleost species. However, the large degree of sequence variation and gene copy numbers between species suggests that some NILT genes may have duplicated, diverged and gained novel functions. As zebrafish are highly amenable to gene disruption (knockout) technologies, we predict that if NILTs play an important role in the immune response, reverse genetics studies will quickly reveal a phenotype and lead to additional studies (LaFave et al. 2014; Hwang et al. 2013).

Acknowledgements

We thank James (Thomas) Howard for his technical support. This work was supported by the National Institutes of Health (R01 AI057559 to GWL and JAY and R01 AI23337 to GWL) and the Triangle Center for Evolutionary Medicine (TriCEM).
References


CHAPTER 4: SPOTTED GAR AND THE EVOLUTION OF INNATE IMMUNE RECEPTORS

RESEARCH ARTICLE

Spotted Gar and the Evolution of Innate Immune Receptors

DUSTIN J. WCISEL¹, TATSUYA OTA², GARY W. LITMAN³, AND JEFFREY A. YODER¹,4,5

¹Department of Molecular Biomedical Sciences, North Carolina State University, Raleigh, North Carolina, USA
²Department of Evolutionary Studies of Bioystems, SOKENDAI (The Graduate University for Advanced Studies), Hayama, Japan
³Department of Pediatrics, University of South Florida Morsani College of Medicine, St. Petersburg, Florida, USA
⁴Comparative Medicine Institute, North Carolina State University, Raleigh, North Carolina
⁵Center for Human Health and the Environment, North Carolina State University, Raleigh, North Carolina, USA

ABSTRACT

The resolution of the gar genome affords an opportunity to examine the diversification and functional specialization of immune effector molecules at a distant and potentially informative point in phylogenetic development. Although innate immunity is effected by a particularly large number of different families of molecules, the focus here is to provide detailed characterization of several families of innate receptors that are encoded in large multigene families, for which orthologous forms can be identified in other species of bony fish but not in other vertebrate groups as well as those for which orthologs are present in other vertebrate species. The results indicate that although teleost fish and the gar, as a holostean reference species, share gene families thought previously to be restricted to the teleost fish, the manner in which the members of the multigene families of innate immune receptors have undergone diversification is different in these two major phylogenetic radiations. It appears that both the total genome duplication and different patterns of genetic selection have influenced the derivation and stabilization of innate immune genes in a substantial manner during the course of vertebrate evolution. J. Exp. Zool. (Mol. Dev. Evol.) 00:1–19, 2017. © 2017 Wiley Periodicals, Inc.

INTRODUCTION

It has been suggested that the high level of diversity between innate immune receptor genes from different species reflects the evolutionary history of pathogen exposures for that species' lineage (Jack, 2019). In addition, the coevolution with commensal bacteria has likely further refined the host's innate immune system (Thakur et al., 2016). As the genomes of more diverse animals are reported, it has become clear that many species...
possess immune system strategies that differ dramatically from mammalian species (Buenoocore and Gerold, 2016). The merits of novel immune strategies include strong selection pressures that can influence rates of speciation and animal diversification (Malmström et al., 2016). Ray-finned fish (Actinopterygii) comprise the largest group of vertebrate species with over 30,000 members (Echmeyer et al., 2016) and nearly all of these (>99.8%) are teleost fishes (Nelson et al., 2016), which include model organisms, such as zebrafish (Danio rerio) and Japanese medaka (Oryzias latipes) and agriculturally important species (e.g., carp, catfish, and salmonids). Genomic analyses have revealed variation in the sequence and numbers of innate immune receptors in diverse teleost species with much of this diversity arising from recent gene birth and death events within multiple multigene families of innate immune receptors (Malmström et al., 2016; Wcislo and Yoder, 2016).

The diversification of immune-related genes can be accelerated through gene duplication (gene birth) events followed by species-specific adaptations. Whole genome duplications (WGD) provide a massive number of genes the freedom to evolve. The fate of duplicated genes may be nonfunctionalization (the acquisition of new functions), subfunctionalization (all functions of the original gene are maintained, but divided between the gene copies), or non-functionalization (acclimatization of deleterious mutations resulting in gene death) (Kuraku and Meyer, 2010). Two rounds of early vertebrate genome duplication (VGD1 and VGD2) (Dehal and Boore, 2005) and a subsequent teleost genome duplication (TGD) (Amores et al., 1998; Taylor et al., 2003; Glassner and Noltenius, 2014) have likely contributed to the great diversity of innate immune receptors present in teleosts (Malmström et al., 2016; Wcislo and Yoder, 2016). In addition to the TGD, certain teleost lineages (e.g., salmonoids, carp, and suckers) have experienced additional, more recent WGD events (Pasquier et al., 2016).

Recently, the genome of spotted gar (Lepisosteus oculatus) (hereafter gar), a ray-finned fish but a holostean outgroup to the TGD, was annotated and reported (Bransch et al., 2016). It was demonstrated that gar has the ability to bridge the gap between teleost and mammalian genomes, facilitating cross-species comparisons and illuminating hidden orthologies. The gar genome, when compared to teleosts, is considered slowly evolving; thus, it has preserved a more ancient genome structure. This permits a unique perspective into the evolution of vertebrate immunity and the impact TGD may have laid on immune-related loci.

Here, we offer a detailed description of multiple families of innate immune receptors encoded by gar including Toll-like receptors (TLRs) and families of fish-specific immunoglobulin (Ig) domain containing innate immune receptors (IIRs) (Fig. 1). We demonstrate that while many gar TLR genes share more sequence similarities with TLRs encoded by teleost fish (e.g., zebrafish), their genomic organization often resembles that of human TLRs. We characterize gar sequences that encode inhibitory and activating forms of novel immune-type receptors (NITRs), which are predicted to function as natural killer cell receptors (Yoder and Litman, 2011). We identify gar sequences that encode diverse immunoglobulin domain-containing proteins (DCPs) and demonstrate their linkage to major histocompatibility complex (MHC) sequences in cyprinids (zebrafish), Lepisosteiformes (gar), and Latimeria (coelacanth), a lobe-finned fish that branched from ray-finned fishes ~450 mya and shares a common ancestor more recently with mammals than teleosts. The failure to identify orthologs of specific tetrapod innate immune receptors indicates that, although gar is an outgroup to the TGD, its innate immune receptor repertoire more closely resembles that of teleosts than tetrapods. We have integrated these observations into an updated view of the evolutionary origins of these families of innate immune receptors.

**METHODS**

Genes initially were identified by searching Ensembl’s database (release 79) of zebrafish and human TLRs, NITRs, and DCPs genes for gar orthologs identified by the Ensembl’s annotation and Gene Orthology/Paralogy pipeline (Villemia et al., 2009; Aken et al., 2016). For NITRs and DCPs, the surrounding genomic loci of the gar genome (LepOcu1) also was searched with tBLASTn using both zebrafish and gar NITR/DCP protein sequences for the presence of genes mixed by prediction models (Altschul et al., 1990; Camacho et al., 2009). Synthentic comparisons were performed using Genomica v8.4.0 (Lobos et al., 2015). Multiple sequence alignments were performed with Clustal Omega (Sievers et al., 2011). Phylogenetic analyses were performed in MEGA7 (Kumar et al., 2016). Neighbor-joining trees were constructed using JTT matrix-based method to compute evolution distance, with pairwise deletion of ambiguous positions and 1000 bootstrap replicates (Saitou and Nei, 1987; Jones et al., 1992). Protein domains were identified using SMART (Letunic et al., 2015).

**RESULTS AND DISCUSSION**

Gar TLRs Share Evolutionary Histories of Both Humans and Zebrafish

The family of TLRs, which are present in vertebrates and invertebrates, recognize a variety of pathogen-associated molecular patterns (PAMPs) providing one of the initial immune responses to infection (Ademre and Ulevitch, 2008). The TLR family includes 10 receptors in humans (TLR1-TLR10) and 12 in mice (TLR1-TLR13), lacks TLR8 and 10. Teleost genomes can encode more than 20 TLR genes (Quintinou et al., 2013), which may reflect nonfunctionalization after the TGD. For example, zebrafish and channel catfish (Ictalurus punctatus) contain duplicate copies of genes encoding TLR4, TLR5, and TLR8 and at least six other "nonmammalian" TLR genes (Jault et al., 2004; Pošík, 2011; Jamaina et al., 2013; Kanwal et al., 2014). Vertebrate TLRs can be organized into six subfamilies based on amino acid sequence similarities: TLR1, TLR3, TLR4, TLR5, TLR7, and TLR11, each
of which can contain multiple members (e.g., TLR7, TLR6, and TLR8 are all part of the TLR7 subfamily) (Roach et al., 2005).

TLRs recognize a variety of PAMPs through an N-terminal ectodomain consisting of numerous leucine-rich repeats (LRR) (Fig. 1). The number of LRRs varies between TLR members and assigning TLR gene homology across species based on LRR sequence alone can be difficult due to the variable length of extensions within LRR motifs (Matsushima et al., 2005, 2007). Usually, cysteine clusters cap the N-terminal side of the LRRs and flank the C-terminal side and are termed LRRN and LRRC, respectively, and are thought to stabilize the hydrophobic LRRs (Gay and Gangloff, 2007).

TLRs are type I transmembrane proteins and are expressed either on the cell surface or on endosomes. The cytoplasmic C-terminal region contains a TOLL/IL1 receptor (TIR) domain (Fig. 1). Upon ligand recognition, TLRs recruit cytoplasmic adapter molecules such as MyD88, MAL, TIRF, TIRF-related adapter molecule (TRAM), and SARM in order to initiate an immune response (reviewed in Kawai and Akira, 2006). While the N-terminal LRR-containing regions are highly variable between species, the TIR domain is highly conserved and amenable to phylogenetic analyses (Beutler and Reh, 2002; Joule et al., 2004; Quiniiou et al., 2013; Boudriot et al., 2014). Ensembl’s annotation and Gene Orthology/Paralogy pipeline (Vilella et al., 2009) predicts 17 TLR-related genes within the gar genome (Table 1). Thirteen of these 17 gar sequences were annotated using zebrafish TLRs as reference, while four gar TLR sequences remain uncharacterized (tlrd, tlrh, tlrl, tlrb).

Based on phylogenetic analysis of TIR domains (Fig. 2), we confirmed all but one of the annotations and account for the remaining uncharacterized TLRs utilizing West Indian Ocean cokelanth (Latimeria chalumnae) and channel catfish reference sequences. We find that representative members of all six TLR subfamilies are encoded in the gar genome. Here, we illustrate the genomic context of gar TLRs, infer synteny relationships between human and zebrafish using gar as a reference, and make predictions of the evolutionary histories of these gene families.

**TLR1 Subfamily.** Members of TLR1 subfamily bind a range of ligands, which often originate from bacteria (reviewed in Gay and Gangloff (2007) and Zhang et al., 2014). For example, mammalian TLR1 is able to heterodimerize with TLR2 in order to recognize triacyl lipopeptides originating from mycobacteria, whereas TLR2/TLR6 heterodimers recognize diacylated lipopeptides (Takeuchi et al., 2002). Similar to other vertebrates, teleost TLR1 and TLR1 subfamily members often lack the LRRNT flanking sequence (Zhang et al., 2014). This absence of LRRNT is thought to be an important feature allowing TLR1 and TLR2 molecules to dimerize and suggests that TLR1/TLR2 heterodimers might also occur in fish (Quiniiou et al., 2011).

The TLR1 subfamily has been described as containing more species-specific adaptations than any other TLR subfamily.
### Table 1. Gliar Toll-like receptors (TLRs)

<table>
<thead>
<tr>
<th>TLR subfamily</th>
<th>Protein name in Figure 2</th>
<th>Recommended gene name</th>
<th>Ensembl identifier</th>
<th>Location</th>
<th>Syntenic support</th>
<th>Protein domains*</th>
</tr>
</thead>
<tbody>
<tr>
<td>TLR 1</td>
<td>Tir1</td>
<td>tir1</td>
<td>ENSLGC000000012910</td>
<td>L64:62,158,233-62,160,668:1</td>
<td>ZF</td>
<td>SP/ED/CT/ TM/TIR</td>
</tr>
<tr>
<td></td>
<td>Tir2</td>
<td>tir2</td>
<td>ENSLGC000000018220</td>
<td>L64:21,718,623-21,726,022:1</td>
<td>Hu</td>
<td>SP/ED/CT/ TM/TIR</td>
</tr>
<tr>
<td></td>
<td>Tir18</td>
<td>tir18</td>
<td>ENSLGC000000007892</td>
<td>L624:6,943,613-6,955,636:1</td>
<td>–</td>
<td>SP/ED/CT/ TM/TIR</td>
</tr>
<tr>
<td></td>
<td>Novel (Tir[a])</td>
<td>tir25a</td>
<td>ENSLGC00000016881</td>
<td>L68:48,599,630-48,602,032:1</td>
<td>–</td>
<td>SP/ED/CT/ TM/TIR</td>
</tr>
<tr>
<td></td>
<td>Novel (Tir[b])</td>
<td>tir25b</td>
<td>ENSLGC00000017911</td>
<td>L68:25,066,396-25,068,994:1</td>
<td>–</td>
<td>SP/ED/CT/ TM/TIR</td>
</tr>
<tr>
<td></td>
<td>Novel (Tir[c])</td>
<td>tir27</td>
<td>ENSLGC00000017732</td>
<td>L10:1,032,739-1,035,096:1</td>
<td>–</td>
<td>SP/ED/ TM/TIR</td>
</tr>
<tr>
<td></td>
<td>Novel (Tir[d])</td>
<td>Pseudogene</td>
<td>ENSLGC00000017625</td>
<td>L19:1,189,904-1,191,199:1</td>
<td>–</td>
<td>TIR/ TM/ TM</td>
</tr>
<tr>
<td>TLR 3</td>
<td>Tir3</td>
<td>tir3</td>
<td>ENSLGC00000013826</td>
<td>L64:70,374,961-70,381,739:1</td>
<td>HuZF</td>
<td>SP/ED/CT/ TIR</td>
</tr>
<tr>
<td>TLR 4</td>
<td>Tir4a</td>
<td>tir4</td>
<td>ENSLGC00000003751</td>
<td>L621:5,778,346-5,781,017:1</td>
<td>Hu</td>
<td>ED/ TM</td>
</tr>
<tr>
<td>TLR 5</td>
<td>Tir5a</td>
<td>tir5</td>
<td>ENSLGC00000018000</td>
<td>L61:32,406,982-32,409,585:1</td>
<td>HuZF</td>
<td>ED/ CT/ TM/ TIR</td>
</tr>
<tr>
<td>TLR 7</td>
<td>Tir7</td>
<td>tir7</td>
<td>ENSLGC00000009977</td>
<td>L61:21,059,390-21,065,834:1</td>
<td>HuZF</td>
<td>SP/NT/ED/ CT/ TIR</td>
</tr>
<tr>
<td></td>
<td>Tir7a</td>
<td>tir7a</td>
<td>ENSLGC00000009990</td>
<td>L61:21,075,973-21,079,392:1</td>
<td>HuZF</td>
<td>SP/NT/ED/ CT/ TM/ TIR</td>
</tr>
<tr>
<td></td>
<td>Tir7b</td>
<td>tir7b</td>
<td>ENSLGC00000008982</td>
<td>L61:21,068,100-21,073,657:1</td>
<td>HuZF</td>
<td>SP/ED/CT/ TM/ TIR</td>
</tr>
<tr>
<td></td>
<td>Tir8a</td>
<td>tir8a</td>
<td>ENSLGC000000014202</td>
<td>LG6:48,972,573-48,981,170:1</td>
<td>HuZF</td>
<td>SP/ED/CT/ TM/ TIR</td>
</tr>
<tr>
<td></td>
<td>Tir8b</td>
<td>tir8b</td>
<td>ENSLGC000000014430</td>
<td>LG6:50,478,035-50,482,646:1</td>
<td>–</td>
<td>SP/ED/ TM/ TIR</td>
</tr>
<tr>
<td>TLR 11</td>
<td>Tir19</td>
<td>Novel</td>
<td>ENSLGC00000015687</td>
<td>LG1:13,198,517-13,200,058</td>
<td>–</td>
<td>ED/ CT/ TM/ TIR</td>
</tr>
<tr>
<td></td>
<td>Tir21/22</td>
<td>tir22</td>
<td>ENSLGC00000018316</td>
<td>L66:38,022,636-38,026,533:1</td>
<td>–</td>
<td>SP/ED/ CT/ TM/ TIR</td>
</tr>
</tbody>
</table>

*Evidence for conserved synteny presented in Supplementary Fig. S1 and Table S1. ZF: zebrafish; Hu: human.

*Protein domains: SP = signal peptide, NT = N terminal LRR cap, ED = ectodomain, CT = C terminal LRR cap, TM = transmembrane, TIR = TOL/TL1 receptor domain.
Figure 2. Phylogenetic relationships between gar, teleost, mammalian, avian, reptilian, and coelacanth TLR TIR domains. Four letter codes are used for each species: spotted gar (Epinephelus coioides, Lcog), coelacanth (Latimeria chalumnae, Lach), zebradish (Danio rerio, Zdb), human (Homo sapiens, Hsna), mouse (Mus musculus, Mmum), anole lizard (Anolis carolinensis, Ancra), chicken (Gallus gallus, GaGa), and channel catfish (Ictalurus punctatus, Icpa). Coelacanth TLRs were described by Boudinot et al. (2014b). Coelacanth TLR paralogs ENSLACG0000000008376 and ENSLACG000000004773 were included in this analysis, although they may represent pseudogenes: ENSLACG0000000008376 possesses a limited number of LRRs and ENSLACG000000004773 encodes a partial TIR domain. Channel catfish TLRs were annotated by Quiniou et al. (2015). Gar Tlr4 lacks a TIR domain and is excluded from this analysis. Bootstrap values are indicated next to the branches; values less than 50 are not shown. Sequences are provided in Supplementary Table S2.
(Roach et al., 2005). Seven gar genes were determined to belong to the TLR1 family based on the phylogenetic analysis of their TIR domains (Table 1, Fig. 2). Gar trl1 displays some synteny to zebrafish trl1, however there is no evidence for conserved synteny with humans or any other nonfish species (Table 1). The four gar TLRs, which were not given an Ensembl annotation, likely arose through gene- or fish-specific adaptations as compared with zebrafish and humans revealed only weak synteny (Supplementary Fig. S1 and Table S1). Although gar trl1 (ENSLGCG000000017625) groups out with a coelacanth TLR gene, it is unlikely either function as receptors given the arrangement of detected protein domains (TLR/MD/MD, Table 1). Through phylogenetic analyses of the TIR domains (Fig. 2), two additional unannotated gar TLRs (trl4/30/ENSLGCG00000016891 and trl6/6/ENSLGCG00000017711) appear related to TLR25, which was first identified in carpids but determined to also be present in atraciichthyids, cichlids, cyprinids, and plecostomids, suggesting its presence early in fish evolution (Quinioiu et al., 2013). Gar trl4/30 shares significant synteny with tilapia TLR25 (not shown). Interestingly, the tilapia TLR25, ENSONGCG00000004149, which is encoded on an unplaced scaffold in the latest Orel9.1 genome assembly (Brawand et al., 2014), is considered synonymous with the gene name TLR1. TLR25 sequences from tilapia, carpids, and gar (trl4/30 and trl6/6) lack the IRINT domain, which may allow for the dimerization with TLR2, increasing the repertoire of ligand specificities. Together, this suggests that TLR25 resulted from a gene duplication event of TLR1 loci. Its presence in gar, a much more basal species, suggests that a TLR1 duplication and subsequent divergence may have predated the EDG, which may explain why so many TLR1 subfamily members are observed in teleosts (Fig. 2). Gar also encodes trl1 (ENSLGCG00000017711) that clusters very tightly with both coelacanth and elephant shark (Galeorhinus milli) TLR27 (Boudinot et al., 2014b; Wang et al., 2015). TLR10 was estimated to have diverged from its TLR1-like precursor ~300 million years ago (mya), nearly 150 million years after the divergence of ray-finned fishes (Beutler and Rehli, 2002). In agreement with this, gar and teleost genomes do not encode orthologs of TLR10. Likewise, TLR6 is thought to have emerged 130 mya and thus is also absent from fish genomes.

TLR3 Subfamily. It has been demonstrated that both mammalian and teleost TLR3 proteins are able to recognize and respond to dsDNA, the genetic material of many viruses, to induce the expression of type I interferons (Roach et al., 2005; Huang et al., 2011; Samanta et al., 2013). Phylogenetic analyses place gar trl3 (ENSLGCG0000001826) in a group with teleosts (Fig. 2), however, it shares an almost equivalent amount of synteny with human TLR3 loci (Table 1, Supplementary Fig. S1). Although it is tempting to predict that gar TIR is also able to recognize and respond to dsDNA, caution should be used when generalizing the function of TLRs across species. For example, sea cucumber

TLR3 (Apostichopus japonicus) is transcriptionally upregulated following exposure to bacterial PAMPs, suggesting a dual or alternative role of TLR3 (Lu et al., 2013). Many differences in TLR response to PAMPs are observed across species and are well illustrated in the case of TLR4.

TLR4 Subfamily. Mammalian TLR4, in conjunction with myeloid differentiation protein 2 (MD-2), initiates an immune response against lipopolysaccharide (LPS), the major cell wall component of Gram-negative bacteria (Beutler, 2000). Additionally, LPS-binding protein (LBP) and CD14 aid in separating LPS aggregates and delivering LPS to the TLR4/MD-2 complex (Wright et al., 1990; Beutler, 2000; Su et al., 2000; Kim et al., 2000). Prolonged activation of TLR4 can lead to endotoxic shock, a situation that can be fatal to the host and is characterized by the overproduction of cytokines such as tumor necrosis factor-a and interleukin-6 (Morrison and Ryan, 1997). Fish and amphibians display notable resistance to LPS-induced endotoxic shock when compared to mammals (Bencze and Bereczki, 1966; Beutler and Rehli, 2002) suggested that LPS sensing by TLR4 may be a unique adaptation to mammals as TLR4 was not found in the available fish genomes. Following the completion of a quality draft of the zebrafish genome, Jautl et al. (2004) determined zebrafish contains 19 putative TLR variants including two copies of TLR4 (tria and trib), although accessory proteins LBP, MD-2, and CD14 remain absent. Although zebrafish tria and trib do not recognize LPS, overexpression of zebrafish tria/TIR domain negatively regulates NF-kB activation (Sepulcre et al., 2009; Sullivan et al., 2009). As morpholino knockdowns of tria, trib, and myd88 in zebrafish was not sufficient to disrupt the response to LPS, Sepulcre et al. (2009) concluded that LPS sensing in zebrafish occurs independently of TLR4/MD-2 pathway.

Gar encodes a TLR4 ortholog (ENSLGCG000000031791), however it shares minimal synteny with humans and none with zebrafish (Table 1, Supplementary Fig. S1 and Table S1). The predicted protein sequence of gar TIR4 lacks a TIR domain while the ectodomain displays a high degree of sequence divergence from other TLR4 genes (less than 31% identical to human). Accessory proteins MD-2 and CD14 are not readily discernible in gar, although LBP was identified (ENSLGCG0000006158). Therefore, we predict that LPS sensing by gar TIR4 is unable to activate an immune response through either the Myd88-dependent pathway or the TRAM signaling pathway as observed in mammals (Miggion and O’Neill, 2016). TLR4 likely diverged from a TLR2/TLR4 precursor around the time of the emergence of vertebrates, approximately 500 mya (Beutler and Rehli, 2002). The absence of TLR4 from coelacanth and lack of TIR domain and accessory proteins in gar suggest that TLR4 function was dispensable in fish or that LPS sensing by TLR4 may be a mammalian innovation (Boudinot et al., 2014a).

TLR5 Subfamily. Both mammalian and teleost TLR5 recognize bacterial flagellum (Munoz et al., 2013). In some teleosts, such
as rainbow trout (Onchorhynchus mykiss) and gilthead seabream (Sparus aurata), two TLR5 genes have been described: one encodes a membrane-bound TLR5M and the other encodes a soluble TLR5S that lacks the TIR domain necessary for classical TLR signaling. Even in the absence of a TIR domain, TLR5S is believed to enhance the immune response in rainbow trout (Tsujita et al., 2004). However, in gilthead seabream, it was observed that the binding of TLR5S to flagellum blocks TLR5M activation and therefore downregulates the immune response (Mutoh et al., 2013). Coelacanths also are found to encode TLR5S and TLR5M (Boulebnet et al., 2014b); however, since TLR5S lacks a TIR domain, orthology to teleost TLR5S was not examined. Nonetheless, the presence of TLR5S in coelacanths indicates an ancient origin for this sequence.

In other teleost species, such as Atlantic cod (Gadus morhua), TLR5 genes have been lost altogether. Atlantic cod has many immune system abnormalities including the apparent loss of MHC class II and large expansion of MHC class I molecules (Star et al., 2011; Buonocore and Gerdol, 2016). The single copy of gar thr5 encodes a transmembrane domain and TIR domain and is therefore most similar to TLR5M. TLR5S was not detected in gar, and may be restricted to certain teleost species. Adjacent genomic regions of TLR5 remain highly conserved among human, gar, and zebrafish (Supplementary Fig. S1 and Table S1).

**TLR7 Subfamily.** Mammalian members of the TLR7 subfamily recognize nucleic acid motifs; however, it is not known if the ligands are conserved in fish (Rauta et al., 2014). The TLR7 subfamily, which includes TLR7, TLR8, and TLR9, have arisen prior to the rise of vertebrates, some 500 million years ago and is therefore found throughout vertebrate taxa (Beutler and Rehli, 2002). Gar thr7, thr8a, and thr8b are found in tandem on LG14 in the same basic organization as observed in humans (Supplementary Fig. S1). While zebrafish thr7 and thr8 are encoded in tandem on chromosome 9, zebrafish thr8b is encoded on chromosome 10. This may be reflective of a single gene transposition as syntenic genes surrounding zebrafish thr8b are not observed (Supplementary Table S1). Duplicated copies of TLR7 are found on distinct loci on gar LG5 (Table 1). Gar TLR7 genes have diverged significantly and are only approximately 40% identical to each other. An expansion of TLR9 genes has been observed in Atlantic cod, which may be a compensatory adaptation in response to the loss of TLR5 and MHC class II sequences (Star and Jenoe), 2012). While gar does encode TLR5 and MHC class II sequences, an expansion of TLR9 may be advantageous for some fish lineages.

**TLR11 Subfamily.** The TLR11 subfamily includes TLR11, TLR12, TLR13, and many others that may be species specific. Humans lack a functional TLR11 ortholog, whereas other vertebrate species have many TLR11 members. For example, TLR21 is common to birds, amphibians, and fish but absent from humans. The ligands for most TLR11 subfamily members remain unknown; however, it has been determined that mouse TLR11 and TLR12 recognize ligands originating from protozoan parasites while TLR31 from chicken and zebrafish recognizes CpG DNA common to bacteria (Koetsier et al., 2010; Yeh et al., 2013a; Zhang et al., 2014) and TLR22 in pufferfish (Takifugu rubripes) recognizes dsRNA (Matsumo et al., 2008). There are a number of fish-specific TLRs that belong to the TLR11 subfamily (e.g., garfish TLR36). The gar genome encodes two TLR11 subfamily members: thr19 (ENSLGC00000015687) and thr21 (ENSLGC00000018314), neither of which share synteny with zebrafish nor humans (Table 1). Gar thr19 is highly diverged and is supported by low (34%) bootstrap values (Fig. 2). Therefore, this gene may represent a gar-specific adaptation and we recommend it remain classified as novel (Table 1).

As mentioned above, zebrafish thr21 is functionally similar to mammalian TLR9 in that it recognizes CpG DNA, although the specific CpG sequences recognized differ significantly between the two types of receptors (Yeh et al., 2013b; Zhang et al., 2014). Interestingly, gar thr21 appears to share an evolutionary history with the TLR11 subfamily, instead of the TLR7 subfamily that might be expected based on shared functions with TLR9 (Fig. 2). While the functions of gar TLR genes have not yet been determined, it is interesting that gar possesses two divergent copies of TLR9 as well as a TLR21. This may indicate that gar has a heightened response to CpG and is able to recognize different CpG DNA motifs.

"Teleost-Specific" Inflammatory Receptors Are Present in Gar

Although TLRs are ubiquitous throughout the vertebrate kingdom and represent an ancient mechanism of innate immunity (Beutler and Rehli, 2002; Roach et al., 2005; Rauta et al., 2014), natural killer receptor (NK) gene families have been described as recently and rapidly evolving (Manter et al., 2015; Carrillo-Bustamante et al., 2016). The recent evolution of NKRs has led to many species-specific adaptations. For example, in humans the killer cell immunoglobulin-like receptor (KIR) gene family encodes activating and inhibitory NKRs that bind MHC class I or MHC class I-like molecules to differentiate between normal cells ("self") and infected or transformed cells ("nonself") (Jamal and Khakoo, 2011; Grethlein et al., 2015). Engagement of activating KIRs can result in the direct killing of target ("nonself") cells. In contrast, mice utilize the C-type lectin family of Ly49 NKRs to differentiate between "self" and "nonself" (Rakim and Makrigiannis, 2015). While fish species do not encode genetic orthologs of either the KIR or Ly49 families, it has been hypothesized that the NTR family, which belongs to the larger IR-like group of receptors and has been well described in teleosts, may function as NKs in fish (Yoder, 2009; Yoder and Litman, 2011). Telesost NTRs are similar in structure to mammalian KIRs and when recombinantly activating or inhibitory NTRs are expressed on human cells they can initiate NKR signaling pathways (Yoder et al., 2001; Wei et al., 2007). In addition to NTRs,
Teoleost genomes encode other families of IIIRs that are similar in structure to NITRs such as DzCzPs, polymeric immunoglobulin receptor like proteins (PIGLRs), leukocyte immunoglobulin-like receptors (LITRs), and novel immunoglobulin-like transcripts (NILTs) (Weisel and Yoder, 2014). Teoleost IIIR gene families encode membrane-bound and secreted proteins, which possess one or more extracellular Ig domains that are predicted to bind undefined ligands (Fig 1). Membrane-bound IIIRs typically possess activating or inhibitory signaling motifs. Activating IIIRs may possess a cytoplasmic immunoreceptor tyrosine-based activation motif (ITAM: Yxx/I[LxA]-Yxx/V) or a charged residue within the transmembrane domain, which permits the association with an adaptor protein, such as DAP12, which possesses a cytoplasmic ITAM. Inhibitory IIIRs possess one or more immunoreceptor tyrosine-based inhibition motifs (ITIM: S/TXY/LYxxIVL). ITIM signalling opposes T-cell receptor signalling and the balance between these signals can determine the activation state of an immune cell (Barrow and Trowsdale, 2006).

Although genomic DzCzP sequences have been reported from the coelacanth (Boudinot et al., 2014a) NITRs, PIGLRs, LITRs, and NILTs have previously been reported only in select teleost genomes (Weisel and Yoder, 2016). Here, we investigated the gene for homologs of these gene families. We predict that the genome encodes multiple inhibitory NITRs and a single activating NITR gene. We describe several DzCzP sequences from gar and substantiate the claim that DzCzPs are present in coelacanth. Lastly, we report that the putative single copy pigr gene in gar (ENSLOC00000001.203.3) did not undergo an expansion of PIGL-like genes as reported in certain teleost species (Kortum et al., 2014). We are currently unable to identify conclusive homologs of teleost NILTs or LITRs in gar (see below).

We also investigated the gene for homologs of a number of mammalian IIIR gene families. Human CD100s, orthologs to CMRFs in model species such mice, are an IIIR family composed of eight to ten members, although many more CD100-like genes exist (reviewed in (Borrego, 2013). Ensembl identifies three potential orthologs to human CD100 on gar LG3. CD100s share many sequence similarities to teleost NILTs, which have only been described in species that did not have genomic resources at that time (Siret et al., 2005; Koch and Fischer, 2008; Østergaard et al., 2009; Østergaard et al., 2010). Synteny was not observed between LG1 and any mammanlian or teleost model in which NILTs have been described. Therefore, it is not clear at this time if these IIIR-like genes on gar LG3 are CD100s or NILTs. FeR-like molecules, which are found in mammals and teleost fish, are readily identifiable in gar (e.g., ENSLOC000000009867, ENSLOC000000009868, ENSLOC000000009869, ENSLOC000000009870, ENSLOC000000009871, ENSLOC000000009866, and ENSLOC000000009865) and dispersed throughout the gar genome (LG3, LG5, and one unplaced scaffold). As mentioned above, it is unclear if definitive orthologs of the teleost LITR gene family, a subset of which are markers for cytotoxic T cells (Taylor et al., 2016), are present in gar. Transcripts encoding inhibitory and activating LITRs have been described extensively in catfish (Stafford et al., 2006, 2007) and LITR Ig domains can be identified in the zebrafish genome (Rodriguez-Nunez et al., 2014). BLAST searches of the gar genome with catfish LITR sequences as queries identify multiple sequences annotated as FeR-like. The sequence similarity between LITR Ig domains and those of FeR (as well as FeR-like, KiR, LIR, and NK46) has been reported (Stafford et al., 2006) and confounds the ability to definitively identify LITRs in gar. Further, the limited information on the genomic organization and syntetic relationships of teleost LITR gene cluster(s) currently prohibits the characterization of these genes in gar. Gar orthologs to human KIRs, Ly49s, NKp44, NKp46, and LILRs were not detected in our analyses.

Novel Immune-Type Receptors. Teoleost NITRs are expressed by lymphocytes (Yoder et al., 2010), function in allore cognition (Cannon et al., 2008), and, when expressed as recombinant proteins on mammalian cells, can influence NK cell function (Yoder et al., 2001; Wei et al., 2007). These features have led to their description as putative "functional orthologs" of mammalian NKR receptors (Yoder and Litman, 2011). NITRs commonly possess two extracellular Ig domains: one of the variable (V) type and one of the intermediate (I) type (Yoder, 2009). NITR V domains are highly similar to V domains encoded by immunoglobulin and T-cell receptor genes. A defining feature of NITR I domains is the presence of six cysteines that may influence the folding of this domain via intramolecular disulfide bonds. NITRs with a single Ig domain (V or I) have been reported. We previously identified I5 NITR sequences from gar that are encoded at two distinct genomic regions (Braasch et al., 2016). Here, we provide an in-depth analysis of these sequences and identify additional gar NITR sequences. In total, we have identified 17 gar ENSEMBL sequences that possess V and/or I domains that can be classified as NITRs (Fig 3). Thirteen sequences possess V and I domains, one lacks a V domain, and three lack an I domain: these observations suggest that the presence of a V and an I domain is an ancient feature of NITRs. The original description of teoleost NITRs noted the presence of sequences similar to joining (J) segments (EAAgEXExV1) that contribute to immunoglobulin (antibody) and T-cell receptor diversity via VDJ recombination (Strong et al., 1999). Teoleost NITR J segments are germline joined within the same exon as the V and I domains resulting in V-J or I-J segments without evidence for recombination. Most gar NITRs possess V-J and I-J sequences, suggesting that this sequence organization precedes the divergence of teoleosts from other ray-finned fish (Fig 3). In addition, gar NITR genes encode V-J and I-J domains in adjacent exons, which reflect the exon organization in zebrafish. This suggests that the organization of V-J-I-J in a single exon, which is observed in southern
SPOTTED GAR INNATE IMMUNE RECEPTORS

Leader | V domain
---|---

V domain | J | I domain

I domain | J | TM

TM | JTM

J. Exp. Zool. (Mol. Dev. Evol.)
pufferfish (Sphaeroides nolfei) NTRs and in some Japanese medaka and European sea bass (Dicentrarchus labrax) NTRs, is a derived feature in these species (Strong et al., 1999; Desai et al., 2008; Ferreiros et al., 2009).

Most telost NTRs encode a transmembrane domain, although some are secreted. Membrane bound NTRs can be classified as activating, inhibitory, or functionally ambiguous (Fig. 1; Yoder, 2009). Inhibitory NTRs encode ITIMs in their cytoplasmic tail. Activating NTRs possess an intramembrane charged residue that promotes the association with an activating adapter protein (Wei et al., 2009). Based on these features, 14 of the 55 NTR sequences are predicted to encode membrane proteins and, of these, ten are predicted to function as inhibitory NTRs, one is predicted to function as an activating NTR, and three lack these signaling motifs and are classified as functionally ambiguous (Fig. 3). Based on studies with an activating telost NTR (Wei et al., 2009), we predict that the putative activating NTR with a positively charged arginine in the transmembrane domain (ENSCOLOCG00000000529) will associate with and signal via the adapter protein, DAP12 (ENSCOLOCG00000003181). Three of these NTR sequences do not possess a transmembrane domain and are predicted to encode secreted proteins, although it is possible that there are errors in the ENSEMBL predictions.

In order to investigate possible conserved syntenies between the gar NTR loci and the human and zebrafish genomes, gar NTRs were placed in genomic context (Fig. 4). Twelve NTRs are clustered on LG14 and five are present on an unplaced scaffold JH919199.1. Only NTR sequences were identified on scaffold JH919199.1, providing no strategy for investigating conserved syntenies. In contrast, gar sequences positioned within or adjacent to NTR clusters on LG14 can be identified and include a single copy gene that encodes a secreted Ig domain and multiple G-coupled protein receptors including olfactory receptors. The secreted Ig domain encoded by this single copy gene (ENSCOLOCG00000010704) is more similar to members of the B7 family of proteins than Ig is to telost NTRs. BLASTp searches of the human and zebrafish protein databases using this sequence as a query identify B7-H4 (encoded by the V-set domain-containing T-cell activation inhibitor 1-like (VTCN1) gene) as the most similar protein in both human and zebrafish; however, the gar *vtcn1* gene has been placed on LG17, suggesting that this single copy gene encodes a gar-specific B7 protein. Orthologs of the olfactory receptors adjacent to the gar NTRs map to a region of human chromosome 13, which lacks sequences with identifiable similarity to NTRs. Although a number of olfactory receptor genes are linked to the T-cell receptor (TCR) alpha/delta locus, which undergoes V(D)J recombination, in human and mouse (Lance et al., 2002), the gar TCR alpha/delta locus is on LG24 (Braasch et al., 2016). Additional efforts to identify conserved syntenies between NTRs and the human genome were unsuccessful as Genomics and Ensembl indicate that NTRs have no mammalian homolog. Additionally, the genomic regions of gar that encode NTRs were compared to the genomic regions of teleosts that encode NTRs and no evidence for conserved syntenies was apparent (not shown). Despite the lack of identifiable conserved syntenies between NTR loci of gar and teleosts, their sequence conservation and presence of key NTR features provide compelling evidence that these regions of the gar genome encode homologs of teleost NTRs.

In order to examine the diversity of gar NTRs and compare them to teleost NTRs, one-to-one sequence comparisons were completed using 16 gar NTRs, 37 zebrafish NTRs, 44 Japanese medaka NTRs, and 24 southern pufferfish NTRs and represented as a heatmap (Fig. S). This analysis utilized NTR V domains as they are highly variable and amenable to phylogenetic analyses. NTRs within each teleost species have been classified into subgroups, with the V domains in each subgroup sharing >70% sequence identity. For example, the 39 described zebrafish NTR genes are grouped into 14 subfamilies (named NTR1 through NTR14) (Yoder et al., 2008). The observation that any single NTR gene within a select species is more likely to be more similar to any other NTR within the same species, than to any NTR from a different species, demonstrates that there are no one-to-one NTR orthologs between teleosts and suggests lineage-specific gene birth and death events (Desai et al., 2008; Ferreiros et al., 2009; Yoder, 2009). In addition, the presence of NTR subgroups likely reflects recent gene duplications that differ between species. In contrast, gar NTR genes appear to be evolving more slowly. Sixteen identified gar NTR V domains are all less than 70% identical to each other; however, the identity between a few pairs of gar NTRs approach 70% (Fig. S). Although it is clear that the gar NTR genes have not undergone...
a large degree of diversifying selection, it could be stated that
gar encodes three two-member subfamilies and 11 single copy
NITRs. However, we propose defining gar NITRs as single copy
genes based on the observation that the putative NITR gene
duplication events (generating possible subfamilies) in gar are more
ancient than those observed in teleost NITR subfamilies (Brausch
et al., 2016).

Diverse Immunoglobulin Domain-Containing Proteins. Defini-
tive DICPs have been identified in cyprinid fishes: zebrafish,
glass carp (Crenilabrus gilberti), and common carp (Cyprinus
carpio) (Rodriguez-Nuñez et al., 2016). A number of DICPs
share many similarities to members of the mammalian CD300
family including the same basic protein structure (Fig. 1) and the
ability to bind lipids (Haier et al., 2012). Each DICP possesses one
or more extracellular Ig domains and two types of Ig domains
have been described, D1 and D2. A defining feature of both Ig
domains is the presence of four cysteine residues that likely in-
fluence their folding via intramolecular disulfide bonds (Weis et
and Yoder, 2016). Secreted and membrane-bound DICPs have
been described from zebrafish. As described for NITRs (above),
membrane-bound DICPs include putative activating (identified
by an intramembranous charged residue), inhibitory (identi-
fied by cytoplasmic ITIM), and functionally ambiguous forms
(Fig. 1) (Weis et al., 2016). Although DICP-like sequences
have been identified in Atlantic salmon (Salmo salar), pufferfish
(T. rubripes and Tetraodon nigroviridis), and Nile tilapia (Ore-
ochromis niloticus), which belong to the Salmonidae, Tetraodon-
tidae, and Cichlidae families (Haier et al., 2012), it is unknown
if they share a common evolutionary origin. Recently, DICP-
like sequences were reported from coelacanth (Boudriot et al.,
2014).

Here, we identify two gar homologs of teleost DICPs on LG29
(ENSLOCG0000000000745) and on unplaced scaffold JH591438
(ENSLOCG000000000810). Both of these DICP sequences are
predicted to encode two Ig domains (D1 and D2) with no additional
identifiable features. Using the predicted protein sequences from
these gene models as BLASTn queries against the spotted gar
genome, we identified eight additional Ig domains on LG29
and scaffold JH591438 (Fig. 6). These gar DICP Ig domains
include six D1 domains and four D2 domains (Fig. 7). An align-
ment of gar DICP Ig domains with sequences from zebrafish,
carps, salmon, pufferfish, tilapia, and coelacanth reveals that each species, except coelacanth, possesses at least one DICP Ig domain with all conserved cysteines (Fig. 6). It will be of interest to determine if these gar and coelacanth genomic sequences encode inhibitory and activating DICPs as described in zebrafish.

The identification of DICP sequences in gar and coelacanth indicates that DICPs are more ancient than NTRs. Two partial DICP D2 domain sequences (Fig. 6) were identified within two unplaced scaffolds (JH:30400.1 and JH:30632.1) of the coelacanth genome (Boudriot et al., 2014a). Although the coelacanth D2 domains lack one of the four conserved cysteines, we identify evidence for conserved synteny that suggests that coelacanth, gar, and zebrafish DICP sequences share a genetic ancestry (Fig. 8). The zebrafish DICP gene cluster present on chromosome 3 is adjacent to MHC class I genes (Haire et al., 2012; Rodríguez-Núñez et al., 2016). Similarly, one of the coelacanth DICP sequences also is adjacent to MHC class I sequences (Fig. 8). Although neither of the gar DICP clusters is found to be linked to MHC class I sequences, this may be due to the incomplete assembly of the gar genome. For example, a gar MHC class I sequence was identified in the NCBI genomic Sequence Read Archive (SRX090800) that was omitted from the final genomic assembly (Grimholt et al., 2015). Indeed, Brausch et al. (2016) reported several gar MHC class I transcripts of the MHC class I Z

**J. Exp. Zool. (Mol. Dev. Evol.)**
and P lineages that did not map to the reference genome. These MHC class I lineages were previously thought to be restricted to teleost genomes (Bircher et al., 2014; Bircher and Yoder, 2014). Although DCP sequences are most similar to mammalian CD100 sequences, our analyses of the genomic region containing gar DCP sequences revealed no synteny with the human CD100 gene cluster on chromosome 17. Nevertheless, there is evidence linking gar DCPIPs to coelacanth DCPs—the DCP sequences on gar scaffold J159143.1 and one of the coelacanth DCP sequences are both linked to the ddx2 (dimethylaminohydrolase 2) gene (Fig. 6). Interestingly, DDAH2 is encoded within the human MHC locus on chromosome 6 (Flajnik and Kasahara, 2001). These observations support the model that zebrafish, gar, and coelacanth DCP sequences share an evolutionary origin within the proto-MHC.

CONCLUSIONS

Multiple teleost species are being employed as model species for human diseases (Schlacht, 2014); however, as seen here, the genomes of these species can be quite diverse from each other as well as from human. It is likely that the TGD combined with more recent lineage-specific expansions and diversification of a range of innate immune receptor gene families are responsible, at least in part, for these differences, especially in the molecular mediators of the immune system. Here, we have shown that although gar did not undergo the TGD, gar nonetheless possess a heightened repertoire of TLRs similar to teleosts. The gar genome also encodes NITRs and DCPs—two multigene families previously believed restricted to teleost species. This suggests that both NITRs and DCPs are characteristics of innate immunity that existed prior to the TGD. The evolutionary context of these gene families with respect to vertebrate evolution are summarized in Figure 9. As NITRs are absent in coelacanth, but present in gar and teleost, we have placed their birth after the split of ray-finned fishes from lobe-finned fishes. As DCPs are present in teleost, gar, and coelacanth, they have been placed in a more basal position, as shown in Figure 9. The “arms race” between host and pathogen as well as coevolution between commensal microorganisms and host has likely shaped the genomic repertoire of immune response genes. The environment inhabited by
Figure 7. Evolutionary relations of DCP D1 and D2 domains. The classification of the DCP sequences as D1 or D2 domains is supported by phylogenetic analysis of Ig domains shown in Figure 6. Sequence nomenclature is as in Figure 6 with the addition of zebrafish NitrA Ig domains (V and I) as outgroups. Bootstrap values are indicated next to the branches; values less than 50 are not shown.

Fish ensures intimate interactions with microorganisms and may explain the expansions of some of these immune gene families. IMR families often have a few activating receptors but many inhibitory receptors; the ectodomains are often highly similar, but amino acid sequence differences may have evolved to tolerate the presence of certain ligands (e.g., commensal bacteria). The sequencing of additional fish genomes will likely reshape our view on the evolutionary history of some of these genes. For instance, TLR22 has been identified in coelacanth, gar, and elephant shark (Wang et al., 2015), but not yet described in teleosts, therefore, we have TLR22 lost in teleost, but it may be identified in select teleost lineages at a later date. It should be kept in mind that sequence similarity does not always reflect functional similarity. For example, TLRs in Drosophila melanogaster and Caenorhabditis elegans play critical roles in embryonic development as well as immune function (Leclerc and Lemaître, 2008).

J. Exp. Zool. (Mol. Dev. Evol.)
Figure 8. DICP genes of zebrafish, spotted gar, and coelacanth are linked to genes associated with the MHC. Syntenic relationships of DICP-containing loci are displayed for (A) zebrafish, (B and C) spotted gar, and (D and E) coelacanth. Gene positions are intended only to give a sense of order and orientation and are not drawn to scale. DICP genes are shaded black. Genes with orthologs to the four paralogous human MHC loci (Flajnik and Kasahara, 2001) are shaded in gray. Novel genes or genes not known to be linked with MHC loci are white. (A) The linkage of zebrafish DICPs to MHC class I genes is well documented (Haire et al., 2012). (B–E) Genomic regions encoding gar and coelacanth DICP sequences are shown above with expanded views below. Exons from Ensembl-predicted DICP genes are shaded black, while additional DICP Ig domains are shaded gray, and drawn to scale. Dashed horizontal lines with a double slash represent gaps in genomic sequences introduced to reduce figure size. (D) Coelacanth MHC class I genes were previously described (Saha et al., 2014).
Therefore, functional assignment of orthologous genes may require further experimental investigations.

The gar genome has afforded a unique look into what may be an ancestral precursor of several immune gene families. Gar DICP and NTR genes have not undergone the large expansion and diversification observed in teleosts. In the case of immune genes, different gene families may substitute another gene family if they can function similarly. Perhaps the most famous example are mouse Ly49 and human KIRs. Therefore, it may be that novel mediators of immune function await discovery in gar. As genomic sequencing of additional genes and their functional annotation becomes available, we anticipate exciting revelations regarding the evolutionary processes driving gene gain/loss, gene expansion, and gene diversification. Importantly, understanding the similarities and differences between species is pinnacle to furthering our understanding of health and disease.

**ACKNOWLEDGMENTS**

The authors thank Ingo Branch, John Postlethwait, and all who contributed to sequencing the spotted gar genome. We thank Chris Amemiya, Ronda Litman, Nil Ratan Saha, and Michael Criscitiello for discussions of gar immune genes.

**LITERATURE CITED**


SPOTTED GAR INNATE IMMUNE RECEPTORS

Kortum AN, Rodriguez-Nunez I, Yang Y, et al. 2014. Differential expression and ligand binding indicate alternative functions for...
zebrafish polymeric immunoglobulin receptor (pigr) and a family of prig-like (pigr-like) proteins. microorganisms 66:267-279.


Kuraku S, Meyer A. 2010. Whole genome duplications and the radia-

Lane RP, Raorch JC, Lee SY, et al. 2002. Genomic analysis of the ob-

Letunic I, Doerks T, Bork P. 2015. SMART: recent updates, new develop-


Rauta PR, Samanta M, Dash HR, Nayak B, Das S. 2014. Toll-like rece-


Rodríguez-Nunez I, Weiss DJ, Litman GW, Yoder JA. 2014. Multigene families of immunoglobulin domain-containing innate immune receptors in zebrafish: deciphering the differences. Dev Comp Im-
munol 48:24-34.

Rodríguez-Nunez I, Weiss DJ, Litman RT, Litman GW, Yoder JA. 2016. The identification of additional zebrafish DCP genes reveals hap-
lo-type variation and linkage to MHC class I genes. Immunogenetics 68:295-312.


Stafford JL, Bengtén E, Du Posquier L, et al. 2006. A novel family of diversified immunoregulatory receptors in teleosts is homol-
ogous to both mammalian Fc receptors and molecules encoded within the leucocyte receptor complex. Immunogenetics 58:758-773.
SPOITTED GAR INNATE IMMUNE RECEPTORS


CHAPTER 5: SUMMARY AND FUTURE DIRECTIONS

During the course of my dissertation, I had the opportunity to analyze numerous vertebrate genomes in search of select innate immune receptors, in particular immunoglobulin domain-containing innate immune receptors (IIIRs) and Toll-like receptors (TLRs). While TLRs are known to be ubiquitous from invertebrates to vertebrates, many species-specific adaptations occur (Quiniou, Boudinot, and Bengtén 2013). Therefore, I cataloged the TLR genes in spotted gar, which was the first Actinopterygii (but outgroup to the teleost genome duplication) to be sequenced (Braasch et al. 2016). The IIIR gene families I researched included the teleost specific-novel immune-type receptors (NITRs), diverse immunoglobulin-like proteins (DICPs) and novel immunoglobulin-like transcripts (NILTs) (reviewed in Appendices A and B). The species distribution, sequence diversity and function were largely unknown for these gene families. My research overturned previous assumptions that the NITR and DICP gene families were restricted to teleost species. While there is a great deal of research regarding mammalian receptors with similar protein structures, it was unclear if orthology to these fish IIIRs exists. Identifying orthology between humans and mammalian models (e.g., mouse) to teleost model systems (e.g., zebrafish) has great potential to inform function of highly diversified fish IIIRs and the results of these efforts are summarized below.

Diverse immunoglobulin domain-containing proteins (DICPs)

I identified DICP genes in the genomes of zebrafish (Danio rerio), common carp (Cyprinus carpio) and grass carp (Ctenopharyngodon idellus). While I searched all of the currently available genomes in Ensembl’s database, most DICP-like sequences I identified did not contain significant synteny and sequence homology to known DICPs. I confirmed the report that DICP-like sequences found in coelacanth were DICPs and linked to MHC molecules. This finding is very intriguing as it suggests that DICPs are a much more ancient gene family than previously thought, increasing the age by about 450 million years (based on the most recent common ancestor of coelacanth (Latimeria chalumnae) and teleost species (Braasch et al. 2016)). Given how long the linkage of DICPs to MHC
molecules has survived, it suggests a possible interaction between DICP and MHC proteins, which is reminiscent of mammalian KIRs and MHC (Stewart et al. 2005). The conservation of these ancient genes suggests they may play a critical role in an individual's immune response and are likely present in in additional teleost species.

Additionally, I aided in the identification of the unprecedented level of variation between DICP genes of different lines of zebrafish. Numerous DICP transcripts were identified that did not match the reference genome, highlighting a need for complementary sequencing of matching zebrafish genomes.

**Novel immunoglobulin-like transcripts (NILTs)**

Since the annotation of Atlantic salmon NILTs on a bacterial artificial chromosome (BAC) was performed in 2008 by Østergaard et al, a reference genome of the Atlantic salmon was published (Lien et al. 2016). I used this genome to place the BAC on chromosome 3 and identified genes providing evidence of conserved synteny between zebrafish and Atlantic salmon NILT loci. Two NILT genes were previously predicted to be encoded on zebrafish chromosome 1 by Stet et al (2005). Within this genomic region of zebrafish chromosome 1, I identified 105 immunoglobulin-like domains with characteristics of NILTs described in other teleost species. Using a combination of next-generation sequencing and 5'/3' RACE, I annotated transcripts corresponding to 25 NILT genes encoded in this region. Comparatively, the zebrafish NILT repertoire contains many more members than Atlantic salmon. Two alternative haplotypes for NILTs were defined from the Zv9 version of the zebrafish reference genome.

**Lessons from spotted gar**

Continued use of the spotted gar genome (and other non-teleost Actinopterygii) has great potential to aid translation studies utilizing zebrafish to model human diseases. My comparison of spotted gar TLRs highlights many differences and we suggest many hypotheses, the testing of which...
will undoubtedly highlight numerous important features of vertebrate immunity. The expanded repertoire of TLRs in spotted gar (17 in spotted gar compared to 10 in humans) overturns the assumption that the teleost whole genome duplication is responsible for so many fish specific TLRs (Palti et al. 2010). For example, spotted gar TLR repertoire includes a TLR9 duplication that is absent from humans as well as a single copy of TLR21. These TLRs recognize CpG DNA originating from bacteria, suggesting that spotted gar may have a more robust immune response to certain types of bacteria (Yeh et al. 2013).

Additionally, NITRs were identified in spotted gar, the first identification of NITRs in a species outside of teleost. Fifteen NITRs were found with less sequence diversity than in zebrafish and no apparent subfamily structures. This could mean that the TGD aided in the expansion of IIIRs genes and/or the evolutionary histories of teleost versus spotted gar necessitated the expansion and divergence of NITRs. Understanding the role of NITRs in spotted gar may be easier than in zebrafish given that they are more conserved. Sequencing of related species (e.g., longnose gar, bowfin) may reveal many more important features. Comparing the genomes of these ancient species may reveal sites under positive/negative selection, helping to identify which sites are important for function.

**Implications and future directions**

A significant portion of my research focused on describing the amount of gene content variation found in IIIR families; however we still do not know the importance of the various haplotypes or the basic function of these genes. The multiple haplotypes of immune gene families described here are ideal targets for molecular marker development and for investigating their linkage to disease resistance/susceptibility. The identification of molecular markers associated with immune health and their application to improve breeding programs is a major focus both in the aquaculture industry and in conservation efforts for endangered species (Shen et al. 2016; Stone 1991; Guimaraes et al. 2007). For example, the recovery plan for the Wyoming Toad (Anaxyrus baxteri), which is extinct in the wild, calls for a genetically-informed breeding program in order to maximum genetic
diversity and the natural resistance to chytrid, a fungal pathogen partially responsible for the extinction of this species (Stone 1991). In contrast, the genetic selection of Nile tilapia (*Oreochromis niloticus*) has focused on marketable traits (e.g. size and growth) and not disease resistance putting these farm-raised fish at risk (Ponzoni et al. 2011; Li et al. 2015). Concerns regarding excessive application of antibiotics (and resultant antibiotic resistance) have increased interest in breeding programs which select for disease resistance (Gjedrem and Robinson 2014).

As the cost to resequence a genome decreases and the technology improves, the discovery of many more IIIR gene-content haplotypes should be anticipated. Already, very promising technological developments have occurred, including linked-read library preparation for next-gen sequencing (NGS) by 10X Genomics (Pleasanton, CA), and 3rd-gen sequencing technologies by PacBio (Menlo Park, CA) and Oxford Nanopore Technologies (Oxford, UK). Utilizing these technologies will allow for the accurate assembly of large genomic regions, providing a critical dataset for testing models of genome evolution. It is not unprecedented that members of multigene families, such as the IIIR genes described here, rapidly undergo a process of evolution called “gene birth and death” (Nei and Rooney 2005). This model expands on the previous model called concerted evolution and explains why members of a multigene family are more closely related to paralogous members (genes encoded on the same loci) than they are to the same gene in another species (an ortholog). It will be intriguing to test the hypothesis that multigene families encoded in tight clusters on a chromosome (e.g., IIIRs described here) evolve more quickly than dispersed multigene families (e.g., TLRs). Additionally, experiments that seek to identify the function of these genes will be highly informative. Here, I go over a few experiments that may be able to shed light on the function of IIIRs.
**Ligand Search**

The ligand(s) recognized by cell surface receptors such as IIIRs are notoriously difficult to ascertain. For example, the ligand for NKp44, a gene known for almost 15 years to be expressed on the cell surface of human natural killer cells (NK cells), was not found until 2013 (Rajagopalan and Long 2013; Baychelier et al. 2013; Vitale et al. 1998). Baychelier et al (2013) utilized a modified yeast-two-hybrid method that expresses transmembrane bound proteins on the surface of yeast. Yeast-two-hybrid methods utilize a recombinant protein as bait, in this case the extracellular domain of NKp44 and screens potential ligands against a library of proteins translated from pools of mRNA. This approach has great potential to be utilized to find the ligands for IIIRs. However, the method requires that the ligand be transcribed and translated by the selected cells and that the ligand does not originate from a foreign source (e.g., pathogens). Additionally, it requires that if the ligand is the product of an alternatively spliced transcript (as was the case for human NKp44), enough overlap exists to bind to the bait, or conditions to produce the alternative product are imitated.

**Reverse Genetics**

Insight into the function of genes of interest can sometimes be quickly ascertained by disrupting their function (e.g., knock-out/knock-downs) and observing a phenotypic effect (reverse genetics). Historically in zebrafish, this was done by knocking down transcripts with antisense morpholinos (Bill et al. 2009; Nasevicius and Ekker 2000). However, this approach requires that the genes of interest be expressed and functionally relevant early in development as morpholinos are injected at the one-cell stage and only persist for a few days. In order to ascertain function of genes in adult zebrafish, stable gene knockouts using TALENs and ZFNs have been successful (Huang et al. 2011; Doyon et al. 2008). However, these methods suffer from difficulty in designing and assembling reagents to target a specific gene. Recently, CRISPRs have been adapted to knockout genes in zebrafish with large success (Hwang et al. 2013). To provide a model to investigate the role of the
activating NILT in the immune system, I generated a stable line of CRISPR-induced gene knockout (Appendix F).

Once a homozygous knockout line is established, phenotypic effects can be investigated. If there are no detected developmental issues, adults with this genotype could be exposed to a variety of immune assays. Well established protocols include (1) tumor induction, (2) exposure to bacterial pathogens and (3) exposure to viral pathogens (Phelan et al. 2005; Pressley et al. 2005; Taylor and Zon 2009).

**Concluding thoughts**

Zebras provide an excellent model to quickly develop assays and test the effects of gene functions, mutations and knockouts. Additionally, the readily available gene-content haplotypes found in these immune gene families will provide excellent data to test theories of evolution. While the TGD has obfuscated many human-zebrafish gene orthologies, recent genome sequencing of outgroups to the TGD provide a clearer picture. The association of disease susceptibilities with the identified gene-content haplotypes will be a significant resource to guide breeding strategies in both aquaculture and conservation efforts. As functional studies of IIIR genes are performed from zebrafish, spotted gar and other models, our understanding of the evolution of the immune system and how to manage disease will undoubtedly improve greatly.
References


Shen, Yubang, Keyi Ma, Feng Liu, and Gen Hua Yue. 2016. “Characterization of Two Novel gadd45a Genes in Hybrid Tilapia and Their Responses to the Infection of Streptococcus Agalactiae.” *Fish & Shellfish Immunology* 54 (July): 276–81.


APPENDIX A: MULTIGENE FAMILIES OF IMMUNOGLOBULIN DOMAIN-CONTAINING INNATE IMMUNE RECEPTORS IN ZEBRAFISH: DECIPHERING THE DIFFERENCES

Developmental and Comparative Immunology 46 (2014) 24–34

Contents lists available at ScienceDirect

Developmental and Comparative Immunology

Journal homepage: www.elsevier.com/locate/dci

Review

Multigene families of immunoglobulin domain-containing innate immune receptors in zebrafish: Deciphering the differences

Iván Rodríguez-Nunez a, Dustin J. Wcisel a, Gary W. Litman b,c,d, Jeffrey A. Yoder e,a

a Department of Molecular Biomedical Sciences and Center for Comparative Medicine and Translational Research, North Carolina State University, 1060 Willow Moore Drive, Raleigh, NC 27695, USA
b Department of Pediatrics, University of South Florida College of Medicine, USF-CH (Children’s) Research Institute, 1417 7th Avenue South, St. Petersburg, FL 33701, USA
c Department of Molecular Genetics, All Children’s Hospital, 601 6th Avenue South, St. Petersburg, FL 33701, USA
d H. Lee Moffitt Cancer Center and Research Institute, 12902 Magnolia Drive, Tampa, FL 33612, USA

ARTICLE INFO

Article history:
Available online 15 February 2014

Keywords:
Novel immune-type receptors
Diverse immunoglobulin domain-containing proteins
Polymeric immunoglobulin receptor-like proteins
Novel immunoglobulin-like transcripts
Leucine-rich immune-type receptors

ABSTRACT

Five large multigene families encoding innate-type immune receptors that are comprised of immunoglobulin domains have been identified in zebrafish, which four do not possess definable mammalian orthologs. The members of some of the multigene families exhibit unusually extensive patterns of divergence and the individual family members demonstrate marked variation in interspecific comparisons. As a group, the gene families reveal striking differences in domain type and content, mechanisms of intracellular signaling, basic structural features, haplotype and allelic variation and ligand binding. The potential functional roles of these innate immune receptors, their relationships to immune genes in higher vertebrate species and the basis for their adaptive evolution are of broad interest. Ongoing investigations are expected to provide new insight into alternative mechanisms of immunity.

© 2014 Elsevier Ltd. All rights reserved.

Contents

1. Introduction ................................................................. 24
2. Immunoglobulin (Ig) domains .......................................... 25
3. Inhibitory and activating receptors .................................... 25
4. Families of Ig domains containing innate immune receptors in zebrafish ................................................................. 25
  4.1. Novel immune-type receptors (NETRs) ......................... 26
  4.2. Diverse immunoglobulin domain-containing proteins (DIPs) ......................................................................................... 26
  4.3. Polymeric immunoglobulin receptor-like proteins (PIRs) ............................................................................................. 28
  4.4. Novel immunoglobulin-like transcripts (NILTs) .......... 29
  4.5. Leucine-rich immune-type receptors (LIRRs) .......... 30
5. Defining the Ig domain differences ...................................... 30
6. Evidence for haplotype variation ......................................... 33
7. Conclusions .................................................................... 33
Acknowledgements .............................................................. 33
References ........................................................................ 33

1. Introduction

Although many effectors of innate immunity have experienced relatively minimal change over extended periods of evolutionary time, others, many of which are encoded in large multigene families, exhibit striking degrees of variation in both form and function. In mammals, natural killer (NK) cell function is mediated through the killer-cell immunoglobulin-like receptor (KIR), which are encoded in a diversified immunoglobulin gene family in man; however, in mouse, equivalent function is effected through lectins encoded in the Ly49 multigene family. Additional molecules
encoded in multigene families mediate immune effector function in mammals and include: CD3D and CD3G-like proteins, triggering receptors expressed in myeloid cells (TREM's and TREM-like proteins), stialic acid binding proteins, Ig-like lectins (SIGLEC's) and the natural cytotoxicity triggering receptors (NCR's). The gene families encoding these molecules have been well defined genetically; however, their functions and particularly the scope of their interactions with other components of the immune system are less well understood. As a group, these molecules exhibit significant interspecific differences as evidenced, in some cases, by the lack of recongnizable orthologs between mouse and man. Although it is recognized that multiple gene birth and death events account for the rapid evolution and functional variation of this large and diverse group of molecules, the nature of the genomic structures and selective pressure events that drive the process are not at all clear.

Several years ago, as part of an investigation of alternative forms of immune function, we initiated studies to examine the broad question of innate immune receptor diversity in the members of taxonomic classes outside of the mammals. These efforts primarily have been focused on the ostechyths, specifically telost fish, which represent the largest subclass of animals and possess particularly remarkable morphological variation as well as physiological diversity that reflect the extensive variation in their habitats and responses to selective pressures. These studies, which are based primarily on the recognition of genetic homologies, were initiated with a search for diversified receptors that contain variable-like (V) multigene immunoglobulin domains and later were extended to include multigene families encoding molecules possessing other types of domains. In the course of these investigations, some of the gene families that have been identified have shown to represent orthologs of immune effector genes in mammals, whereas others are unique to bony fish. Taken together with studies from other laboratories as well as extensive work in mammals, a picture has emerged that reveals particularly complex patterns of dynamic genetic change that likely are driven by selective pressures and host adaptations that are unique to major animal groups and overall reflect the enormous complexity and extensive system of immune mediators.

2. Immunoglobulin (Ig) domains

An individual Ig domain is approximately 100 amino acids in length and forms an "ig-fold" structure consisting of two anti-parallel β-sheets packed face to face. Four sets of Ig domains have been described: variable-like domains (V), constant-like domains (C1 and C2) and intermediate domains (I) (Harrop and Choithia, 1994). Although Ig domains vary in the number and size of strands in the β-sheets and in the size and conformation of the links between the strands, one β-sheet is anchored by strands A, B and E while strands G, F, and C form the center of the second β-sheet (Barclay, 2003; Harrop and Choithia, 1994). For example, most members of the V set possess ten strands forming two β-sheets with strands A, B, E and D forming one β-sheet and strands A, G, F, C, and C' forming the other. In contrast, most members of the C1 set are smaller with one β-sheet formed by strands A, B, E, D and O and the other by strands G, F, C and C (Harrop and Choithia, 1994).

Two highly conserved cysteines, C61 and C64 (numbering based on the IMGT system; Lefrancois et al., 2009) form an intrachain disulfide bond between the B and F strands stabilizing Ig-fold conformation. The inter-cysteine distance in V domains ranges from 65 to 75 residues and is appreciably shorter in C1 and C2 domains (55-60 residues) (Williams and Barclay, 1988). Intermediate (I-type) Ig domains possess structural features of V domains but exhibit shorter inter-cysteine distances (Harrop and Choithia, 1994). As discussed below, additional cysteines within Ig domains also can contribute to intrachain disulfide bridges.

3. Inhibitory and activating receptors

Many mammalian innate immune receptor families include receptors with opposing functions. Engagement of ligands by some receptors activates the immune cell, whereas other receptors inhibit this process. The KIR, leukocyte immunoglobulin-like receptor (LIR), CD300, TREM and SIGLEC multigene families encode Ig domain-containing transmembrane proteins that include both activating and inhibitory receptors.

Some activating receptors possess a short cytoplasmic tail and a charged residue within the transmembrane domain which permits partnering with and signaling through an adaptor protein, such as DAP12, DAP10, CD11; and FcγRII. These adaptor proteins possess a cytoplasmatic activation motif such as an immunoreceptor tyrosine-based activation motif [ITAM: (D/E)(Y)(X)(Y)(X)(Y)(X)(L/U)]. Other activating receptors possess a cytoplasmatic tail that includes an ITAM (e.g. select FcR- Like Molecules; Ehrhardt and Cooper, 2011).

Inhibitory receptors typically possess one or more immunoreceptor tyrosine-based inhibition motif [ITIM: (I/V)XX[L/I]XX[L/I]XX[L/I]XX[L/I]]. Within their cytoplasmic tail and do not utilize an adaptor protein (Zsebo and Lanier, 2000; Vey and Vivier, 1997). For example, ligand engagement by activating KIR-DAP12 complexes on natural killer cells results in the phosphorylation of the DAP12 ITAM through JAK activation. These ITIMs recruit Shc family phosphotyrosine-binding adapter proteins to inhibit this signaling cascade. Shifts in the balance of these opposing signals determine if a natural killer cell will attack or ignore a cell or pathogen.

Some receptors can function both in an activating or inhibitory manner depending on the cellular context. A cytoplasmic immunoreceptor tyrosine-based switch motif [ITSM: TXXXXX(V/I)] is found in some of these receptors (Ochrochloch and U. 2000; Shlapatska et al., 2001). Multigene families that encode activating and inhibitory receptors and their associated adaptor proteins have been identified from the genomes of various teleost species including zebrafish (Danio rerio) (Haire et al., 2012; Montgomery et al., 2011; Stafford et al., 2006; Stett et al., 2000; Yoder et al., 2007; Yoder, 2009).

4. Families of Ig domain-containing innate immune receptors in zebrafish

The genomes of teleost species, including zebrafish, encode hundreds to thousands of Ig domains. Many of these domains belong to the immunoglobulin (e.g. IgM, IgG, IgZ) and T cell receptor genes that undergo V(D)J recombination as part of the adaptive immune system (Hikima et al., 2011), whereas other Ig domains belong to genes encoding cell adhesion molecules, growth factor receptors (e.g. VEGFRs and FGFRs) and innate immune receptors.

The best described Ig domain-containing innate immune receptor families in teleosts are the novel immune-type receptors (NITRs), leukocyte immunoglobulin-like receptors (LIRs), novel immunoglobulin-like transcripts (NILTs), diverse immunoglobulin domain-containing proteins (DIDPs), and polymorphic immunoglobulin receptor-like proteins (PGRls); we have described the zebrafish NITRs, DICPs and PIGRLs (Haire et al., 2012; Kortum et al., 2014; Yoder et al., 2001, 2004, 2008, 2010). NILTs and LIRs have been characterized in other fish species, but can be identified in the zeb-
The roles of secreted variants of these proteins remain to be elucidated; however, in mammals, secreted TREM-1 has been shown to compete with membrane bound TREM-1 for ligand binding often dampening the immune response (Dervie et al., 2010; Gibot et al., 2004).

4.1. Novel immune-type receptors (NITRs)

The NITR genes were identified initially as V domain-containing transmembrane receptors from the compact genome of the Southern pufferfish (*Spheroidees niphobles*): 26 NITR genes were identified from a single genomic clone of ~100 kbp (Rast et al., 1995; Strong et al., 1999). NITRs subsequently were identified in two gene clusters on zebrafish chromosomes 7 and 14 (Yoder et al., 2001, 2004, 2008). To date, 36 NITR genes have been described on zebrafish chromosome 7 and three on chromosome 14. NITRs possess one or two Ig domains. For NITRs with two Ig domains, the membrane distal Ig domain is of the variable (V) type and the membrane proximal Ig domain is of the intermediate (I) type. Although all NITR genes encode a V domain, some NITR genes lack an I domain that results in proteins with a single V domain. NITR proteins that lack a V domain are generated by alternative mRNA splicing (Shah et al., 2012). Based on sequence similarities between V domains, NITR form 14 groups, which are designated numerically (within an individual species) based on the order of their discovery, e.g., the NITR1, NITR2, NITR3 and NITR4 groups were the first to be described; the numbering does not reflect interspecific sequence homology. Individual gene names include a letter indicating the order of their description, e.g., nitr3b was the second gene described from group 3. A pairwise sequence comparison demonstrates a higher level of sequence variability between the NITR V domains as compared to the more conserved I domains (Fig. 1). In addition, the 15-mem-

![Image of heatmap](image-url)

**Fig. 1.** Heat map of pairwise genetic distances between zebrafish NITR, DCP, PIGR, NILT, and LTR Ig domains. The entire dataset (418 Ig domains) as well as two pIGR Ig domains (included with the PIGR) were aligned using default settings of Clustal Omega (Stevens et al., 2011). Poisson corrected distances were calculated using Meta-PCA (Hofackers and Middendorf, 2010). The heat map was generated using Microsoft Excel 2013. Minimum and maximum distances were 0.4423 and 7.072, respectively.
The zebrafish NITR family includes a single gene that encodes three activating receptor isoforms, multiple genes that encode inhibitory receptors possessing ITIM and/or ITIM-related (itim) sequences, and genes that encode receptors with no identifiable signalling motifs and secreted proteins (Fig. 2; Yoder, 2009). In the context of human NK cells, engagement of the ITIM-containing...
NitrA [previously Nitr3.1] has been shown to down regulate MAPK in an ITIM-dependent manner (Yoder et al., 2001). The single activating gene, nitrA, was described originally as encoding one V domain, one I domain and a positively charged residue within the transmembrane domain. After the discovery of mRNA splice variants that partially or completely lack the exon encoding the V domain, the original isoform was renamed Nitr9-long (or Nitr9L), whereas the proteins encoded by the splice variants were designated Nitr9-short (Nitr9S) and Nitr9-super short (Nitr9SS) (Wei et al., 2007). Within the context of mammalian cell culture, Nitr9L partners with and signals through Dap12 via pairing of reciprocally charged residues in their transmembrane domains (Wei et al., 2007; Yoder et al., 2004).

PCR-based expression studies reveal that transcripts from all NITRs can be detected from zebrafish lymphocytes but not from myeloid cells (Yoder et al., 2010). Studies employing activating NITRs from channel catfish (Ictalurus punctatus) and NK-like catfish cell lines indicate that NITRs are involved in allorecognition (Cannon et al., 2008); however, specific molecular ligands have not been identified. These observations and signaling similarities to mammalian KIRs led us to hypothesize that NITRs function as NK receptors in teleosts (Yoder and Litman, 2011).

NITR genes or transcripts have been identified in most major lineages of teleosts including tetradontiforms (putterfish), perciforms (stickleback) ovulartaria (medial), gadiforms (cod), salmoniforms (trout), siluriforms (catfish), cypriniforms (zebrafish), cichliformes (lake whitefish), as well as in a Holostean fish (gar); however, NITRs have not been identified in cartilaginous fish or other vertebrate lineages (Yoder, 2009). Comparisons of NITR Ig domain sequences between fish species indicate that NITR gene clusters expand in species-specific manners. Different numbers of NITR genes are found in different species and any given NITR within a species is more related to other NITRs within that species than to NITRs in a different species (Yoder, 2009). Despite the species-specific differences in NITR sequences and gene number, NITRs likely are present in all teleosts.

4.2. Diverse immunoglobulin domain containing proteins (DCPs)

Efforts to identify NITRs from cartilaginous fish resulted in the identification of the Ig domain-containing modular domain immunoglobulin-type receptor (MDIR) gene family from needlefish (Rajula eglaetana) (Cannon et al., 2006). Using the MDIRs as search queries, three gene clusters were identified in the zebrafish reference genome (Zv8) that are distinct from NITRs and designated DCPs (Cannon et al., 2006; Haire et al., 2012). DCPs possess one or two types of Ig domains referred to as D1 and D2. All DCP genes encode at least one D1 domain which can be divided into three groups based on sequence similarity and chromosomal location. This relationship is reflected in their gene nomenclature. The DCP1, DCP2 and DCP3 groups are encoded in clusters on chromosomes 3, 14 and 16, respectively, and individual genes are named in the order that their D1 domains were identified (e.g. dcp1.2 is encoded by the DCP1 group and encodes the second D1 domain identified in this cluster). Some DCP genes encode a single D1 domain, whereas others encode a D1 domain and membrane proximal D2 domain, although differential mRNA splicing can produce DCPs with a single D2 domain (Haire et al., 2012). As observed in Fig. 1, the DCP1 D1 and D2 domains each are highly conserved in sequence and share little sequence similarity to Ig domains from other families.

Multiple DCP transcripts are predicted to encode a single putative activating receptor, multiple receptors with two or three cytoplasmic ITIM or ITIM sequences, membrane bound receptors with no identifiable signaling motifs and secreted forms (Fig. 2), reminiscent in part of the structural variation observed in NITRs. The single activating receptor, Dicp1.2 possesses a negatively charged residue within the transmembrane domain and is predicted to partner with and signal through an adaptor protein as observed for mammalian CLM-5 (CDSKLD) (Haire et al., 2012). Thus, the zebrafish DCP family is predicted to include both inhibitory and activating forms. Expression patterns of DCPs have not yet been reported.

Recombinant DCP Ig domain were shown to bind various phospholipids and lipid extracts from various bacteria in an ELISA-based assay (Haire et al., 2012). Of the seven DCP Ig domains examined, it was shown that the D1 domain of Dicp1.5 bound phospholipids with the broadest specificity whereas the D1 domain of Dicp1.4 displayed a narrow specificity for binding extracts from Mycobacterium. Although the ability to bind phospholipids is conserved between the DCP1 and DCP2 families (e.g. the D2 domain of Dicp1.5 binds with broad specificity), no lipid binding was detected for the D1 domain of Dicp3.1. The capacity to bind phospholipids is shared with the mammalian CD300 and TREM family of receptors and suggests they may share a common or at least related function (Cannon et al., 2012; Choi et al., 2011).

4.3. Polymeric immunoglobulin receptor-like proteins (PIGRs)

In mammals, the polymeric immunoglobulin receptor (pIgR) is an integral transmembrane glycoprotein that transports soluble polymeric Iggs such as pIgA across mucosal epithelial cells. For example, pIgA is expressed by plasma cells present in the lamina propria underlying the intestinal epithelium and is bound by pIgR on the basolateral surface of epithelial cells, pIgR transports pIgA through the epithelial cell by transcytosis (Asano and Komiyama, 2011; Kaetzel, 2005). The transport of pIgA by pIgR to the intestinal lumen is essential for protecting the host from invading pathogens and maintaining homeostasis (Johnson et al., 1995). The mammalian pIgR is encoded by a single copy gene (PIGR) that encodes five extracellular Ig domains (D1–D5) (Asano and Komiyama, 2011; Kaetzel, 2005). Alternatively spliced PIGR transcripts lacking the D2 and D3 domains have been reported in rabbit and cow (Deticher and Monnot, 1986; Kusnerth et al., 1995). Chicken and Xenopus pIgR possesses four Ig domains; the D2 domain found in mammals is absent in these latter species (Braathen et al., 2007; Wiesland et al., 2004). Full-length transcripts encoding pIgR homologs, which possess two Ig domains, have been identified in multiple fish species including zebrasfish, fugu (Takifugu rubripes), grouper (Epinephorus coioides), common carp (Cyprinus carpio), Atlantic salmon (Salmo salar) and rainbow trout (Oncorhynchus mykiss) (Feng et al., 2005; Hamuro et al., 2007; Kortum et al., 2014; Rombout et al., 2008; Tadiso et al., 2011; Zhang et al., 2010). Teleost pIgR is expressed by lymphoid organs including mucosal tissues (intestinal, skin and gill) and has been shown to bind both IgM and IgG/IgA/IgE (Feng et al., 2005; Hamuro et al., 2007; Rombout et al., 2008; Tadiso et al., 2011; Zhang et al., 2010). pIgR-like (PIGRl) transcripts were identified originally from Atlantic salmon and common carp (Cannon et al., 2006; Ribeiro et al., 2011; Tadiso et al., 2011). The carp PIGRl protein was shown to be expressed abundantly in macrophages and is secreted upon immune stimulation (referred to as a soluble immune-type receptor, SIR, in: Ribeiro et al., 2011). Recently a PIGRl multigene family was described adjacent to the single copy zebrafish pIgR gene on chromosome 2 (Kortum et al., 2014). Twenty-nine distinct PIGRl genes were identified and transcripts corresponding to ten of these genes encode membrane-bound receptors with two cytoplasmic ITIM or ITIM sequences or secreted forms (Fig. 2). Activating PIGRl receptors have not yet been identified. All PIGRl transcripts encode two Ig domains, D1 and D2. Although PIGRl D2 domains are highly similar in sequence, phylogenetic analyses of the PIGRl D1 domains define
four groups of PIGRL proteins (PIGR1-PIGR4). Individual gene names include the group number and a second number indicates the order in which they were identified (e.g. pigrf2 is the second gene identified in PIGRL group 3) (Kortum et al., 2014).

Zebrafish PIGRL genes are differentially expressed in lymphoid and myeloid cells. PIGRL1 and transcripts are predominantly detected in lymphoid cells whereas PIGRL2 transcripts are predominantly detected in myeloid cells. PIGRL3 and PIGRL4 transcripts are detected from both leukocyte lineages. An examination of PIGRL gene expression in adult tissues after bacterial or viral infection produced contradictory outcomes: bacterial infection led to a generalized increase in PIGRL transcript levels, whereas viral infection led to a generalized decrease in transcript levels of PIGRLs. Although there are a variety of mechanisms that may cause these different responses, one possibility is that PIGRL expression is downregulated through virus-induced immune suppression (Kortum et al., 2014).

Recombinant PIGRL1 Ig domains have been shown to bind phospholipids in a manner similar to DCs (see above). The Ig domains expressed from PIGRL2, PIGRL3 and PIGRL4 proteins did not show lipop binding and their ligands remain unknown (Kortum et al., 2014).

4.4. Novel immunoglobulin-like transcripts (NILTs)

NILTs have been described from carp, trout and salmon and are encoded on zebrafish chromosome 1 (Kock and Fischer, 2008; Ostergaard et al., 2009, 2010; Stet et al., 2005). Individual NILTs encode one or two Ig domains, which exhibit sequence and structural similarity to human TREM/NKp44, human CD300 and Xenopus pigr Ig domains (Kock and Fischer, 2008; Stet et al., 2005). NILTs encode type I transmembrane receptors that possess cytoplasmic ITIMs or ITAMs as well as secreted isoforms. NILTs are encoded by a multi-gene family as evidenced by Southern blot analyses in carp and sequencing of six NILT genes from a single salmon genomic (BAC) clone (Ostergaard et al., 2010; Stet et al., 2005). Although no reports have been published describing zebrafish NILT transcripts, nilt1 and nilt2 (GenBank: BN001234 and BN001235), respectively, are predicted to encode one and two Ig ectodomains and cytoplasmic ITIMs (Ostergaard et al., 2009; Stet et al., 2005). This is likely an under-representation of the NILT genes encoded by zebrafish as 104 NILT-related Ig domains can be identified on chromosome 1 of the current (2.09) zebrafish reference genome.

The two types of carp NILTs are: Cyro-NILT1 which encodes a single Ig domain and a cytoplasmic ITAM, and Cyro-NILT2 which encodes a single Ig domain and a cytoplasmic ITIMs (Fig. 3; Stet et al., 2005). Four types of NILTs have been described in trout. Omny-NILT1 undergoes alternative splicing to generate a membrane-bound isoform that possesses two Ig ectodomains, a cytoplasmic ITIM, a cytoplasmic ITIM and a cytoplasmic ITAM as well as a secreted isoform. Omny-NILT2 encodes a single Ig ectodomain and a cytoplasmic ITAM. Omny-NILT3 encodes two Ig domains and three cytoplasmic ITIM/ITIM sequences. Omny-NILT4 encodes a type I transmembrane protein that possesses four cytoplasmic ITIM/ITIM sequences and undergoes alternative splicing resulting in isoforms that possess one or two Ig ectodomains (Fig. 3; Kock and Fischer, 2008; Ostergaard et al., 2009). There is evidence for more than nine NILT genes in salmon in which the nomenclature has named new sequences sequentially (e.g. Sino-NILT1 was the first described) and the structural variation of salmon NILTs reflects that observed in trout (Fig. 3). As RT-PCR amplicons from a single carp reveal 53 different NILT sequences (Stet et al., 2005), it is highly likely that many more NILTs remain to be described from all of these species and not all structural variants have been recognized. The genomic organization and structural variation of zebrafish NILTs remains to be described and annotated.

The expression of carp and trout NILTs has been described. The highest expression is seen in lymphoid tissues (kidney, spleen and thymus) and it is anticipated that zebrafish NILTs will display similar

---

Fig. 3. Predicted structures of teleost NILT and LIR proteins. Sequences have been reported (Montgomer et al., 2011; Ostergaard et al., 2009, 2010; Stet et al., 2005). Proteins are organized numerically: activating receptors are boxed. The Ig domains of (A) NILT (D1 and D2) and (B) LIR (D1, D2, D3, D4, D5 and D6) proteins are indicated. Cytoplasmic ITAMs, ITIMs, ITIM-like (itim) sequences and prolines as well as charged residues (e) within transmembrane domains are labeled. Proteins encoded by splice variants are indicated with asterisks (*). This is not intended to represent a complete catalog of proteins from these families as additional proteins sequences and structures will likely be identified in the future.
expression patterns (Rock and Fischer, 2008; Østergaard et al., 2009; Steet et al., 2005). Ligands for NITLs remain unknown.

4.5. Leukocyte immune-type receptors (LITRs)

LITRs are type 1 transmembrane proteins that possess two to six Ig ectodomains and have been best described from channel catfish (Fig. 3; reviewed by Montgomery et al., 2011). Catfish LITRs are encoded by a multi-gene family located at multiple loci and include inhibitory and activating forms (Fig. 3; Cortes et al., 2012; Montgomery et al., 2012; Stafford et al., 2006, 2007). LITR Ig domains share similarity to mammalian FcR, FcγR-like, KIR, LR1 and Nkp46 sequences (Stafford et al., 2006). Although zebrafish LTR transcripts have not been identified, putative LTR genes from Zv8 have been described (Stafford et al., 2006). Conservatory estimates suggest at least 137 Ig domains from the current zebrafish reference genome (Zv9) belong to the LTR family which share 33–58% identity with catfish LITRs. This high level of sequence diversity between zebrafish and catfish LITRs confound efforts to predict the number and organization of individual zebrafish LTR genes based only on genomic Ig domain exons.

Expression analysis reveals catfish LTR transcripts to be most abundantly expressed in kidney, gill, spleen and PBLS, as compared to liver, muscle and intestine. LITR transcripts were detected in macrophage, B cell, cytotoxic T cell, and NK-like cell lines (Stafford et al., 2006). In addition, multiple LITR transcripts are increased dramatically in PBLS and cytotoxic T cell lines after exposure to aflatoxin (Stafford et al., 2007). Although ligands have not been identified for LITRs, some LITRs are predicted to bind MHC I (Stafford et al., 2007). Recombinant forms of ITAM-containing LITRs associate with SHP-1 and SHP-2 following treatment with the protein phosphatase inhibitor, pervanadate (Montgomery et al., 2002). In the context of transfected mammalian cells, surface expression of a recombinant activating LITR, which possesses a positively charged residue within the transmembrane domain, is increased by co-expression with the ITAM-containing adaptor proteins FlkY and Flkγ-like. Activating LITRs associate with these adaptor proteins as well as with CD3ε-like (Mewes et al., 2009). In addition, recombinant activating LITRs have been shown to form non-covalent homo- and hetero-dimers suggesting a complex mechanism for regulating LTR-mediated signaling.

The nomenclature of catfish LITRs has been based on the structures of the initial LITR transcripts designated LTR1 (which encodes an inhibitory receptor) and LTR2 and LTR3 (both of which are activating receptors) (Fig. 3; Stafford et al., 2006). Subsequent LITR transcripts have been named based on their similarity to the original transcripts (e.g. LTR1-like and LTR2-like) (Montgomery et al., 2011; Stafford et al., 2007). The complete genomic organization and structural variation of LITRs remains to be described and annotated in any fish species including zebrafish.

5. Defining the Ig domain differences

The zebrafish protein families discussed above can be distinguished from one another based on sequence variation (see Fig. 1) and Ig domain structure including the spacing of cysteine residues within Ig domains. The general characteristics of each family are summarized in Table 1. Whereas overall protein structures and signaling capabilities (via ITIMs/ITAMs or adaptor molecules) are conserved between protein families, many unique features distinguish each gene family, even in the absence of complete functional data. Given the remarkably large dataset now available, consisting of 418 different Ig domains, broad classifications can be assigned.

Within the Ig domains of each family, there is a large degree of amino acid sequence variation which is complicated by the observation that each gene family possesses multiple groups of similar

<table>
<thead>
<tr>
<th>Feature</th>
<th>NITR</th>
<th>DICP</th>
<th>PIRGL</th>
<th>NILT</th>
<th>LITR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min. estimate of domains</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Ig domains per molecule</td>
<td>1–2</td>
<td>1–2</td>
<td>1–2</td>
<td>1–2</td>
<td>1–2</td>
</tr>
<tr>
<td>Encoded by zebrafish chromosomes</td>
<td>7, 14</td>
<td>7, 14</td>
<td>7, 14</td>
<td>7, 14</td>
<td>7, 14</td>
</tr>
<tr>
<td>Members with cysteine-rich TM</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Members with charged residue in TM</td>
<td>Yes, +charge</td>
<td>Yes, +charge</td>
<td>Yes, +charge</td>
<td>Yes, +charge</td>
<td>Yes, +charge</td>
</tr>
<tr>
<td>Adaptor proteins</td>
<td>Dapt2</td>
<td>Dapt2</td>
<td>Dapt2</td>
<td>Dapt2</td>
<td>Dapt2</td>
</tr>
<tr>
<td>Members with cysteine-rich ITAM</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Ig types</td>
<td>V, I</td>
<td>V, I</td>
<td>V, I</td>
<td>V, I</td>
<td>V, I</td>
</tr>
<tr>
<td>C² to C² spacing</td>
<td>64-71</td>
<td>64-71</td>
<td>64-71</td>
<td>64-71</td>
<td>64-71</td>
</tr>
<tr>
<td>Notable features</td>
<td>Protein-protein interaction</td>
<td>Protein-protein interaction</td>
<td>Protein-protein interaction</td>
<td>Protein-protein interaction</td>
<td>Protein-protein interaction</td>
</tr>
<tr>
<td>Ligands</td>
<td>Unknown</td>
<td>Phospho-lipids</td>
<td>Phospho-lipids</td>
<td>Unknown</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

* Approximate numbers of NILT and LITR genes were calculated by dividing the total number of Ig domains by two (NILT) or six (LITR).
* Based on transcripts from catfish, carp, trout and salmon.
* NITR 1 domains contain six conserved cysteine residues and are excluded from this analysis.
* No all Ig domains possess an extra cysteine pair.
lg domains. In contrast to the NITR, DCP, and PIGR families which encode two very distinct types of lg domains, the NILT and LITR families possess at least three and six different types of lg domains, respectively (Haire et al., 2012; Kortum et al., 2014; Montgomery et al., 2011; Ostergaard et al., 2010; Yoder, 2009). Pairwise sequence comparisons reveal that homology between lg domains of the same type, for example NITR V domains, can be as low as 45% identical and 50% similar. Despite the relatively reduced level of sequence conservation, detailed comparisons reveal that lg domains group together by lg type within each gene family (Fig. 1), revealing signatures within the predicted structures of the lg domains that are unique to each receptor family.

A. NITR loci

![Diagram of NITR loci]

**Fig. 5.** Evidence for histotypic variation at the NITR, DCP, and PIGR loci. The regions of (A) chromosomes 7 and 14 that encode the NITR genes, (B) chromosome 3, 14 and 16 that encode the DCP genes and (C) chromosome 3 that encodes the PIGR genes are shown in detail. Graphics were adapted from the Ensembl Genome Browser, www.ensembl.org (homo. serr.). The NITR, DCP, and PIGR genes (red and orange) and relevant flanking genes (brown) are annotated above and below the genomic regions. The Database of Genomic Variants archive (DCV) provides the genomic structural variants: regions of copy number variation (CNV) are indicated in black, regions of gene loss are indicated in red, and regions of gene gain are indicated in blue.
The spacing between conserved pairs of cysteine residues is relatively unique among the gene families. As discussed above, the spacing between C\textsuperscript{C1C} and C\textsuperscript{H4C} can be employed to characterize an Ig domain at V, C1, C2, or I and are separated by 44–75 residues (Barclay, 2003; Harpaz and Cho, 1994; Smith and Xue, 1997).

When available, X-ray crystallography or protein structure modeling confirmed the classifications for these protein families (Cannon et al., 2003; Ostrov et al., 2003; Stet et al., 2003). Additional pairs of cysteine residues also appear to be unique among each gene family (Table 1). For example, NTR I domains possess four additional highly conserved cysteines: three that fall between C\textsuperscript{1C2} and C\textsuperscript{1H4} and one at position C\textsuperscript{1H5}, whereas NTR V domains do not possess additional conserved cysteines (Yoder, 2009). In DICP D1 and D2 domains, a wide spacing between two additional conserved cysteines (C\textsuperscript{D1C} and C\textsuperscript{D2C}) is observed. In the case of PIGRL D2 domains, two additional cysteines separated by six residues (C\textsuperscript{6C}) are seen, whereas only the PIGRL1, PIGRL3, and PIGRL4 D1 domains possess two additional cysteines (C\textsuperscript{11C}).
and C^9(P)^9(C^5(C^2)) spaced seven or nine residues apart (C^6(C) or C^6(C)). D1 domains of PIGRIL2 proteins do not possess these additional cysteines (Kortum et al., 2014). NITL Ig domains can be divided into at least three subgroups based on the spacing between cysteine present in the C and C' strands: C^6(C), C^6(C) and C^6(C) (Ostergaard et al., 2010). LITR Ig domains lack additional conserved cysteine pairs.

The Ig domain of MIF and NITL I domains share a 21 residue motif (I^1(V)/Y^1(V/L)X_5(X/Y/S)X_4) that includes C^8e as well as the residues that are conserved in classical V domains (Kock and Fischer, 2008); many PIGRL proteins also contain this sequence. This motif determines the E and F strands in opposing β-sheets of the Ig fold and may be important for defining a binding ligand pocket (Kock and Fischer, 2008).

Many of the gene families also contain unique sequences that do not share across the other families. Many NITR and I domains contain a sequence similar to joining (J) domains (FGKCXXLXCV) present in immunoglobulins and T-cell receptors after V(D)J recombination (Litman et al., 2001; Yoder, 2009). PIGRL proteins share a conserved IPCXX motif at the C^6 position with pIgR molecules (Kortum et al., 2014). D1C D2 domains typically contain a poly-serine stretch in the amino-terminus which is caused by a triplet repeat expansion at the beginning of this exon (Haire et al., 2012).

6. Evidence for haplotype variation

The human KIR gene cluster displays polymorphic and copy number variation between individuals resulting in over one hundred known KIR haplotypes (Middleton and Gonzalez, 2010). Fig. 4 demonstrates that the NITR, D1C, PIGRL, NITL and LITR gene clusters are located in chromosomal regions with high levels of copy number variation. In addition, there is growing evidence that the polymorphic NITR, D1C, PIGRL and KIR gene clusters display multiple haplotypes that encode different genes (Haire et al., 2012; Kortum et al., 2014; Yoder et al., 2004, 2008). A comparison of the NITL, D1C and PIGRL loci to the Database of Genomic Variants archive (DGV; Lappalainen et al., 2013) reveals that these gene clusters are indeed in chromosomal regions that display high levels of copy number variation, suggestive of gain and loss of haplotypes between individuals (Fig. 5). Detailed analyses of the zebrafish NITL and LITR clusters remain to be examined, but are expected to display multiple haplotypes as well.

Haplotypic variation may explain the difficulties experienced in assembling these regions of the reference genome as DNA from multiple individuals were merged into the reference genome. For example chromosome 7 genomic scaffolds 90B and 909B that encode numerous highly similar NITL genes (Fig. 5A) show different organizations for these genes: in scaffold 90B the nitr7 genes are flanked by nitr5 and nitr7 genes while in scaffold 909B the nitr5 genes are flanked by nitr2 and nitr7 genes. Although this could be a result of recent gene duplication events, it may also be the result of misassembled scaffolds resulting from joining contigs derived from different haplotypes. In addition, chromosome 7 scaffold 910 joins two contigs that both encode the mpuk421, men7 and r3 genes adjacent to two nitr5 genes. In this case, it seems likely that each contig may represent a different haplotype. A similar scenario is observed for scaffold 1953 that joins two contigs that both encode nitr2p, d1cp2, and phox2b (Fig. 5B). It is likely that the haplotype variation at these loci is far more complex than described here and may play a significant role in inter-individual immune health.

7. Conclusions

Investigations of large multigene families encoding receptor-like molecules that are predicted to effect innate immune function have been carried out in a number of different species of bony fish. In most cases, orthology with molecules identified in higher vertebrates cannot be recognized. In the case of the NITRs, the most extensively characterized of these gene families, the patterns of divergence in disparate groups are similar, although the genetic variants are extensively diverged both within and between multigene families. Taken together with the extensive differences in domain types and numbers, putative mechanisms of intracellular signaling and in some cases haplotype and extensive allelic variation, it appears that as chromosomal mechanisms are in place that facilitate rapid divergence and functional specialization of different classes of immune-type receptors even though species are presumed to be subject to generally equivalent selective pressures. Further exploration of the functional variation in such molecules will rely heavily on the use of new methods of targeted gene disruption. Investigations into the mechanisms of genetic variation that occur in these receptor gene families and how they are stabilized in the different species represent key focus areas for future research on alternative pathways of innate immunity.

Acknowledgements

We are very grateful to Charles Lee (The Jackson Laboratory Institute for Genomic Medicine) for providing the chromosomal map of copy number variation and to Sab Prowy for editorial assistance. The authors are supported by the National Institutes of Health (R01 AI057559 to G.W.L. and J.A.Y. and R01 AI23337 to G.W.L.).

References


APPENDIX B: THE CONFounding COMPLEXITY OF INNATE IMMUNE RECEPTORS WITHIN AND BETWEEN TELEOST SPECIES

The confounding complexity of innate immune receptors within and between teleost species

Dustin J. Wiesel a, Jeffrey A. Yoder b, c, *

a Department of Molecular Biomedical Sciences, North Carolina State University, Raleigh, NC 27607, USA
b Comparative Medicine Institute, North Carolina State University, Raleigh, NC 27607, USA
c Center for Human Health and the Environment, North Carolina State University, Raleigh, NC 27607, USA

Article Info

Article history:
Received 30 February 2016
Received in revised form 3 March 2016
Accepted 15 March 2016
Available online 17 March 2016

Keywords:
Gene content haplotypes
Novel immune-type receptors
Diverse immunoglobulin domain-containing proteins
Polymeric immunoglobulin receptor-like proteins
Novel immunoglobulin-like transcripts

ABSTRACT

Teleost genomes encode multiple multigene families of immunoglobulin domain-containing innate immune receptors (IIRs) with unknowns function and no clear mammalian orthologs. However, the genomic organization of IIR gene clusters and the structure and signaling motifs of the proteins they encode are similar to those of mammalian innate immune receptor families such as the killer cell immunoglobulin-like receptors (KIRs), leukocyte immunoglobulin-like receptors (LIRs), FC receptors, triggering receptors expressed on myeloid cells (TREMs) and CD38s. Teleost IIRs include novel immune-type receptors (NITRs), diverse immunoglobulin domain containing proteins (DCPPs), polymeric immunoglobulin receptor-like proteins (PIGRLs), novel immunoglobulin-like transcripts (NILTs) and leukocyte immunoreceptor-type receptors (LITRs). The accumulation of genomic sequence data has revealed that IIR gene clusters in zebrafish display haplotype and gene content variation. This interspecific genetic variation, as well as significant interspecific variation, frequently confounds the identification of definitive orthologous IIR sequences between teleost species. Nevertheless, by defining which teleost lineages encode (and do not encode) different IIR families, predictions can be made about the presence (or absence) of specific IIR families in each teleost lineage. It is anticipated that further investigations into available genomic resources and the sequencing of a variety of multiple teleost genomes will identify additional IIR families and permit the modeling of the evolutionary origins of IIRs.

© 2016 Elsevier Ltd. All rights reserved.

Contents

1. Introduction ........................................................................................................................................... 25
2. Features of zebrafish immunoglobulin domain-containing innate immune receptors (IIRs) ............................................................................................................. 26
   2.1. Zebrafish novel immune-type receptors (NITRs) ........................................................................ 26
   2.2. Zebrafish diverse immunoglobulin domain containing proteins (DCPPs) ........................ 27
   2.3. Zebrafish polymeric immunoglobulin receptor-like proteins (PIGRLs) ............................... 27
   2.4. Zebrafish novel immunoglobulin-like transcripts (NILTs) ....................................................... 27
3. Gene content haplotypes of immunoglobulin domain-containing innate immune receptor (IIR) gene clusters ......................................................................................... 28
   3.1. Haplotypic variation of zebrafish novel immune-type receptor (NITE) gene clusters ............ 28
   3.2. Haplotypic variation of zebrafish diverse immunoglobulin domain containing protein (DCP) gene clusters ............................................................................. 28
   3.3. Haplotypic variation of the polymeric immunoglobulin receptor-like (PILGR) gene cluster in zebrafish ......................................................................................... 30
   3.4. Haplotypic variation of the novel immunoglobulin-like transcript (NILT) gene cluster in zebrafish ........................................................................................................ 30

* Corresponding author. Department of Molecular Biomedical Sciences, North Carolina State University, 1060 William Moore Drive, Raleigh, NC 27607, USA.
E-mail address: Jeff.Yoder@ncsu.edu (J.A. Yoder).

http://dx.doi.org/10.1016/j.fsi.2016.03.054
1080-6068/© 2016 Elsevier Ltd. All rights reserved.
1. Introduction

Numerous multigene families encoding immunoglobulin (Ig) domain-containing, innate immune receptors (IIgRs) of unknown function have been described from diverse teleost species (bony fish) [1]. The unifying features of these receptors are that they (1) possess one or more extracellular Ig domain, (2) are encoded in multigene clusters and (3) share signaling motifs with mammalian innate immune receptors. Most IIgRs are type I transmembrane proteins with Ig domains at the amino-terminus and a cytoplasmic tail at the carboxyl-terminus; however, multiple IIgR transcripts have been reported that lack a transmembrane domain and are predicted to encode secreted Ig domains [1]. The general protein structure of teleost IIgRs is highly similar to well characterized families of mammalian innate immune receptors in that they contain activating and inhibitory signaling motifs which, in mammals, mediate innate cell function. These mammalian families include the killer cell immunoglobulin-like receptors (KIRs), leukocyte immunoglobulin-like receptors (LIRs), Fc receptors, triggering receptors expressed on myeloid cells (TREM) and CD300s [2–7].

The hallmark of many mammalian activating immune receptors is the presence of a charged residue within the transmembrane domain which permits the physical partnering with an adapter protein (e.g. DAP12, DAP10, FcγR and CD3) that possesses a reciprocally charged transmembrane residue and a cytoplasmic activation motif. For example, certain activating KIRs (on NK cells) and TREM-1 (on neutrophils and macrophages) partner with the adapter protein DAP12 to activate specific immune cell lineages, whereas multiple Fc receptors associate with the adapter protein FcγR. The most common activation motif has been termed an immunoreceptor tyrosine-based activation motif (ITAM) with the consensus sequence YXX(L/V)_{a,b}YXX(L/V)_{1,2} (where x is any residue) [8,9]. Cytoplasmic ITAMs are present in DAP12, FcγR and CD3; but can also be found in certain activating immune receptors (e.g. FcγRIIA, CD79A/B) negating the need for an adapter protein. Once an activating receptor engages its ligand, the tyrosines in the ITAM are phosphorylated by Src family protein tyrosine kinase which provides a docking site for a Syk family protein tyrosine kinase, such as ZAP-70 or Syk [8]. Syk kinases then phosphorylate a variety of substrates which can lead to calcium influx, Ras activation, stimulation of the ERK/MAPK pathway and cellular activation (e.g. cytokine release, release of cytolytic granules, etc.) [10]. In contrast, mammalian inhibitory immune receptors possess one or more cytoplasmic immunoreceptor tyrosine-based inhibition motif (ITIM) with the consensus sequence S(IV/L)VXxxI/I.L. Ligand engagement by inhibitory receptors leads to tyrosine phosphorylation in the ITIM by Src family kinases and the recruitment of phosphotyrosine phosphatases (SHP-1, SHP-2 or SHIP) that interfere with the ITAM signaling pathway [8,10]. Thus, these families of receptors receive and transmit extracellular signals that either activate or suppress immune cell function.

In mammals, the balancing between receptor mediated ITAM and ITIM signaling plays important roles in multiple immune cell lineages. ITAM-mediated ITAM and ITIM signaling influences whether an NK cell will be activated to kill an infected or transformed cell [11]. B-cell activation through B-cell receptor-mediated ITAM signaling can be countered by ITIM-containing receptors such as the Fc receptor FcγRIIB [12]. Certain members of the CD300 family bind phospholipids and likely mediate immune-mediated clearance of apoptotic cells [6]: CD300c and CD300a have been shown to provide ITAM and ITIM mediated signaling in mast cells and monocytes [13,14].

Clear genetic orthologs of mammalian KIRs, LIRs, Fc receptors, CD300s and TREMs have not been identified from any bony fish, presumably because of the recent and rapid in specie diversification of these gene clusters. However, five multigene families of IIgRs have been described in representative teleosts that share structural features and signaling motifs with these mammalian multigene families [1]. These gene families include the novel immune-type receptors (NITRs) [15]; diverse immunoglobulin domain containing proteins (DIPCs) [16,17]; polymeric immunoglobulin receptor-like proteins (PIGRIs) [18,19]; novel immunoglobulin-like transcripts (NILTs) [20–22] and leukocyte immune-type receptors (LITRs) [23,24]. Teleosts also encode orthologs to the mammalian adaptor proteins DAP12, DAP10, CD3ε and FcγR [25,26]. It is cautioned that the identification of specific IIgR families and adaptor proteins in a number of teleost genomes does not indicate that all of these genes and gene families are present in all teleosts. As discussed below, it is likely that some IIgR gene families are derived features of specific teleost lineages while other IIgR appear to be of an ancient origin but may be lost in different lineages.

Fig. 1. General structures of IIgR proteins. Ig domains are shaded tan. Activating receptors possess either a cytoplasmic ITAM (depicted by "ITAM" shaded green) or utilize a charged residue (depicted by "-" within a transmembrane domain) to associate with an adapter protein that possesses an ITAM or other activation motif. Inhibitory receptors possess one or more cytoplasmic ITIM (depicted by "ITIM" shaded red). Functionally ambiguous IIgRs act identical signaling motifs. Secreting IIgRs lack transmembrane domains. Although two Ig domains are shown for these proteins, the number of Ig domains in a single IIgR ranges from one to six (see Fig. 2B).
specific species or lineages. Zebrasfish has the most complete telost reference genome with NTRs, DICPs, PIGRLs and NLTs readily identifiable. Although LITRs can be identified in the zebrasfish genome, they have been more thoroughly described from channel catfish (Ictalurus punctatus) [23]. Therefore, for the purpose of this review, we will first describe NTRs, DICPs, PIGRLs and NLTs from zebrasfish with subsequent comparisons to other teleost genomes.

2. Features of zebrasfish immunoglobulin domain-containing innate immune receptors (IIIRs)

The uniting feature of IIIRs is that they all share the same generalized protein structure. All IIIRs consist of a leader peptide and one or more extracellular Ig domains which are each approximately 100 amino acids in length and form an “Ig fold” structure. Four sets of Ig domains have been described: variable-like domains (V), constant-like domains (C1 and C2) and intermediate domains (I) [27] and different IIIRs possess different combinations of these Ig domains. For clarity, we refer to each Ig domain based on its position within the IIIR – for example, the Ig domain most distal to the cell membrane is referred to as D1 and more proximal Ig domains are numbered consecutively (e.g. D2, D3 etc). These Ig domains can be further characterized through phylogenetic analyses, which often reveal that D1 domains form their own clade, and D2 domains typically form a clade separate from D1 domains. This has led to the identification of some IIIR members that contain three or more domains in a D1-D2-D2 array instead of a D1-D2-D3 array. Most IIIRs also encode a transmembrane domain and a cytoplasmic tail; while other IIIRs lack a transmembrane domain and are predicted to be secreted. Putative activating IIIRs possess either a charged residue within their transmembrane domain and are predicted to signal via an adapter protein or possess an ITAM within their cytoplasmic tail. Predicted inhibitory IIIRs possess one or more cytoplasmic ITIMs. Functionally ambiguous membrane IIIRs lack a charged residue within their transmembrane domain (Fig. 1).

2.1. Zebrasfish novel immune-type receptors (NITRs)

A maximum haplotype of 36 NITR genes have been described from overlapping genomic (BAC and PAC) clones that map to a single gene cluster on zebrasfish chromosome 7 [28]. A maximum haplotype of three additional NITRs have been mapped to a second gene cluster on chromosome 14 (reference genome Zv8; Fig. 2A) [29]. NTR proteins possess one or two Ig domains which are of two types and easily resolved via phylogenetic analyses [15]. For NITRs that possess two Ig domains, the membrane distal D1 domain is always a V domain and the membrane proximal D2 is always an I domain [15]; however, there are instances where a NITR gene encodes both a V and an I domain but alternative mRNA splicing removes the V domain [30]. All NITR genes that encode a single Ig domain encode a V domain. Zebrasfish NITRs can be organized into 14 subgroups based on sequence similarities of the V (D1) domain. For example, the largest zebrasfish NTR subgroup is NITR1, which includes 14 described genes, nitr1.1, nitr1.2, nitr1.3 etc. whose V domains are >70% similar. Unique among the IIIRs discussed in this review, many NITRs also possess a joining (J) region-like motif directly after the V and/or I domain [31]. The joined V-J sequences in NITRs are reminiscent of T-cell receptors that undergo V-J recombination; however, the V-J sequences of NITRs are encoded in single exons and do not undergo recombination [32]. Another feature of the NITRs that distinguishes them from other IIIR families is the presence and position of six conserved cysteine residues in their I domains [1]. In addition to C85 and C90, which are features of nearly all Ig domains [33], NITR 1 domains possess four additional cysteines which are commonly located at C170/175, C173/174, C193 and C199 [15].

Zebrasfish NITRs include activating, inhibitory, secreted and functionally ambiguous forms (Fig. 2B) [15]. We have predicted that the NITRs may be “functional orthologs” of mammalian NK receptors such as the KIR family. NITR transcripts have been recovered from hematopoietic tissues and from isolated lymphocytes, but not from isolated myeloid cells [34]. When activating and inhibitory NITRs are expressed as recombinant proteins on the surface of human cell lines and cross-linked, they can activate and inhibit the ERK/MAPK pathway, respectively [35,36]. In addition, cross-linking a recombinant activating NITR (zebrasfish Nitr9) on the surface of a human NK cell line resulted in increased target cell killing [26]. Finally, certain activating NITRs have been shown to
play a role in all recognition, although the exact ligand remains elusive [37]. As most of the sequence diversity of NTRs is observed in hypervariable regions of the V domain, it is likely that the NTR family recognizes a wide range of ligands [15]. Limited sequence diversity in the I domain may indicate that these domains play more of a structural role or that the I domains interact with another as of yet unidentified, slowly evolving molecule domain.

2.2. Zebrafish diverse immunoglobulin domain containing proteins (DICPs)

Twenty-nine DCP genes have been described on zebrafish chromosomes 3, 14 and 16 (reference genome Zv8; Fig. 2A) [16,17]. DCP proteins possess one or two Ig domains which are of two types and easily resolved via phylogenetic analysis. For DCPs that possess two Ig domains, the membrane proximal D2 domain is always a V or V-like domain and the membrane proximal D2 domain is most similar to an I domain [17]. Certain DCP genes encode a D1 and D2 domain but alternative mRNA splicing removes the D1 domain [16,17]. All identified DCP genes that encode a single Ig domain always encode a D1 domain. Amplification of partial cDNAs suggests that DICPs with four Ig domains in a D1-D2-D1-D2 configuration may exist, however, full-length transcripts have not been recovered. Zebrafish DICPs can be organized into three subgroups based on sequence similarities of the D1 domain and these subgroups correlate with the chromosome on which they are encoded. The DICP1, DICP2 and DICP3 subgroups are encoded on chromosomes 3, 14 and 16, respectively. The DICP1 gene cluster is the largest which includes at least nineteen genes, dcp1.1, dcp1.2, dcp1.3 etc.

Zebrafish DICPs include putative inhibitory, activating, secreted and functionally ambiguous forms (Fig. 2B). Although transcripts of most DICPs have been recovered from both lymphoid and myeloid cells, a few DICPs are more abundantly expressed in the lymphoid (dcp3.2) or myeloid (dcp3.3) lineages [16]. The high level of sequence similarity among zebrafish DICPs [1] suggests that DICPs experienced a recent zebrafish-specific expansion or are under selective pressure to maintain similar sequences. Certain recombinant DCP Ig domains bind bacterial phospholipids [17] and an ability shared with the mammalian CD300/TREM family of receptors [38–40], which may provide a mechanism for recognizing pathogens or apoptotic cells. DICPs share a modest amount of sequence similarity to mammalian CD300 molecules, however true orthology has not been demonstrated. DICPs on chromosome 3 have been shown to be linked to a MHC class I*Z gene cluster but it is unclear if this reflects a functional relationship [16,41,42].

2.3. Zebrafish polymeric immunoglobulin receptor-like proteins (PIGRRLs)

Teleost PIGRRL sequences were originally named soluble immune-type receptors (SITRs) [19]. However, because some members of this family are membrane bound (e.g. not soluble) and because this gene cluster is adjacent to the single copy pigr gene in zebrafish, we refer to them as the pigr-like (PIGRRL) family [18].

In mammals, the polymeric Ig receptor (pigr) is a membrane-bound protein that transports soluble polymeric antibodies (pigs) such as piga across mucosal epithelial cells [43–45]. In zebrafish, pigr is encoded by a single copy gene (pigr) on chromosome 2 that possesses two extracellular Ig domains. Twenty-nine PIGRRL genes have been described in a single cluster adjacent to the zebrafish pigr gene (reference genome Zv8; Fig. 2A) that, based on sequence similarities, have been organized into four subgroups (PIGR1, PIGRL2, PIGRL3 and PIGRL4) [18]. PIGRL3 is the largest subgroup with twelve described genes (pigr3.1, pigr3.2, pigr3.3 etc.).

Zebrafish PIGRL proteins share modest sequence homology to mammalian pigr as well as Nkp44 and members of the CD300 and TREM families [18]. With the single exception of PIGRL3.10, PIGRL proteins possess two Ig domains which are of two types (D1 and D2) and easily resolved via phylogenetic analyses: PIGRL3.10 displays a D1-D2-D2 organization [18]. Although not recognized as authentic V domains, D1 and D2 domains of PIGRLs share multiple core residues that are conserved with V domains.

No activating PIGRL has been identified from the zebrafish genome: putative inhibitory and secreted forms have been described (Fig. 3B). Transcripts of PIGRLs are not detected in the developing embryo or larvae and it is not until adulthood that PIGRL expression is detected in hematopoietic tissues. In contrast, pigr transcripts are detected as early as 36 h post fertilization (hpf). While transcripts of the PIGRL1 subgroup are more abundant in lymphoid cells, transcripts of the PIGRL2 subgroup are more abundant in myeloid cells and PIGRL3 and PIGRL4 transcripts are expressed in both lineages: pigr transcripts have not been recovered from either hematopoietic lineage [18]. In general, transcript levels of PIGRLs are increased during infection with Streptococcus iniae; whereas, infection with snakehead rhabdovirus resulted in a general decrease in levels of PIGRL transcripts, in contrast to pigr which binds antibodies [46–48]. It was demonstrated that certain PIGRLs bind phospholipids [18], an ability shared with DCPs (above) and the mammalian CD300 and TREM families of receptors [38–40]. Based on these differences in ligands and expression patterns, it is clear that PIGRL proteins provide a distinctly different function than pigr.

2.4. Zebrafish novel immunoglobulin-like transcripts (NILTs)

NILTs have been described from common carp (Cyprinus carpio), rainbow trout (Oncorhynchus mykiss) and Atlantic salmon (Salmo salar) and predicted from the zebrafish genome [120–22]. Sequence comparisons revealed homology between NILTs and human Nkp44 and members of the TREM and CD300 families [22,49]. Although this similarity to human proteins is reminiscent of DICPs and PIGRRLs (see above), phylogenetic comparisons indicate significant divergence between zebrafish IIR families [1]. In zebrafish, NILTs are encoded in a large cluster on chromosome 1, approximately 7 Mbps downstream of MHC class I*Z lineage genes (Fig. 2A) [22,41,42]. Linkage of the NILT gene cluster to a MHC class I*Z gene cluster is yet another similarity shared between NILTs and DICPs. Further characterization is required to determine if NILTs and DICPs share a common origin.

NILT transcripts recovered from carp, rainbow trout and Atlantic salmon encode one or two Ig domains and reflect inhibitory, activating and secreted forms [reviewed in 1, 23]. In rainbow trout, it was demonstrated that NILTs are expressed in the egg and 48 hpf embryo [21] and transcripts of Atlantic salmon NILTs have been recovered from hematopoietic tissues [22]. Transcriptional profiling of one activating NILT and one inhibitory NILT from carp (NILT1 and NILT2, respectively) revealed that neither was detectable in freshly isolated peripheral blood leukocytes, but transcript levels of the activating NILT was dramatically increased after overnight culture – it has been suggested that cytokines present in the culture media may stimulate the transcription of this receptor [22]. Ig structures of carp NILTs were inferred from homology modeling against Nkp44’s crystal structure [22]. NILT Ig domains are predicted to form a disulfide bond between two characteristic cysteines which are spaced 3, 6 or 7 residues apart (CxxC, CxxxC or CxxxC) [22]. This disulfide bond is in addition to and internal to C23 and C24 which are spaced ~65 residues apart: this spacing of C23 and C24 is characteristic of V domains [27]. To our knowledge, the CxxC motif is unique to NILT Ig domains, whereas the CxxxC motif is
common among mammalian CD300 proteins. The Cx6C motif is found commonly among Ig domains including those present in human NKp44 and NKp8 and zebrafish pIgR and pIgRL.

By mining the zebrafish reference genome (GRCz10), we have identified 74 NITL-related Ig domains within a single cluster on chromosome 1 and sequenced cDNAs defining nine genes (unpublished observations), including one transcript encoding a NITL with six Ig domains (Fig. 2B). Ensembl gene annotation (release 83) [56], which utilizes transcriptional evidence such as expressed sequence tags (ESTs) and next-generation sequencing data, predicts no overlap of known or predicted genes with most of these Ig domains. Due to the highly variable nature of this gene family, D1 and D2 domains are not easily resolved through phylogenetic analyses. This region of the zebrafish genome remains poorly characterized, possibly due to low levels of NITL expression (as suggested by the lack of EST data) under normal, healthy conditions.

3. Gene content haplotypes of immunoglobulin domain-containing innate immune receptor (IRR) gene clusters

Multigene families experience a multitude of evolutionary forces including gene duplication, gene loss, mutation, recombination, micro-recombination, gene conversion, genetic drift, and gene flow [51–54]. Within a species, inter-individual genetic difference at immune gene clusters can include: (1) polymorphic variation (sequence differences between the same gene); (2) polygenic variation (multiple copies of highly similar genes); (3) haplotype variation (different combinations of polymorphic variations in neighboring genes); and (4) gene content variation (different combinations and/or numbers of genes within a locus). For these reasons, members of immune-related gene families are among the highest polymorphic regions of the genome. For example, more than 500 different KIR genotypes have been described from approximately 18,000 humans [55] and this variation includes differences in KIR gene content (Fig. 3) [56].

Gene content variation in certain immune-related genes families has been found to be associated with disease susceptibility. An expanded repertoire of activating KIRs in humans, for example, reduces the rate of CMV infection in kidney transplant recipients [57]. Conversely, haplotypes containing two or more activating KIR genes were found more frequently in patients with systemic lupus erythematosus [58]. When both KIR and HLA (ligands for KIRs) haplotypes are considered, many more associations with disease susceptibility can be made [59,60], including: AIDS [61], Crohn’s disease [62], malignant melanoma [63], pre-eclampsia [64], and type I diabetes [65,66]. As discussed below, there is growing evidence that zebrafish IRR gene families also display gene content variation; however, the extent and the functional consequences of this variation remain to be resolved. A detailed examination of IIR sequence differences between individual zebrafish from different genetic backgrounds will be required to determine the extent of gene content variation at these loci.

3.1. Haplotype variation of zebrafish novel immune-type receptor (NITL) gene clusters

Alternate haplotypes that include gene content variation have been described for both NITL gene clusters on zebrafish chromosomes 7 and 14 [28,29]. These haplotypes were identified from the sequencing of individual genomic clones of the RUSM 1 PAC library which was derived from the polymorphic AB strain of zebrafish and from additional genomic clones of the Danio rerio BAC library which was derived from the polymorphic TU strain of zebrafish [67]. Interestingly, the AB library was generated using approximately 200 individual zebrafish [68] which is likely why multiple NITL haplotypes were identified. This genetic diversity is one reason why the zebrafish genome project employed genomic libraries derived from single individuals [69].

Gene content variation has been described from three different regions of the NITL gene cluster on chromosome 7 (Fig. 4A) [28,29]. At the first region, the sequence between the nitr3a and nitr3d genes encodes either the nitr3b gene, the nitr3c gene or no detectable gene sequence. At the second region, the sequence between the nitr1d and nitr1l genes includes either the combination of the nitr1f, nitr2e and nitr1l genes or the combination of the nitr1e pseudogene and the nitr2d and nitr1g genes. At the third region, the sequence between nitr1m and nitr6a includes either the combination of the nitr1o, nitr5 and nitr6a genes or the combination of the nitr1n, nitr1a, nitr5, and nitr4a genes along with the nitr4k pseudogene [28]. Similarly, gene content variation at the NITL gene cluster on chromosome 14 has been described in which the nitr14a gene is either present or absent (Fig. 4A) [29]. These alternate haplotypes on chromosome 14 were identified from the sequences of one BAC clone of the Danio rerio library and one BAC clone from the RPT1 library (which was generated from 7000 TU embryos).

It is noted that the gene content variation at the NITL gene clusters is accompanied by dramatic differences in polymorphic variation and intergenic distances. Although the inter-individual sequence diversity of NITLs remains to be strategically investigated, we have sequenced numerous NITL transcripts that do not map to the reference genome suggesting that more haplotypes and genes remain to be described.

3.2. Haplotype variation of zebrafish diverse immunoglobulin domain-containing protein (DICP) gene clusters

Alternate haplotypes that include gene content variation have been described for the DICP gene clusters on zebrafish chromosomes 3 and 16 [16,17]. These haplotypes were identified by comparing the reference genome (derived from sequences of TU zebrafish) to implanted genomic scaffolds (for other genomic scaffolds see Table 1) to identify insertions and/or deletions. Of these, 19 insertions and/or deletions were found to be associated with disease susceptibility. The most common of these insertions and/or deletions were found in the DICP gene cluster on chromosome 16 (reference genome 2.56). Polymorphic variation was evident at this gene cluster as the reference genome predicted

---

*Fig. 3. Gene-content variation of the human KIR gene cluster on chromosome 18q13. Each horizontal line represents one gene content haplotype with each KIR gene represented by a circle. Seven described gene content haplotypes are shown. Framework genes are present in all haplotypes and shaded black. Figure adapted from Ref. [59]. Intergenic distances are not to scale, but gene order is accurate.*
that dip1.4 was a pseudogene based on a frame shift in the D2 exon, whereas genomic sequence from the CHORI-73 individual lacked this frame shift and was predicted to encode a functional dip3.4 gene (Fig. 4B) [17]. In addition, six DICP transcripts were recovered from polymorphic EKX strain of zebrafish that did not map to the TU reference genome [17]. This observation supported the hypothesis that the reference genome was not complete at this locus and/or not reflective of the inter-individual sequence diversity of DICPs within the zebrafish species.

Recently, the sequencing of nearly 200 individual DICP cDNA amplicons from individual zebrafish of the TU, AB, and EKX lines led to the description of eleven new DICP sequences [18]. Two of these eleven sequences correspond to the dip1.22 and dip1.23 genes which are adjacent to the gimap4, cdc13a1 and mhc1z6a genes on unplaced scaffold NA310 from the reference genome (GRCz10). As the previously described dip1.1—dip2.21 gene cluster is adjacent to the gimap4, cdc13a1 and mhc1z6a genes in the reference genome, dip1.22 and dip1.23 likely represent part of an alternate haplotype on chromosome 3 (Fig. 4B); however, the remaining nine DICP sequences fail to map to any publicly available zebrafish genome sequence [18]. This suggests that these nine new DICPs may map to extended regions of the dip1.22/dip1.23 haplotype or possibly to additional haplotypes.
3.3. Haplotypic variation of the polymorphic immunoglobulin receptor-like (PIGRL) gene cluster in zebrafish

Two alternate haplotypes have been described for the PIGRL gene cluster on zebrafish chromosome 2, twenty-nine PIGRL genes have been identified on three TUG genomic scaffolds from the zebrafish reference genome (24); scaffolds 234 and 235 mapped to zebrafish chromosome 2 while scaffold 3509 was unplaced in the reference genome [18]. Interestingly, no sequence overlap was identified between scaffolds 234 and 235 suggesting that a gap may exist between these scaffolds and that additional PIGRL genes could be present in such a gap. Scaffold 234 encodes the single copy dad1 and pigr genes which likely denote one end of the PIGRL gene cluster and scaffold 235 encodes the ircc24 gene that likely denotes the other end of this gene cluster. Scaffold 3509 encodes the ircc24-like (ircc34) gene and a few adjacent PIGRL genes that share high homology to PIGRL genes that are adjacent to ircc24 gene in scaffold 235. This high level of sequence identity across approximately 50 kbp indicates that scaffolds 235 and 3509 represent (partial) alternate haplotypes of the PIGRL gene cluster [18]. Although the sequence and organization of the pigr33.8, pigr33.9, pigr44.1, pigr44.2 and ircc24 genes on scaffold 235 are highly similar to the pigr33.11, pigr33.12, pigr44.3, pigr44.4 and ircc44 genes on scaffold 3509, the adjacent PIGRL genes share very little homology (Fig. 4C) [18]. It is likely that scaffold 3509 was removed from the subsequent reference genome assembly (GR6210) because it represents an alternate gene content haplotype of the PIGRL gene cluster.

3.4. Haplotypic variation of the novel immunoglobulin-like transcript (NILT) gene cluster in zebrafish

NILTs were first reported from common carp in 2005 and mining the zebrafish reference genome (most likely reference genome 24) identified multiple putative NILT genes on zebrafish chromosome 1 [22]. This report noted that numerous additional Ig domains were present in the genomic regions flanking the zebrafish NILT sequences and we have confirmed that the current version of the reference genome, GR6210, contains at least 74 Ig domains with sequence homology to carp NILTs. By data mining previous versions of the zebrafish genome, we have identified a possible alternate gene-content haplotype for the NILT gene cluster (unpublished observations). For example, in the ninth version of the zebrafish reference genome (24), NILT Ig domains are present on scaffolds 100 and 101 of chromosome 1. These two scaffolds contain highly similar genes but differ in multiple single nucleotide polymorphisms and the number of predicted NILT genes (unpublished observations). It is likely that scaffold 100 was removed from the latest version of the zebrafish genome (GR6210) because it reflects an alternate gene content haplotype of the NILT gene cluster.

4. Identifying immunoglobulin domain-containing innate immune receptor (IIRs) in teleosts

To the best of our knowledge, the IIRs described above have not been identified outside of teleosts, with the exception of NITRs that have been identified in the more basal Holostei lineage (spotted gar, Lepisosteus oculatus) [76]. However, the high level of IIR sequence diversity between teleosts can confound the identification of orthologous sequences between teleost genomes. This diversity is further complicated by the fact that Ig domains possess only a few highly conserved residues making it sometimes difficult to distinguish between different gene families when comparing sequences of divergent species. Therefore, defining orthologous families of IIRs between teleosts may require the combination of sequence homology, the presence of protein family-specific conserved residues [1] and evidence for conserved synteny. Identifying syntenic regions between teleost genomes can become challenging because of ancient whole genome duplications (WGDs) that hypothetically led to multiple gene clusters of each IIR family. However, different teleosts may have different numbers of IIR gene clusters. For example, NITRs are encoded by two gene clusters in zebrafish, but three gene clusters in medaka (Oryzias latipes) [25,71]. One explanation for variable numbers of IIR gene clusters in these species, is that there has been selective pressure to maintain duplicated IIR gene clusters in these species, but not in others. The “2K” hypothesis suggests that there were two vertebrate-specific whole genome duplication events (WGDs) [72] and more recent evidence suggests that there was an additional “3K” teleost-specific whole genome duplication (TSGD) [73]. Evidence for the TSGD is provided by the identification of four extant paralogous regions of the MHC which are predicted to be derived from an ancient “proto-MHC” [74]. Evidence for the TSGD is provided by the observation that numerous mammalian single copy genes have two definitive orthologous genes in teleosts [73]. Within various teleost lineages, additional rounds of WGD have occurred and can further complicate the identification of genes. For example, salmonids likely experienced an additional “4R” WGD [75] and common carp (Cyprinus carpio) appear to have experienced a very recent (~4.2 million years ago) “4R” WGD [76]. Thus, it may not be surprising if more IIR gene clusters are found in “4R” salmonids and common carp than other “3R” teleost genomes.

Concerted evolution offers an additional explanation as to why identification of IIR orthologs between species is often so difficult. Concerted evolution posits that gene copies are used as templated repair for family members, thereby reversing divergence of the family but in a species specific manner [77]. Each species will continually evolve independently of other species, resulting in species-specific sorting of gene sequences. For example, NITRs of zebrafish, while displaying as low as 70% sequence identity among other zebrafish NITR sequences, are often more closely related to any other zebrafish NITR than they are to any NITR from another species [15,71]. These species-specific differences in IIRs are reflective of recent and rapid gene duplications (within each phylogenetic lineage) and diversification under pressures of concerted evolution.

4.1. Novel immune-type receptors (NITRs) throughout teleosts

Most NITR genes can be readily identified by the presence of a single V (D1) domain and a single I (D2) domain that contains six highly conserved cysteines (see above) NITR genes that lack an I domain can be identified by the presence of a single V domain and by being encoded within the same genomic region as I domain-containing NITR genes [1,19]. Twenty-six NITR genes were first described from a single BAC clone from southern pufferfish (Sphoeroides peploides) [78]. Full-length NITR transcripts (without corresponding genomic sequences) were subsequently identified from zebrafish, rainbow trout and channel catfish [35,29-31] and mining EST databases revealed NITR transcripts from Atlantic salmon, Atlantic cod (Gadus morhua), Atlantic halibut (Hippoglossus hippoglossus), lake whitefish (Coregonus clupeaformis) and three-spined stickleback (Gasterosteus aculeatus) [15]. In addition to the southern pufferfish, detailed descriptions of NITR gene clusters have been reported from zebrafish (described above), medaka, European sea bass (Dicentrarchus labrax) and minnow eel (Miichthys minnow) [71,82,83]. NITR sequences have also been identified from the genomes of fugu (Takifugu rubripers) and the green spotted pufferfish (Tetraodon nigroviridis) [28,71].

NITR sequences have been identified from both major radiations
of teleosts. Ostariophysi and Euteleostei (Fig. 5). Ostariophysi known to encode NTRs include: cypriniformes (zebrafish and carp), clupeiformes (whitefish), and siluriformes (catfish). Euteleostei known to encode NTRs include: Salmonidae (salmon and trout), Gadiformes (cod), Cichlidae (tilapia), Atherinomorpha (medaka), Perciformes (stickleback) and Tetraodontiformes (pufferfish). Although NTRs are not identifiable in coelacanths or other tetrapods, they have been identified in the more basal Holostei lineage (spotted gar) [70]. The presence of NTRs in all examined teleosts but in Holostei, indicates that NTRs were present prior to the teleost-specific WGD and are likely present in all teleosts (although it is possible that they may be lost in certain species or lineages). At teleosts comprise nearly half of all known vertebrates [84] it is possible that NTRs provide immune function in ~50% of all species within this subphylum.

4.2. Diverse immunoglobulin domain containing proteins (DICPs) in Cypriniformes

Most DICP genes can be identified by the presence of a single D1 domain, a single D2 domain, and the spacing of two conserved cysteines (located between C2 and C3) in both Ig domains (Cys5-Cys[1]). DICP genes that lack a D2 domain have also been described [16,17].

Definitive DICP species have only been identified from three cyprinid species: zebrafish, grass carp (Ctenopharyngodon idella) and common carp (Cyprinus carpio). DICP-like sequences have been identified in multiple Euteleostei including Atlantic salmon, fugu, green spotted pufferfish, medaka, and Nile tilapia (Oreochromis niloticus) [16,17]. However, due to independent clade formation in phylogenetic analysis, these DICP-like sequences remain unclassified as true DICP orthologues. These observations suggest that definitive DICPs may be restricted to Cypriniformes or to the larger group of Ostariophysi – it is unknown if Siluriformes (e.g., catfish) or any other Ostariophysi lineage encodes DICPs. The sequencing of additional Ostariophysi genomes will help resolve this issue.

Although there are three DICP gene clusters in zebrafish, one of them is adjacent to a MHC class I Z lineage gene cluster. Importantly, genomic scaffolds of DICP genes in grass carp and common carp also encode MHC Class I Z lineage genes [16]. This suggests that DICPs may have originated from an ancestral MHC loci and may be coevolving with MHC Z genes/haplotypes. It remains to be determined if DICP-like genes in non-cyprinds are linked to MHC sequences.

4.3. Polymeric immunoglobulin receptor-like (PIGR) sequences in select teleosts

The zebrafish PIGR gene cluster is adjacent to the single-copy pig gene (see above) and presumed to have arisen through multiple gene duplication events. All defined PIGR genes can be identified by the presence of a single D1 domain, a single D2 domain and the spacing of highly conserved cysteines in the D2 domain CysX-Cys[1]. Most PIGR D1 domains also possess cysteines with conserved spacing (CysX-CysX). Single definitive PIGR transcripts have been identified from common carp and Atlantic salmon (Fig. 5) [16]. Interestingly the D1 domain of the salmon sequence is most similar to the D1 domain of zebrafish PIGR proteins, whereas its D2 domain is more similar to the D2 domain of the pig proteins from multiple teleosts. This may suggest that (1) there has been selective pressure on the D2 domain of this salmon PIGR to maintain function or (2) this sequence reflects a recombination event between the PIGR gene and a PIGR gene.

As described above, the zebrafish pig gene and the PIGR gene cluster are flanked by the daf1 and irr24 genes (Fig. 4C). The reference genomes of fugu, Nile tilapia and stickleback all display conserved synteny to this genomic region with their PIGR genes.
flanked by the DADJ1 and LRRC24 genes but with no evidence for 
PIGR1 sequences. Although the medaka genome lacks PIGR genes at 
this locus, four tandem copies of the PIGR gene are present be-
 tween the DADJ1 and the LRRC24 genes. The presence of PIGR1 
sequences in both Ostariophysi (zebrafish in Cypriniformes) and 
Euteleostei (salmon in Salmonidae) suggests that these genes were 
present in the ancestral teleost. If true, then the specific absence of 
PIGR genes in representative species of related Euteleostei line-
eges (Cichlidae, Atherinomorpha, Perciformes, Tetraodontiformes) 
indicates that these genes must have been lost in these lineages 
after their split from other Euteleostei lineages (Salmonidae and 
possibly Esociformes, Osteichthyes, and Gadiformes) [Fig. 5]. The 
sequencing of additional teleost genomes will help resolve this 
issue.

4.4. Novel immunoglobulin-like transcripts (NILTs) in select teleosts

NILTs were first identified in common carp [22] and subse-
quently identified in rainbow trout [21,49], and Atlantic salmon 
[20]. NILTs described from these species possess either a single D1 
domain or a D1 domain and a D2 domain [1]. These NILTs can be 
divided into at least three subgroups based on the spacing between 
conserved cysteines in the Ig domains = Cys/C, Cys/C, and Cys/C 
[20]. However, zebrafish appear to encode NILT proteins with more 
than two Ig domains and more than three subgroups of Ig domains 
(unpublished observations). Future comparisons of the NILT gene 
clusters of trout and salmon to that of zebrafish may reveal 
important insights into their evolution.

While extensive multi-species studies of NILTs have not yet been 
reported, initial estimates suggest that NILTs have expanded their 
numbers greatly in cyprinids (>56 sequences in carp, 74 Ig domains 
in zebrafish) whereas only a few (~10 genes) have been found in 
salmonids. This is interesting, because the salmonid-specific WGD 
[48] would have generated more NILT gene clusters in these spe-
cies [72]. Although NILTs have been reported from both teleost 
radiations (Cypriniformes in Ostariophysi and Salmonidae in 
Euteleostei) they were explicitly not identified in the fugu genome 
(Fig. 5) [22], suggesting that NILTs may have been lost in this 
and related species.

5. Conclusions and notes for future directions

Of the four ILIR gene families discussed above, the NTR genes are 
likely the most ancient as they have been identified in all teleost 
species and radiations examined as well as in the more basal spe-
cies gar. Although the D1CR and D1CR-like genes appear to be 
present throughout the teleosts, more research is required to de-
fine the evolutionary relationship between these sequences. PIGR1 
and NILTs have been identified from Cypriniformes (zebrafish and 
carp) in the Ostariophysi lineage and from Salmonidae (salmon and/or 
trout) in the Euteleostei lineage, but specifically absent in other 
Euteleostei species.

It is highly likely that many more ILIR gene families exist. For 
example, the linkage of D1CRs and NILTs to MHC class I 
genes supports the model that similar genes were linked in the 
ancient "proto-MHC" [85,86]. This model places the primordial 
MHC genes in the same chromosomal region as Ig domains con-
taining receptors, which would be the precursor to modern 
mammalian receptors encoded at the human leukocyte receptor 
complex (e.g. KIRs etc.) [74]. In addition, at least one NTR has 
been shown to function in allorrecognition and may paralize the 
function of mammalian NK receptors that bind MHC class I mole-
cules as "self" [37].

In contrast, PIGR1 likely arose from gene duplications of the 
PIGR gene, which is not believed to be part of the "proto-MHC", but 
shares many sequence similarities with both D1CRs and NILTs. 
Determining if PIGR1s, D1CRs, and NILTs share a common evolu-
tionary origin may shed insight into their function and suggest 
possible mammalian orthologues. The sequencing of additional 
diverse teleost genomes should help resolve these issues.

Although much sequence information has accumulated for all 
these gene families, their specific functions remain elusive. While 
lipid binding studies have demonstrated that certain D1CRs and 
PIGR1s bind phospholipids, the function of this interaction remains 
unknown — it may be that these receptors recognize pathogen-
specific lipids or membrane lipids associated with apoptotic cells. 
Although one NTR has been shown to mediate allorrecognition, the 
lipid for this receptor remains undefined.

The gene content variation of ILIR between individual zebrafish 
suggests that ILIR sequence diversity is present in other teleosts. As 
more reference genomes are sequenced from diverse species, it will 
be important to remember that a single reference genome does not 
define all the ILIR genes within a species. Although, the gene con-
tent variation of ILIRs may provide immunological strength to a 
population, it is likely that different combinations of ILIRs in 
different individuals makes them more susceptible or resistant to 
different pathogens. Similarly, different species of fish with 
different combinations of ILIR families may display different levels 
of susceptibility to the same immune challenge. Although it is 
evident that no single fish species can be employed as a model to 
understand all ILIR in all teleosts, we anticipate that genome 
engineering strategies (e.g. CRISPR/Cas9) in amenable species (e.g. 
zebrafish and medaka) will provide important functional insights 
to the ILIR families which they possess.

Acknowledgements

Dustin Wcisel is supported by the National Evolutionary Synthesis 
Center (NESCent), NSF REU-0935606.

References

1. J. Rodriguez-Nunez, D.J. Wcisel, C.W. Lmn, J.A. Yoder, Multigene families of 
immunoglobulin-domain-containing lineage-immune receptors in zebrafish 
2. K. Hara, H. Arna, Functional and genetic diversity of leucine 
immunoglobulin-like receptor and implication for disease associations. 
3. M. Carluccio, F. Moers, The ILIR Gene Cluster (National Library of Medicine 
(TN) NCBi, Bethesda, MD, 2005).
4. D.G. Augsten, M. Petri-Eber, KIR and HLA under pressure: evolution of 
5. J.W. Ford, D.W. McVean, TDM and TRED-like receptors in inflammation and 
6. F. Borrego, The CD94NOK molecules: an emerging family of regulators of 
7. C.E. Harrold, M.J. Rose-Zerilli, L.R. Mackall, C. Hynynen, F.J. Huber, M.S. 
Craig, et al., FuguFrost receptors: genetic variation, function, and disease. 
8. A.P. Brown, T. Roderick, You say ITAM and I say ITIM, let’s call the whole 
1648–1655.
9. A. Bicak, N.S. van Oers, T-cell receptor signal transduction: what gives 
10. J.A. Yoder, C.W. Lmn, The phylogenetic origins of natural killer receptors 
and recognition relationships, possibilities, and realities. Immunogenetics 63 
(2011) 133–141.


APPENDIX C: CHAPTER 2 SUPPLEMENTARY MATERIAL

- Electronic Supplementary Material –
- Online Resource 3 -

Immunogenetics

The identification of additional zebrafish DICP genes reveals haplotype variation and linkage to MHC class I genes

Ivan Rodríguez-Nunez*, Dustin J. Wcise†, Ronda T. Litman†,
Gary W. Litman*∥,* and Jeffrey A. Yoder∗∗

* Department of Molecular Biomedical Sciences and Center for Comparative Medicine and Translational Research, North Carolina State University, 1060 William Moore Drive, Raleigh, NC 27607, USA
† Department of Pediatrics, University of South Florida Morsani College of Medicine, USF/ACH Children's Research Institute, 140 7th Avenue South, St. Petersburg, FL 33701, USA
∥ Department of Molecular Genetics, All Children's Hospital Johns Hopkins Medicine, 501 6th Avenue South, St. Petersburg, FL 33701, USA
∗∗ H. Lee Moffitt Cancer Center and Research Institute, 12902 Magnolia Drive, Tampa, FL 33612, USA
*Corresponding author at: Jeff_Yoder@ncsu.edu

Contents

DICP sequence variation .............................................................................................................. 2
DICP exon-intron architecture .................................................................................................. 3
Exon-intron architecture of previously described DICPs ........................................................ 3
Exon-intron architecture of newly described DICPs ................................................................. 6
Fig. S1. Dicp1.7 ......................................................................................................................... 9
Fig. S2. Dicp1.8 ......................................................................................................................... 9
Fig. S3. Dicp1.17 ...................................................................................................................... 10
Fig. S4. Dicp1.22 ...................................................................................................................... 11
Fig. S5. Dicp2.1 ....................................................................................................................... 12
Fig. S6. Dicp3.1 ....................................................................................................................... 13
Fig. S7. Phylogenetic comparison of zebrafish DICP Ig domains ............................................ 15
Fig. S8. Zebrafish DICP protein sequences ............................................................................. 17
Fig. S9. Zebrafish DICP protein structures ............................................................................ 18
Fig. S10. Zebrafish chromosome 3 DICP haplotype dot-plot .................................................. 19
Fig. S11. DICP-MHCI linkage in grass carp and common carp .............................................. 20
Fig. S12. Carp DICP and MHCI sequence comparisons ....................................................... 22
Fig. S13. Carp protein sequences ......................................................................................... 24
Fig. S14. Tissue distribution of DICP transcripts. ................................................................. 25
Table S3. Distribution of DICP transcripts from individual zebrafish ...................................... 26
Table S4. Carp sequence information .................................................................................... 27
Reference List .......................................................................................................................... 28
DICP sequence variation

In order to investigate the allelic variation and polymorphism of DICP genes, a high fidelity DNA polymerase and a non-high fidelity DNA polymerase were used for RT-PCR and several DICP transcripts were cloned and sequenced. The inter-individual distribution of the transcripts generated with a high fidelity DNA polymerase is represented in Online Resource 3 – Table S3. These transcripts are identical or nearly identical to the previously reported sequences (Online Resource 2 – Table S2).

For example, amplicons dicp1.76953 and dicp1.83984 sequences are 100% identical to the previously described dicp1.70233 and dicp1.84282 sequences, respectively (Online Resource 2 – Table S2; Online Resource 3 – Figs. S1 and S2). Although, the dicp1.83984 sequence was 98% identical to the previously reported dicp1.84282 it includes an additional 105 nt between the D1 and transmembrane domain, likely reflecting an exon shuffling (see below). Predicted functional splice variants also are observed for dicp1.17 and dicp1.22 (Online Resource 3 – Figs. S3 and S4). Transcripts of the putative activating dicp2.1 are highly conserved, with the exception of a few single amino acid substitutions and one transcript (dicp2.1e19150) was identified in which the D1 domain is not present, presumably deleted by mRNA splicing (Online Resource 3 – Fig. S5). The dicp3.1 transcripts recovered using a high fidelity DNA polymerase encode highly similar D2 domains, but encode one of two distinctly different D1 domains (Online Resource 3 – Figs. S6 and S7). These differences reflect the two different dicp3.1 haplotypes described originally from the genomic scaffold 1952 (dicp3.11952) and BAC CH73-34N11 (dicp3.124011) (Haire et al. 2012). The group of transcripts most similar to dicp3.11952 share 98-100% identity and the group most similar to dicp3.124011 share 96-100% identity. These sequences, which likely reflect different alleles of the same gene as dicp3.11952 and dicp3.124011, are adjacent to the same gene, setdb1b (Haire et al. 2012). The sequence identities between the dicp3.11952 and dicp3.124011 groups range from 92-96%.
DICP exon-intron architecture

Exon-intron architecture of previously described DICPs

Comparing the genomic architecture of DICP genes in the Zv9 zebrafish reference genome with the sequences of DICP transcripts identified in this study, the exon-intron architecture of these DICPs has been predicted (Fig. 6).

The *dicp1.1* gene in the reference genome possesses one exon encoding a leader peptide, one exon encoding an Ig D1 domain, four exons encoding an extracellular stalk domain, one exon encoding a transmembrane domain and three exons encoding the cytoplasmic tail. Transcripts sequenced in this study span the exons that encode the leader peptide and the transmembrane domain. Most of the transcripts share the exon content of the *dicp1.1* gene from the reference genome, except the *dicp1.1* transcript which skips one exon by alternative mRNA splicing. Several transcripts reflect an intron retention resulting in a PTC.

The *dicp2.2* gene in the reference genome possesses one exon encoding a leader peptide, one exon encoding an Ig D1 domain, one exon encoding an Ig D2 domain, one exon encoding a transmembrane domain and four exons encoding a cytoplasmic tail. The *dicp2.2* transcript identified in this study spans the exons that encode the leader peptide and the transmembrane domain. This transcript reflects an alternative 3′ splice site that introduces: i) a 234 nt deletion in the exon that encodes the Ig D2 domain and ii) an intron retention between the exons encoding the Ig D2 domain and the transmembrane domain resulting in a PTC.

The *dicp1.7* gene in the reference genome possesses one exon encoding a leader peptide, one exon encoding an Ig D1 domain, three exons encoding an extracellular stalk domain, one exon encoding a transmembrane domain and three exons encoding a cytoplasmic tail. The *dicp1.7* transcript identified in this study spans the exon that encodes the leader peptide and the second exon that encodes the cytoplasmic tail. This transcript reflects the exon content of the *dicp1.7* gene from the reference genome.

The *dicp1.8* gene in the reference genome possesses one exon encoding a leader peptide, one exon encoding an Ig D1 domain, three exons encoding an extracellular stalk domain, one exon encoding a transmembrane domain and three exons encoding a cytoplasmic tail. The *dicp1.8* transcripts identified in this study span the exon that encodes the leader peptide and the second exon that encodes the cytoplasmic tail. Most of the transcripts match the exon content of the *dicp1.8* gene from the reference genome, except for the *dicp1.8* transcript which includes sequence that corresponds to two exons encoding the stalk region in the *dicp1.17* gene. The *dicp1.8* transcript was amplified from EKW zebrafish 3 and was the only *dicp1.8* sequence recovered from this
individual. As described below, two dicp1.17 transcripts, (dicp1.17\textsuperscript{2880} and dicp1.17\textsuperscript{2882}) were recovered from this same fish; these transcripts are identical except that one (dicp1.17\textsuperscript{2882}) has three more nucleotides (CAG) than the other which, based on its position at the end of an exon, may be present due to an alternative splice site. Both of these dicp1.17 transcripts and dicp1.8\textsuperscript{3984} (from EKW zebrafish 3) share identical sequences with the previously described dicp1.17 exons (Haire et al. 2012). Based on these observations dicp1.8\textsuperscript{3984} may reflect an exon shuffling or recombination event and represent an allelic variant of dicp1.8 or an undescribed DICP gene representative of an alternate haplotype.

The dicp1.9 gene in the reference genome possesses one exon encoding a leader peptide, one exon encoding an Ig D1 domain, one exon encoding an extracellular stalk domain, one exon encoding a transmembrane domain and four exons encoding a cytoplasmic tail. The dicp1.9\textsuperscript{2822} transcript identified in this study spans the exons that encode the leader peptide and the transmembrane domain. This transcript encodes three additional exons (that encode the stalk domain) which are not predicted in the reference genome.

The dicp1.16 gene in the reference genome possesses one exon encoding a leader peptide, one exon encoding an Ig D1 domain, one exon encoding an Ig D2 domain, one exon encoding a transmembrane domain and four exons encoding a cytoplasmic tail. The dicp1.16\textsuperscript{2878} transcript identified in this study spans the exons that encode the leader peptide and the transmembrane domain. This transcript encodes one additional exon (that encodes a stalk domain) which are not predicted in the reference genome.

The dicp1.17 gene in the reference genome possesses one exon encoding a leader peptide, one exon encoding an Ig D1 domain, two exons encoding an extracellular stalk domain, one exon encoding a transmembrane domain and three exons encoding a cytoplasmic tail. The dicp1.17 transcripts identified in this study span the exon that encodes the leader peptide and the second exon that encodes the cytoplasmic tail. Most of the transcripts reflect the exon content of the dicp1.17 gene from the reference genome. However, the dicp1.17\textsuperscript{2550} and dicp1.17\textsuperscript{2558} transcripts lack the second exon that encodes the stalk domain; the dicp1.17\textsuperscript{2525}, dicp1.17\textsuperscript{2580}, dicp1.17\textsuperscript{2580} and dicp1.17\textsuperscript{2569} transcripts lack both exons that encode the stalk domain; and the dicp1.17\textsuperscript{2569} transcript lacks the second exon that encodes the stalk domain and lack the exon that encodes the transmembrane domain (thus encoding a secreted protein). The dicp1.17\textsuperscript{2581} transcript reflects an alternative 3' splice site that introduces a deletion of 80 nt in the exon that encode the Ig D1 domain and a PTC.

The dicp1.19 gene in the reference genome possesses one exon encoding a leader peptide, one exon encoding an Ig D1 domain, one exon encoding an Ig D2 domain, three exons encoding an extracellular stalk domain, one exon encoding a transmembrane domain and four exons encoding a cytoplasmic tail. The transcripts identified in this study span the exons that encode the leader peptide and the transmembrane domain. Most transcripts share the exon content of the dicp1.19 gene from
the reference genome. The dicp1.19 transcript reflects an alternative splicing event in which the exon encoding the leader peptide splices to part of the intron just 5' of the exon encoding the Ig D1 domain resulting in a 14 nt insertion and a PTC.

The dicp2.1 gene in the reference genome possesses one exon encoding a leader peptide, one exon encoding an Ig D1 domain, one exon encoding an Ig D2 domain, one exon encoding an extracellular stalk domain, one exon encoding a transmembrane domain and one exon encoding a short cytoplasmic tail. The transcripts identified in this study span the exons that encode the leader peptide and the transmembrane domain. Several transcripts reflect the exon content of the dicp2.1 gene from the reference genome. However, other transcripts present a 15 nt insertion in the middle of the exon that encode the Ig D1 domain. The dicp2.1, dicp2.1592 and dicp2.1598 transcripts possess this insertion as well as a 132 nt deletion in the middle of the exon that encodes the Ig D2 domain. The Dicp2.1592 transcript lacks the exon that encodes the Ig D1 domain. Several transcripts possess PTCs introduced by intron retention.

The dicp3.1 gene in the reference genome possesses one exon encoding a leader peptide, one exon encoding an Ig D1 domain, one exon encoding an Ig D2 domain, one exon encoding an extracellular stalk domain, one exon encoding a transmembrane domain and five exons encoding a cytoplasmic tail. The transcripts identified in this study span the exons that encode the leader peptide and the Ig D2 domain. Most of the transcripts match the exon content of the dicp3.1 gene from the reference genome. The dicp3.14756 transcript reflects an alternative 3' splice site that introduces a 108 nt deletion in the exon that encode the Ig D2 domain. The dicp3.14494, dicp3.14494, dicp3.15280 and dicp3.152813 transcripts possess an intron retention between the exons that encode the Ig domains without creating a PTC. However, similar transcripts display a two nucleotide deletion in this intron resulting in a PTC.

The dicp3.2 gene in the reference genome possesses one exon encoding a leader peptide, one exon encoding an Ig D1 domain, one exon encoding an Ig D2 domain and two exons encoding an extracellular stalk domain. The transcripts identified in this study span the exons that encode the leader peptide and the Ig D2 domain. Most of the transcripts possess an intron retention between the exons that encode the Ig domains that results in a PTC. The dicp3.21638 transcript possesses this intron retention as well as an 11 nt deletion in the middle of the Ig D1 domain that results in a PTC.

The dicp3.3 gene in the reference genome possesses one exon encoding a leader peptide, one exon encoding an Ig D1 domain, one exon encoding an Ig D2 domain and one exon encoding an extracellular stalk domain. The transcripts identified in this study span the exons that encodes the leader peptide and the Ig D2 domain. Most of the transcripts match the exon content of the dicp3.3 gene from the reference genome. The dicp3.31617 and dicp3.36028 transcripts possess a 6 nt insertion in the middle of the exon encoding the Ig D2 domain. Some transcripts reflect a 81 nt deletion in the middle of the exon encoding the Ig D1 domain. The dicp3.36027 transcript reflect a 102 nt deletion in
the middle of the exon encoding the Ig D1 domain. The \textit{dicp3.3^{380}} and \textit{dicp3.3^{368}} transcripts possess an intron retention that results in a PTC.

The \textit{dicp3.6} gene in the reference genome possesses one exon encoding a leader peptide, one exon encoding an Ig D1 domain, one exon encoding an Ig D2 domain, one exon encoding an extracellular stalk domain, one exon encoding a transmembrane domain and one exon encoding a cytoplasmic tail. The transcripts identified in this study span the exons that encode the leader peptide and the Ig D2 domain. The \textit{dicp3.6^{687}} transcript matches the exon content of the \textit{dicp3.6} gene from the reference genome. Most of the transcripts reflect an intron retention that results in a PTC.

\textit{Exon-intron architecture of newly described DICPs.}

Eleven new DICP sequences were identified in this study (\textit{dicp1.22} – \textit{dicp1.30}, \textit{dicp3.7} and \textit{dicp3.8}). The genes encoding \textit{dicp1.22} and \textit{dicp1.23} were identified in unplaced genomic scaffold NA310 (GenBank NW_003336703.1) permitting the definition of their genomic organization. For the remainder of these sequences, the genomic architecture of previous DICPs was used to predict their exon-intron organization (Fig. 7).

The \textit{dicp1.22} gene in scaffold NA310 possesses one exon encoding a leader peptide, one exon encoding an Ig D1 domain, one exon encoding an Ig D2 domain, three exons encoding an extracellular stalk domain, one exon encoding a transmembrane domain and three exons encoding a cytoplasmic tail. The transcripts identified in this study span the exon that encodes the leader peptide and the second exon that encodes the cytoplasmic tail. The \textit{dicp1.22^{3805}} and \textit{dicp1.22^{3615}} transcripts share the exon content of the \textit{dicp1.22} gene from scaffold NA310. The \textit{dicp1.22^{3798}}, \textit{dicp1.22^{3797}}, \textit{dicp1.22^{3800}} and \textit{dicp1.22^{3802}} transcripts as well as a RACE clone (6058) lack the first exon that encodes the extracellular stalk domain. Transcripts with an intron retention and a PTC were also identified.

The \textit{dicp1.23} gene in scaffold NA310 possesses one exon encoding a leader peptide, one exon encoding an Ig D1 domain, one exon encoding an Ig D2 domain, four exons encoding an extracellular stalk domain, one exon that would encode a transmembrane domain and four exons encoding a cytoplasmic tail. The transcripts identified in this study span the exons that encode the leader peptide and the potential transmembrane domain. Despite possessing an exon that would encode a transmembrane domain, a frameshift in \textit{dicp1.23} (as compared to other DICP transmembrane domain exons) results in this exon using a different reading frame that would not encode a transmembrane domain. The \textit{dicp1.23^{687}} and \textit{dicp1.23^{6648}} transcripts along with the RACE sequence (6648) match the exon content of the \textit{dicp1.23} gene from scaffold NA310. Other transcripts reflect i) an intron retention that results in a PTC, and ii) an alternative 5' splice site that yields a 41 nt deletion in the exon that encode the Ig D1 domain resulting in a PTC.
The genomic architecture of *dicp1.24* was predicted based on the sequence of similar DICP genes. The *dicp1.24* isoforms, *dicp1.24* and *dicp1.24* transcripts along with a RACE sequence (6134) are predicted to possess one exon encoding a leader peptide, one exon encoding an Ig D1 domain, one exon encoding a transmembrane domain and four exons encoding a cytoplasmic tail. The *dicp1.24* and *dicp1.24* transcripts likely reveal 24 and 36 nt deletions, respectively, in the sequence that encodes the Ig D1 domain. The *dicp1.24* transcript possesses additional sequence that may represent an exon that would encode a stalk domain. The *dicp1.24* transcript possesses this same stalk sequence as well as additional sequence that likely represent part of an intron with no PTC. This putative partial intron sequence is present in other transcripts within a longer putative intron sequence that does introduce a PTC, e.g. *dicp1.24* and *dicp1.24*.

The genomic architecture of *dicp1.25* was predicted based on the sequence of similar DICP genes. The *dicp1.25* and *dicp1.25* transcripts along with a RACE sequence (6655) are predicted to possess one exon encoding a leader peptide, one exon encoding an Ig D1 domain, one exon encoding an Ig D2 domain, one exon encoding an extracellular stalk domain, one exon encoding a transmembrane domain and four exons encoding a cytoplasmic tail. The *dicp1.25* transcript is predicted to possess a second exon contributing to the stalk domain.

The genomic architecture of *dicp1.26* was predicted based on the sequence of similar DICP genes. The *dicp1.26* transcript is predicted to possess one exon encoding a leader peptide, one exon encoding an Ig D1 domain, one exon encoding an Ig D2 domain, one exon encoding an extracellular stalk domain and at least one exon encoding a transmembrane domain.

The genomic architecture of *dicp1.27* was predicted based on the sequence of similar DICP genes. The *dicp1.27* transcript is predicted to possess one exon encoding a leader peptide, one exon encoding an Ig D1 domain, one exon encoding an extracellular stalk domain, one exon encoding a transmembrane domain and at least two exons encoding a cytoplasmic tail. The *dicp1.27* transcript likely reflects an intron retention resulting in a PTC.

The genomic architecture of *dicp1.28* was predicted based on the sequence of similar DICP genes. The *dicp1.28* transcript is predicted to possess one exon encoding a leader peptide, one exon encoding an Ig D1 domain, one exon encoding an extracellular stalk domain, one exon encoding a transmembrane domain and at least two exons encoding a cytoplasmic tail.

The genomic architecture of *dicp1.29* was predicted based on the sequence of similar DICP genes. The *dicp1.29* transcript is predicted to possess one exon encoding a leader peptide, one exon encoding an Ig D1 domain, one exon encoding an Ig D2 domain, two exons encoding an extracellular stalk domain, one exon encoding a transmembrane domain and at least two exons encoding a cytoplasmic tail.
The genomic architecture of *dicp1.30* was predicted based on the sequence of similar DICP genes. The *dicp1.30* transcript along with a RACE sequence (6670) are predicted to possess one exon encoding a leader peptide, one exon encoding an Ig D1 domain, four exons encoding an extracellular stalk domain, one exon encoding a transmembrane domain and three exons encoding a cytoplasmic tail.

The genomic architecture of *dicp3.7* was predicted based on the sequence of similar DICP genes. The *dicp3.7* transcript is predicted to possess at least one exon encoding an Ig D1 domain and one exon encoding an Ig D2 domain. Several transcripts were identified that likely reflect an intron retention resulting in a PTC.

The genomic architecture of *dicp3.8* was predicted based on the sequence of similar DICP genes. The *dicp3.8* and *dicp3.8* transcripts are predicted to possess at least one exon encoding an Ig D1 domain and one exon encoding an Ig D2 domain; however, these transcripts likely reflect an intron retention resulting in a PTC.
### Fig. S1. Dicp1.7
Alignment of the protein sequences encoded by a *dicp1.7* transcript identified in this study with a previously described allele (GenBank AFO8159.1). Identical residues are shaded black. Protein domains were predicted by SMART software. TM – transmembrane.

<table>
<thead>
<tr>
<th>Leader peptide</th>
<th>Ig D1 domain</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dicp1.7</strong></td>
<td></td>
</tr>
<tr>
<td>transcript_0996</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>YNGGFLFDQVREGQGQSSYVEFYYHVPFFTELEVESVYRRQRKTGDRTVYGLQENFETQPGDQYDRFYDTQV</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

### Fig. S2. Dicp1.8
Alignment of the protein sequences encoded by the *dicp1.8* and *dicp1.8* transcripts identified in this study with a previously described allele (*dicp1.8*) encoded by genomic scaffold 262 (GenBank HU_001878770.2). Identical residues are shaded black and structurally related residues are shaded grey. Protein domains were predicted by SMART software. TM – transmembrane.

<table>
<thead>
<tr>
<th>Leader peptide</th>
<th>Ig D1 domain</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dicp1.8</strong></td>
<td></td>
</tr>
<tr>
<td>transcript_0996</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>ESVYVESEAEESVYERPAH</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

Page 9 of 28
<table>
<thead>
<tr>
<th>Leader peptide</th>
<th>Ig D1 domain</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>D1</strong></td>
<td><strong>TM domain</strong></td>
</tr>
</tbody>
</table>

**Fig. S2. Diocl17** Alignment of the protein sequences encoded by multiple diocl17 transcripts identified in this study with a previously described diocl17 allele (GenBank NP.014245105.1). Identical residues are shaded black and structurally related residues are shaded gray. Protein domains were predicted by SMART software. TM – transmembrane.
Fig. S4. Dip1p.22 Alignment of the protein sequences encoded by multiple dip1p.22 transcripts with the protein encoded by the dip1p.22\textsuperscript{203C} allele present in genomic
scaffold NA10 (GenBank NAA_003556703.1). Identical residues are shaded black and structurally related residues are shaded gray. Protein domains were predicted by SMART software. TM = transmembrane.
### Fig. S5. Dip2.1 Alignment of the protein sequences encoded by multiple dip2.1 transcripts identified in this study with a previously described allele (GenBank AFO6165.1).

Identical residues are shaded black and structurally related residues are shaded gray. Protein domains were predicted by SMART software. TM – transmembrane.

<table>
<thead>
<tr>
<th>Leader peptide</th>
<th>Ig D1 domain</th>
<th>Ig D2 domain</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dip2.1 (AF06165.1)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>transcript 6006</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>transcript 6073</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>transcript 6074</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>transcript 6079</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>transcript 6590</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td><strong>Dip2.1 (AF06165.1)</strong></td>
<td>119</td>
<td></td>
</tr>
<tr>
<td>transcript 4066</td>
<td>101</td>
<td></td>
</tr>
<tr>
<td>transcript 4073</td>
<td>101</td>
<td></td>
</tr>
<tr>
<td>transcript 4074</td>
<td>101</td>
<td></td>
</tr>
<tr>
<td>transcript 4075</td>
<td>104</td>
<td></td>
</tr>
<tr>
<td>transcript 4590</td>
<td>104</td>
<td></td>
</tr>
<tr>
<td><strong>Dip2.1 (AF06165.1)</strong></td>
<td>237</td>
<td></td>
</tr>
<tr>
<td>transcript 4066</td>
<td>222</td>
<td></td>
</tr>
<tr>
<td>transcript 4073</td>
<td>222</td>
<td></td>
</tr>
<tr>
<td>transcript 4074</td>
<td>222</td>
<td></td>
</tr>
<tr>
<td>transcript 4075</td>
<td>231</td>
<td></td>
</tr>
<tr>
<td>transcript 4590</td>
<td>231</td>
<td></td>
</tr>
<tr>
<td>transcript 6150</td>
<td>231</td>
<td></td>
</tr>
</tbody>
</table>

Page 12 of 28
Fig. S6. *Dicp*1 Alignment of the protein sequences encoded by multiple *dicp*1 transcripts identified in this study with two previously described alleles (*dicp*1*FF* and *dicp*1*FF*) encoded by genomic scaffold 1962 (GenBank NT_001877662.2) and BAC CH73-34H11 (GenBank FP015862, respectively). Identical residues are shaded black and structurally related residues are shaded gray. Protein domains were predicted by SMART software. TM – transmembrane.
Fig. S7. Phylogenetic comparison of zebrafish DICP Ig domains. DICP Ig domains encoded by amplicons generated with a high fidelity DNA polymerase were aligned to all other DICPs and displayed as a phylogenetic tree. Sequences recovered with the high fidelity DNA polymerase are color coded: sequences that group with the DICP1 genes on chromosome 3 are in red text, sequences that group with the DICP2 genes on chromosome 14 are in green text and sequences that group with the DICP3 genes on chromosome 16 are in blue text. Additional sequences include new DICPs from Fig. 5 (indicated by black circles), DICP sequences found in the unmapped scaffold NA310 (GenBank NW_003336703.1: indicated by red triangles) and previously reported DICPs (Haire et al. 2012) including sequences predicted from genomic clones, BAC CH73-34H11 and BAC CH73-322B17 (GenBank FP629011 and FP015862: indicated by blue triangles). The percentage of replicate trees in which the associated taxa clustered together (bootstrap values) are shown next to the branches: values less than 50 are not shown. Note that there are two distinct groups of Dicp3.1 D1 domains (indicated by yellow shading).
Fig. S8. Zebrafish DICP protein sequences. Protein sequences for select newly identified DICPs were predicted by combining sequences obtained from RT-PCR and 3’ RACE. The different highlighted domains were predicted by SMART software (http://smart.embl-heidelberg.de/). Leader peptide – red; Ig D1 domain – green; Ig D2 domain – blue; transmembrane domain – yellow. Immunoreceptor tyrosine-based inhibition motifs (ITIM) and ITIM-like motifs (litm) were identified and highlighted with light and dark grey, respectively.
Fig. S9. Zebrafish DICP protein structures. Schematic representation of the predicted protein structures of DICPs newly identified in this study. Full-length protein sequences were predicted by combining RT-PCR and 3’ RACE sequences (Online Resource 3 – Fig. S8). DICPs for which the 3’ ends were not amplified are indicated with an asterisk (*). Red text for Dicp3.8 indicates that the transcript encoding this protein possesses a premature termination codon (PTC). D1 – D1 Ig domain; D2 – D2 Ig domain; TM – transmembrane region; ITIM – immunoreceptor tyrosine-based inhibition motif; ITIM-like sequence; Y – tyrosine residue.
**Fig. S10. Zebrafish chromosome 3 DICP haplotype dot-plot.** Dot-plot comparison of zebrafish reference genome GRCz10 scaffolds CTG10218 and NA310. The regions of highest similarity are between ccdc134 and ccdc134l and between gimap8 and gimap4l. The predicted ccdc134 and ccdc134l transcripts (GenBank XM_003198004 and XM_003201873) are 95% identical and their protein products are 100% identical. The predicted gimap8 and gimap4l transcripts (XM_001915280 and XM_001920324) are 91% identical and their protein products are 88% identical. Graphic generated with PipMaker software (Schwartz et al. 2000).
Fig. S11. DICP-MHC I linkage in grass carp and common carp. Conserved linkage of DICP and MHC Class I Z sequences in the genomes of zebrafish, grass carp, and common carp. Zebrafish sequences are from the reference genome and reported previously (Haire et al. 2012). Consensus BLAST results for each carp sequence (Online Resource 3 – Table S4), confirmed through phylogenetic analysis (Online Resource 3 – Fig. S12) are reported above gene symbols (triangles). Sequence identity numbers for each carp sequence are below each triangle and can be used to find further sequence information in Online Resource 3 – Fig. S13 and Table S4. The draft genome and gene annotations of grass carp (Ctenopharyngodon idellus) (Wang et al. 2015) were searched for linkage of DICP to MHC class I sequences. Un-linked scaffold CI01000243 was found to contain both DICP (blue triangles) and MHC Class I Z lineage (red triangles) sequences as well as at cod:134 (green triangle) revealing conserved synteny with the zebrafish chromosome 3 DICP locus. The genome (version 2) and gene annotations of the common carp (Cyprinus carpio) (Xu et al. 2014) also was searched and LG38 was found to encode DICP and MHC Class I Z lineage sequences as well as mhc1 (violet triangle) confirming conserved synteny with grass carp scaffold CI01000243. Note that one partial MHC Class I U lineage sequence (yellow triangle) was identified on common carp LG38 in proximity to a MHC Class I Z lineage sequence.
Fig. S12. Carp DICP and MHC I sequence comparisons. Sequence analyses of linked DICPs and MHC class I genes shown in Online Resource 3 — Fig. S11. Phylogenetic analysis of linked (A) DICPs and (B) MHC Class I molecules of zebrafish (Dre, black text), grass carp (Cld, red text), and common carp (Cyca, blue text). Zebrafish MHC Class I lineage sequences were described previously (Dirschwert and Yoder 2014; Dirschwer et al. 2014). Numbers in parentheses refer to the sequence identity (ID) numbers presented in Online Resource 3 — Fig. S11, Fig. S13 and Table S4. The common carp cDNA, Cyca_AB098477, was previously reported as a DICP (Haire et al. 2012). Gene annotations for additional carp sequences (Wang et al. 2015; Xu et al. 2014) are provided in Online Resource 3 — Table S4. In panel A the Ntr9 Ig V and I domains (Wei et al. 2007; Yoder 2009) were used as an outgroup. The percentage of replicate trees in which the associated taxa cluster together (bootstrap values) are shown next to the branches; values less than 50 are not shown. The D1 domains (C) and D2 domains (D) of newly identified DICPs from grass carp (Cld) and common carp (Cyca) are aligned with DICP domains previously reported from zebrafish and common carp (Haire et al. 2012). Positions that are 70% or greater identical are shaded in black and those that are structurally related are shaded in gray. Conserved residues characteristic of immunoglobulin domains are indicated by the IMGT numbering system above the alignments (Gudinelli et al. 2008) with conserved cysteines highlighted orange. Sequences used for analyses are provided in Online Resource 3 — Table S4 and Fig. S13.
Online Resource 3
Rodriguez-Nunez et al.

127
### Fig. S14. Tissue distribution of DICP transcripts

Distribution of the DICP transcripts obtained from the RT-PCR performed with tissues of adult zebrafish, myeloid and lymphoid cells, and embryos at different developmental stages. Green color: functional transcripts, blue color: functional transcripts and transcripts with PTCs, red color: transcripts with PTCs, grey color: transcripts not found.
Table S3. Distribution of DICP transcripts from individual zebrafish

<table>
<thead>
<tr>
<th>Line</th>
<th>Fish</th>
<th>dicp1.7</th>
<th>dicp1.8</th>
<th>dicp1.17</th>
<th>dicp1.22</th>
<th>dicp1.27</th>
<th>dicp1.28</th>
<th>dicp2.11</th>
<th>dicp3.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>TU</td>
<td>1</td>
<td>3797*</td>
<td>3797*</td>
<td>3800*</td>
<td>3803*</td>
<td>3806*</td>
<td>4066</td>
<td>4073</td>
<td>4445</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3810</td>
<td>3797*</td>
<td>3800*</td>
<td>3803*</td>
<td>3806*</td>
<td>4066</td>
<td>4074*</td>
<td>4445*</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3725</td>
<td>3755</td>
<td>3797</td>
<td></td>
<td></td>
<td>4066</td>
<td>4073*</td>
<td>4445</td>
</tr>
<tr>
<td>AB</td>
<td>1</td>
<td>3864</td>
<td>3860*</td>
<td>3863*</td>
<td></td>
<td></td>
<td>4066</td>
<td>4073*</td>
<td>4446</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3864</td>
<td>3863</td>
<td>3872</td>
<td>3875</td>
<td></td>
<td>4066*</td>
<td>4150*</td>
<td>4446</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3864</td>
<td>3810*</td>
<td>3863*</td>
<td></td>
<td></td>
<td>4066*</td>
<td>4150*</td>
<td>4446</td>
</tr>
<tr>
<td>EKW</td>
<td>1</td>
<td>3860*</td>
<td>3863*</td>
<td>3979</td>
<td>3980</td>
<td></td>
<td>4066</td>
<td>4073</td>
<td>4446</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3986</td>
<td>3860</td>
<td>3985</td>
<td>3991</td>
<td></td>
<td>4066</td>
<td>4590</td>
<td>4445*</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3994</td>
<td>3860*</td>
<td>3863*</td>
<td></td>
<td></td>
<td>4066</td>
<td>4073</td>
<td>4446</td>
</tr>
</tbody>
</table>

Bold text indicate sequences that were shared between different individuals within the same line. Italic text indicate sequences that were shared between individual fish from different genetic backgrounds. Asterisks (*) indicate possible alternative mRNA splicing variants from the same allele.
Table S4. Carp sequence information

<table>
<thead>
<tr>
<th>ID</th>
<th>Gene ID</th>
<th>Species</th>
<th>Origin</th>
<th>L1: Gene start/end</th>
<th>Strand (+/-)</th>
<th>Transcript Protein ID</th>
<th>ZF NCBI BLAST Hit</th>
<th>E-value</th>
<th>BLASTP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CI0100243_00018105_00032937</td>
<td>Grass Carp</td>
<td>Wang et al 2015</td>
<td>CI0100243: 32037-19195</td>
<td>-</td>
<td>CI0100243_00018105_00032937</td>
<td>NP_001274282.1</td>
<td>8E-182</td>
<td>M1c1z</td>
</tr>
<tr>
<td>2</td>
<td>CI0100243_00043972_00045237</td>
<td>Grass Carp</td>
<td>Wang et al 2015</td>
<td>CI0100243: 48257-43972</td>
<td>-</td>
<td>CI0100243_00043972_00045237</td>
<td>NP_001274282.1</td>
<td>1E-101</td>
<td>M1c1z</td>
</tr>
<tr>
<td>3</td>
<td>CI0100243_00062228_00062233</td>
<td>Grass Carp</td>
<td>Wang et al 2015</td>
<td>CI0100243: 62423-62229</td>
<td>-</td>
<td>CI0100243_00062228_00062233</td>
<td>NP_001274282.1</td>
<td>2E-29</td>
<td>M1c1z</td>
</tr>
<tr>
<td>4</td>
<td>CI0100243_00097327_001002005_00239961</td>
<td>Grass Carp</td>
<td>Wang et al 2015</td>
<td>CI0100243: 97327-102005</td>
<td>+</td>
<td>CI0100243_00097327_001002005</td>
<td>XP_005198052.1</td>
<td>4E-121</td>
<td>Cod134</td>
</tr>
<tr>
<td>5</td>
<td>CI0100243_00097327_001002005_00239961</td>
<td>Grass Carp</td>
<td>Wang et al 2015</td>
<td>CI0100243: 100305-239961</td>
<td>+</td>
<td>CI0100243_00097327_001002005</td>
<td>XP_009264449.1</td>
<td>0</td>
<td>Reh1</td>
</tr>
<tr>
<td>6</td>
<td>CI0100243_00074762_00075673</td>
<td>Grass Carp</td>
<td>Wang et al 2015</td>
<td>CI0100243: 274762-275673</td>
<td>+</td>
<td>CI0100243_00074762_00075673</td>
<td>NP_001245164.1</td>
<td>3E-91</td>
<td>Disq</td>
</tr>
<tr>
<td>7</td>
<td>CI0100243_00084609_000301564</td>
<td>Grass Carp</td>
<td>Wang et al 2015</td>
<td>CI0100243: 284609-301564</td>
<td>+</td>
<td>CI0100243_00084609_000301564</td>
<td>NP_001245164.1</td>
<td>8E-87</td>
<td>Disq</td>
</tr>
<tr>
<td>8</td>
<td>CAFS_CommonC_G_0756449</td>
<td>Common Carp</td>
<td>Xu et al 2014</td>
<td>LQ38: 510754-512055</td>
<td>-</td>
<td>CAFS_CommonC_T_000207632</td>
<td>AAH7656.1</td>
<td>0</td>
<td>M1c1z</td>
</tr>
<tr>
<td>9</td>
<td>CAFS_CommonC_G_0756449</td>
<td>Common Carp</td>
<td>Xu et al 2014</td>
<td>LG38: 517707-518230</td>
<td>-</td>
<td>CAFS_CommonC_T_000207643</td>
<td>CAA86732.1</td>
<td>4E-51</td>
<td>M1c1z</td>
</tr>
<tr>
<td>10</td>
<td>CAFS_CommonC_G_075685</td>
<td>Common Carp</td>
<td>Xu et al 2014</td>
<td>LG38: 706234-711126</td>
<td>-</td>
<td>CAFS_CommonC_T_000207648</td>
<td>XP_008264447.1</td>
<td>1E-179</td>
<td>Reh1</td>
</tr>
<tr>
<td>11</td>
<td>CAFS_CommonC_G_075439</td>
<td>Common Carp</td>
<td>Xu et al 2014</td>
<td>LG38: 216196220-21618149</td>
<td>+</td>
<td>CAFS_CommonC_T_000207670</td>
<td>NP_001245164.1</td>
<td>2E-75</td>
<td>Disq</td>
</tr>
</tbody>
</table>

b. BLAST consensus
Reference List


APPENDIX D: CHAPTER 3 SUPPLEMENTAL MATERIAL

GROUP ASSIGNMENTS OF IG DOMAINS

>100_35098 group I
LICYLAAKGLMSIVSFAGAALMIKCEHPQHQTGTKTRYICKESSAGCSGEWRKNGDVSVDED
TRAGVLMVFVRELKADAGTSCGVDKSEYTERFTQQLKLVRH
>100_96084 group I
KVGSTVIGYSAGAMIKCKHPQHKTKTRYICKESSAGCSGEWRKNGDVSVDDDSRAGVLM
VFVRELKADAGTSCGVDKSEYTERFTQQLQIRH
>101_199795 group I
EGVSTVTGSSGAALMIKCEHPQHKTNTRYICKESSAGCSSEWMDVDDDSRAGVLM
MVVFRELKADAGTSCGVDKSEYTERFTQQLKLVRH
>101_180232 group I
KVGSTVIGYSAGAMIKCEHPQHKTNTRYICKESSAGCSGEWMKNGDVSVDSEDTRAGVLM
MVVFRELKADAGTSCGVDKSEYTERQQLQIRH
>101_164300 group I
KVGSTVIGYSAGAMIKCEHPQHKTNTRYICKESSAGCSSEWMDVDDDSRAGVLM
VFVFRELKADAGTSCGVDKSEYTERFTQQLQLIRH
>101_234000 group I
KVGSTVIGYSAGAVTMIKCEHPQHKNTKTRYICKESSAGCSGEWRKNGDVSEDSRAGVLM
VFVFRELKADAGTSCGVDKSEYTERFTQQLKLVRH
>100_80735 group I
KVGSTVIGYSAGAVTMIKCEHPQHKNTKTRYICKESSAGCSGEWRKNGDVSEDSRAGVLM
VFVFRELKADAGTSCGVDKSEYTERFTQQLQIRH
>100_97451 group II
GDFHSSAAFIGSVKTLCTIHREKNTVHICENHDKICQNIVSSNKGRQISDSTDGVFTVS
ISNVSVRDAVGYWCGAKTRDTHLTFLSITIKLELT
>100_917701 group II
GDFHSSAAFIGSVKTLCTIHREKNTVHICENRDKICQNIVSSNKGRQISDSTDGVFTVIS
SNVSVRDAVGYWCGAKTRDTHLTFLSITIKLELT
>101_238674 group II
AIIINKTLGGGDNICQTPEEHJAYFCKEDDNCYRNIRQNPFSAAEVPSNPAEFVSIOSNVS
RDAGVYWGCAGTRETHLTFLSITIKLELT
>100_100200 group II
AIIINKTLGGGDNICQTPEEHJAYFCKEDDNCYRNIRQNPFSAAEVPSNPAEFVSIOSNVS
RDAGVYWGCAGTRETHLTFLSITIKLELT
>101_180630 group II
AKYPKVKINSVHLGDVNISCQMAKKMKVYFCKEDNQRCQIMRASKEPQMTPGERVYVNSVRDAGVYWCGAETRDTHTLTFISLTTKIQLSVI
>101_152962 group II
EKYPKVNKTAPGEVISIQIEQVKNYYFCKEDDHKSCQIMRASEEPQLTHEEPVGNEERVFRVSRDAGVYWCGAETRDTHTLTFISLTTKIQLSVIDN
>100_97059 group II
AKYPKVMKESVQLDGEVKISCQILEEQVKhFSCQKTRASEEPQRTPAGSVGESERVTFTSINNSVSRDAGVYWCGAETRDTHTLTFISLTTKIQLSII
>100_35499 group II
GKYQEVMNESVQLGGEVKISCQIPEEQNKNYFCKEDDHKSCQIRRASEDPQLTAPAGSVGSEEVFTFTSINNSVSRDAGVYWCGAETRDTHTLTFISLTTKIQLSVID
>100_81106 group II
GEYPKVMNESVHLGGEVKISCQIPEQQKNYFCKEDDHKSCQIRRASEDPQLTAPAGSVGSEEVFTVISSNASVSRDAGVYWCGAETRDTHTLTFISLTTKIQLSVI
>100_233516 group II
AENPKVKMTESVHLGGEVKISCQIPEEQKNYFCKEDDHKSCQIRRASEDPQLTAPAGSVGEERVFTVISSNASVSRDAGVYWCGAETRDTHTLTFISLTTKIQLSVI
>101_190349 group II
GKYPKMNESVHLGGEVKISCQIIIEEQKVLFYKEEYNQSCQIRRASEDPQLTAPAGSVGEERVFTVISSNASVSRDAGVYWCGAETRDTHTLTFISLTTKIQLSVI
>100_143042 group III
LHYVPFPCSSAVSPVVRREGESAQIFICPYSIYQSKSLCKGCSTRDTRPLDETVEEKERTLTLDVDTSASVTFTGTITGLTAEDAGKYWCAMTLDELYLTHLMVIIKQ
>101_165048 group III
SVHHHLCKVCVCLCAVSPVVRREGESAQIFICPYSIYQSKSLCKGCSTRDTRPLDETVEEKERTLTLDVDTSASVTFTGTITGLTAEDAGKYWCAMTLDELYLTHLMVIIKQ
>101_180999 group III
SVHHHLCKVCVCLCAVSPVVRREGESAQIFICPYSIYQSKSLCKGCSTRDTRPLDETVEEKERTLTLDVDTSASVTFTGTITGLTAEDAGKYWCAMTLDELYLTHLMVIIKQ
>100_190580 group III
SVHHHLCKVCVCLCAVSPVVRREGESAQIFICPYSIYQSKSLCKGCSTRDTRPLDETVEEKERTLTLDVDTSASVTFTGTITGLTAEDAGKYWCAMTLDELYLTHLMVIIKQ
>101_239889 group III
SVHHHLCKVCVCLCAVSPVVRREGESAQIFICPYSIYQSKSLCKGCSTRDTRPLDETVEEKERTLTLDVDTSASVTFTGTITGLTAEDAGKYWCAMTLDELYLTHLMVIIKQ
>100_35874 group III
SVHHHLCKVCVCLCAVSPVVRREGESAQIFICPYSIYQSKSLCKGCSTRDTRPLDETVEEKERTLTLDVDTSASVTFTGTITGLTAEDAGKYWCAMTLDELYLTHLMVIIKQ
>100_72237 group III

133
LCLAGQGLCFKVGCTGGSLLFKCMYSNQKSSDVENDGAYFSDRLNQFADDYEVKLNVK
>101_18000 group III
ASSNSVAPVGGSVTFCSESHFLAFLSTSKYFCRDSCAEENILIQSETGENPTRRRGRYSLNEDKGD
FVTVIADRLSDDSGTYCVAVERLLLKDTYTLTHVT
>101_70547 group III
KITSVQVQAYIRGRFVFKCKFPOKQIKEKNNKKFMRPQAESQKIMLDEQINQWIIHHDNVHMYDNTE
GHLKVFISLDSSANEGETYRCGVNIDQDDLYTEVKN
>101_916583 group III
KITSVQVQAYIRGVSIFKCKFLQKFKMRRASEQKIFMLDEQINQWIIHHDNVHMYDNTE
GHLKVFISLDSSANEGETYRCGVNIDDDRLYETKN
>101_70000 group III
FLPKISSIAITEYEGSITFRTSYTEFGTSYKFSRSDTFFSEKLVESTSESNRWTHGRFAIFDN
TSAHVIIATNLNTMEDSGSYTVGVDTLLPDYDAEIQLTVMR
>101_70900 group III
FLPKISRAITEYEGSITFSTYLESLETNSYKFSRSGYFCEKLVESTSESNRWTHGRFAIFDN
STHLIITTLNTEDSGSYLIVAVDTLWPDEETIELTVMR
>101_116900 group III
GPRIGTVRGYSVAGVHIINSYAIIHKSQKHVCKTGNCISGGAEEHSGRSFISIHDNR
SAGLRLVFIRELNVQDSGEYRIIVHAEPSDFSEDFDL
>101_106914 group III
GPMGKVRYGSGRRNIKIKIRYRDKQKNHVFVCKTGNCISGGAEEHSGRSFISIHD
NRSAAGLLRVFIRELNQVQDSGGRYRIIVRHSDYSSDFSEFELVTD
>101_122700 group III
GPRIGTVRGYSVAGVHIINSYAIIHKSQKHVCKTGNCISGGAEEHSGRSFISIHD
RSADDLRLVFIRELNQVQDSGGRYRIIVHAEPSDFSEFELVTD
>101_35000 group III
RASVCACTITVTAAGEGGAEAVNCYPPGGGYETSYKYFYIIPYKDSTLWQLSYYGESPVFNGRT
LKDHDHAKFLFVRTDLQMNDAEGYNAAGWRDSSRIQLNVRG
>101_41233 group III
RASVCACTITVTAAGEGGAEAVNCYPPGGGYETSYKYFYIIPYKDSTLWQLSYYGESPVFNGRT
LKDHDHAKFLFVRTDLQMNDAEGYNAAGWRDSSRIQLNVRG
>101_99797 group III
VLSSISVTGYSGTVTITCKYDEEYETNKEYFCKGKWSQSEDQIRTEMKHHWIRGFSFSLFD
TTAAVFTVTIRNLQDSDIHYHGKIFKHAFDHGTKVNLK
>100_27164 group III
VLSSISVTGYSGTVTITCKYDEEYETNKEYFCKGKWSQSEDQIRTEMKHHWIRGFSFSLFD
TTAAVFTVTIRNLQDSDIHYHGKIFKHAFDHGTKVNLK
>100_12125 group III
VLSSISVTGYSGTVTITCKYDEEYETNKEYFCKGKWSQSEDQIRTEMKHHWIRGFSFSLFD
TTAAVFTVTIRNLQDSDIHYHGKIFKHAFDHGTKVNLK
>100_106412 group III
VMSSISVTGYSGTVTITCKYDEEYETNKEYFCKGKWSQSEDQIRTEMKHHWIRGFSFSLFD
TTAAVFTVTIRNLQDSDIHYHGKIFKHAFDHGTKVNLK
>100_94382 group III
VVSSISVTGYSGTVTITCKYDEEYETNKEYFCKGKWSQSEDQIRTEMKHHWIRGFSFSLFD
TTAAVFTVTIRNLQDSDIHYHGKIFKHAFDHGTKVNLK
>100_21767 group III
VVSSISVTGYSGTVTITCKYDEEYETNKEYFCKGKWSQSEDQIRTEMKHHWIRGFSFSLFD
TTAAVFTVTIRNLQDSDIHYHGKIFKHAFDHGTKVNLK
>100_12125 group III
LLRCCVTVGVLGSISSVMGILGGVMTCCNIDRVETYNTAKYFCKKAMPQFWCyclk
INSHKWWKDFTSLYDDTAAAIFVTVTIRNLTKQDESIYQCAVDTSGIDQYEVNLK
>101_106412 group III
VLLRCCVTVGVLGSISSVMGILGGVMTCCNIDRVETYNTAKYFCKKAMPQFWCyclk
INSHKWWKDFTSLYDDTAAAIFVTVTIRNLTKQDESIYQCAVDTSGIDQYEVNLK
>101_12215 group III
VLLRCCVTVGVLGSISSVMGILGGVMTCCNIDRVETYNTAKYFCKKAMPQFWCyclk
INSHKWWKDFTSLYDDTAAAIFVTVTIRNLTKQDESIYQCAVDTSGIDQYEVNLK
136
>100_665 group III
SAVSVNISVTGYSGGGVTITCKYDRQYKKNTKYFCGKQKFLSCSLIKTEAKSDNKWVQ
KDRYSLFHTNTAADFTVTIRNLTVWDSGTYYCGVEVYGLDPSTSITKVNLIITA
>100_50284 group III
CWFVGVGYSSMGYSGVGVMITCKYDEQYKTSAYFCFEKGWISPCSELIRINSEEKWVQS
GRFSLFDNRTAAGLVTITRNLTEQDSTGYYHCGVETYGPDPGTGKVNLEVIT
>101_114829 group III
CWFVGVGYSSMGYSGVGVMITCKYDKGFETYPKYFCEQWISPCSELIRTEINSEEKWV
QGRFSLIDNKTASGLTVTIRDLTEQDSTGYYHCGVETYGPDPGTGKVNLEVIT
>101_121901 group III
GVVSSMVGYSSMGYSGVGVMITCKYDRGFETYPKYFCEQWISPCSELIRTEINSEEKWVQSRFS
LIDNTAALFTVTITRDLTEQDSTGYYHCGVETYGPDPGTGKVNLEVIT
>100_42540 group III
CWFVGVGYSSMGYSGVGVMITCKYDRGFETYPKYFCEQWISPCSELIRTEINSEEKWV
QGRFSLIDNKTASGLTVTIRDLTEQDSTGYYHCGVETYGPDPGTGKVNLEVIT
>101_73716 group III
VLSSISVTGYSGGGVTITCKYAAQYKTSAKYFCGKQKFLSCSLIKTEAKSDNKWVQGRFSL
DNTNAAVNFVTIRNLTEDSGTYYHCGVDQISDIGITKVNLEVIT
>101_86327 group III
VLCCFDVSGLSSISVTGYSGGGVTITCKYDEQYKTSAKYFCGKQKFLSCSLIKTEAKSDNKWV
KKDCKFLYDNTTAAVFVTITRNLTEKDSSTGYYHCGVDHITYHTDIYTVNLEVIT
>101_162000 group III
LISDQSFSLTGYSGGSLMLDSGRSSSSNSLYLLKHHTHGWWGLIELSRODWSDGKFTL
YLKNKRMLFLFIRDLTPQDSGRYTRTARLSAVEVMEIHLKVIEGESFQPITFQKNT
>101_176600 group III
LISQGIFSLSVTGYSGGSLMLDSGRSSSSNSLYLLKHHTHGWWGLIELSRODWSDGKFTL
YLKNKRMLFLFIRDLTPQDSGRYTVGLSAFDEMEIHLKVIE
>100_95834 group IV
SCCKDPTVRMVKSGETAASNCKYSQSRFNSVIFKAQQISSIEELEYTRWNSSGAKFTISNDREK
NFSSVKITAVNSFDGAIVLCGIHQYHQLYSYNYNINTVQLHV
>100_34704 group IV
SCCALSKKVKVNSGETAVFSCDYSRSHYDAKVLFKERQNLIEEVIHSTGTRNKERFSISDDDR
RNNFNSRISTKVPDDGTVYLCGVSIRSSIIITLYTVL
>101_189550 group IV
SCCKVPTSMVNNGETASFNCESSQSQINHIIKAKDYIELIYNWSAYNRSVSVSDRPD
QKKLFSVRITAVTPDGGVYFCQFVRNSHYSSISDIDIVLHV
>101_163700 group IV
SCCKRLKRQEVDLRRTGTFCEFSNHINDEKLVFKEKNSIDMLNSTWRMSTFGLINMID
KQHFKVRINAVTDDPGGTVLCGVWVKNHSHYYINTVHLYL
>101_179800 group IV
SCCKRMKRQEVNLGRGTIGCFEYSQNHIDKVVFKKEKNSIDEINSTWKKVRFSWNRFDKR
RHFNVITAVKADDGGTYLCGVWINEHYSSYYINTVHLQII
>101_152345 group IV
SCCGSSKTLVKSREAAATCFEYSNLQTSGGKTFKTNDIFDEVSTTYTWDKERSIBDD
RQRRLLSISITAVTADDGGVYLCGVWVKNHSHYYINTVHLQII
>101_199422 group IV
SCCDVSKRTVNSGETAVFSCYSHINQIHAKVIFERDSVKSIIHTTARIIEGRVHMSDD
GQRNVLSISITAVTADDGGVYLCGVWVRENSYNSTLHHTILHV
>100_80338 group IV
GGAACAGCGTCATACCGCCCCATAGCTCATTTTAAAGACAAAATGCAGCCATATGTTTCTCTGGCTATATAATTGATGATCTACAGAAATTTATACAGGGGTTTGTTTTCAGATTGAGCCTAATAAAATTGACATATGCTGTTATACATTGTCTACATTATGGCTATATAGTAACTGTCACAGACGAATGTGCTGCTATCCTGTATATTCTATAGTGCTGTTATATCAATTGTATATTGCAACTATCAGAAGCAGGTTTCTATATAGCTGTTATATCTATCTATGCTGCTACTTGTTGTTTGTGTCATCTCTATGCCATTACTCTGAGAGTAAATGCAGATAAATTCTCTGTATGTTTGTATATGTGAACTGACAATAAAGCGGTCACTTTAAAT

5'/3' RACE NILT TRANSCRIPTS

>LOC100537163 confirmed via full length 3' RACE
ACACACGTTTTCTATATCAACACTCCAGATCAGATGTACACGGAGCTGAACACCAGACA
GACTGACCTGTATACACAGTCTTACCACCAGGATTCTGCTCAACACGAGGCCATCTATCAG
TATTTGACCAAAAAATAATAAAAAAAAAAAAAAANNNNNAAAAACA

>LOC101882285 confirmed via full length 3' RACE

AGAACTCAAGTGAATAATGAAGATCTTCCTCGTCTTCTTTACTTCCTACCTGATCTCAGGT
CAGAGCTTTTCTCTTTCTGCTNCNGGAATACTCTGGAAGGAGGGCCTGATTCCTGGAGATGCTGGG
AAACTCTGCTTCATGCTGCTAAATACATGGTTAAAACNCAACCCNTGGAGAACCAC
AATAGACNTCAAGAAAAACTAGTAATATGGACTGCTGAAGGAGTTTCACCTTATTTGCA
ACACNCAATGGAAACCTCGTAGTCTCTACAGGGATCTGCAATCNCNCAAGATGCTGGAGATA
TACAGAGGTAGGTTCTTGCACAAGTGTATATTGATATAGTGAATGTTAAAGGAAGG
TTAGCAATTATTATGTGTGTGTGATGCNTCCNTCTTTAAGAACGGCTCAATTTGACAGGTATTA
ACATTTTTGTGTTGGTAAATGTAATGACAGATCAGATCTGCTGCTTCATGCTGCTCAAGAAGATG
ATAATGGAATATGGAGAAAAGCTGCTCTTACGATTCCCCAGAATCATATTGAT
GATGCTAAAGTGTATTCCAAAGAGAGACAAAAACTTTAATTTAAAGAGGTGATNCNCTCCNC
TGGCCCAAGAAAAAAAAAAAAAAAAAAA
APPENDIX E: CHAPTER 4 SUPPLEMENTAL MATERIAL

A. TLR1 ENSL0CG00000012910

B. TLR2 ENSL0CG00000018220

C. TLR3 ENSL0CG00000013826

D. TLR4 ENSL0CG00000003751

E. TLR5 ENSL0CG00000018000

F. TLR7/8a/8b ENSL0CG00000009977/ENSL0CG00000009982/ENSL0CG00000009990

G. TLR9a ENSL0CG00000014202

H. TLR9b ENSL0CG00000014430

I. TLR18 ENSL0CG00000007992

J. TLR21/22 ENSL0CG00000018316
APPENDIX F: GENERATION OF TARGETED AND HERITABLE CRISPR-INDUCED MUTATIONS IN ZEBRAFISH*

*This protocol is adopted from (Varshney et al. 2015). See this publication for complete details.

CRISPR gRNA sequence was chosen from previously identified NILT genomic sequences using Zebrafish Genomics track on the genome browser at [http://genome.ucsc.edu/](http://genome.ucsc.edu/) (LaFave et al. 2014). The target site (GGATGCAGGAACATAC) is located in the second exon of the single identified activating NILT (see chapter 3 for details) in the Zv9 zebrafish reference genome (Fig. 1) and was placed in Oligo 1 (Table 1). The targeted exon encodes the single immunoglobulin (Ig) domain of the targeted NILT protein.

![Activating NILT 24553](image)

**Figure 1.** Exon-Intron boundaries and alternative splicing of an activating zebrafish NILT gene. Alternative splicing which results in wild-type transcript with a stop codon before the ITAM are represented by with red lines and octagons. Approximate location of CRISPR target is marked by an orange bar. Abbreviations are as follows: Ig = immunoglobulin, STP = serine, threonine and proline, TM = transmembrane and ITAM = immunotyrosine activating motif.

**Table 1.** Oligonucleotides utilized in this protocol.

<table>
<thead>
<tr>
<th>Oligonucleotide Type</th>
<th>Sequence (5' -&gt; 3')</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oligo 1</td>
<td>TAAATACGACTCACTATAGGATGCAGGAACATACGTTTTAGAGCTAGAAATAGC</td>
</tr>
<tr>
<td>Oligo 2</td>
<td>AAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAAC</td>
</tr>
<tr>
<td>Forward</td>
<td>CTTGTCCATCATTTGACCACAT</td>
</tr>
<tr>
<td>Reverse</td>
<td>CTGTGATAGTGACTCGTG</td>
</tr>
</tbody>
</table>

Table 1. Oligonucleotides utilized in this protocol.
Protocol:

*dsDNA template construction for transcription of gRNA*

Promega GoTaq (cat#M300)

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume (µl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H₂O</td>
<td>11.5</td>
</tr>
<tr>
<td>5x GoTaq Buffer</td>
<td>5</td>
</tr>
<tr>
<td>1.25mM dNTPs</td>
<td>4</td>
</tr>
<tr>
<td>10µM Oligo 1</td>
<td>2</td>
</tr>
<tr>
<td>10µM Oligo 2</td>
<td>2</td>
</tr>
<tr>
<td>Promega GoTaq</td>
<td>0.5</td>
</tr>
<tr>
<td>Total Volume</td>
<td>25</td>
</tr>
</tbody>
</table>

Incubate in thermocycler, 2 minutes at 98 °C, 10 minutes at 50 °C, 10 minutes at 72 °C

*gRNA synthesis*

New England Biolab’s HiScribe T7 High Yield RNA Synthesis Kit (cat#E2040S)

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume (µl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H₂O</td>
<td>4</td>
</tr>
<tr>
<td>10x Reaction Buffer</td>
<td>2</td>
</tr>
<tr>
<td>100mM ATP</td>
<td>2</td>
</tr>
<tr>
<td>100mM GTP</td>
<td>2</td>
</tr>
<tr>
<td>100mM UTP</td>
<td>2</td>
</tr>
<tr>
<td>100mM CTP</td>
<td>2</td>
</tr>
<tr>
<td>Annealed Oligo</td>
<td>4</td>
</tr>
<tr>
<td>T7 RNA Polymerase Mix</td>
<td>2</td>
</tr>
<tr>
<td>Total Volume</td>
<td>20</td>
</tr>
</tbody>
</table>

Incubate at 37 °C for >4hr (up to overnight). Add 1µl DNase and incubate 37 °C for 20 minutes

*gRNA purification*

Follow the manufacturer’s protocol for Zymo RNA Clean and Concentrate 5 Kit. Elute in 15µl H₂O.

*Microinjection*

gRNA was quantified using a NanoDrop2000 (Thermo Fisher Scientific) by measuring the absorbance at 260 nm. gRNA was diluted to 50ng/µl with water, distributed into 2µl aliquots and stored at -80°C. 50 picograms of gRNA were microinjected into the yolk of 1-cell stage embryos with 1/10 dilution of Cas9 protein (New England Biolabs cat# M0386M) phenol red (Sigma cat# P-0290) into the NHGRI-1 line of zebrafish (LaFave et al. 2014) using the following recipe:

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume (µl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 ng/µl gRNA</td>
<td>1</td>
</tr>
<tr>
<td>Cas9 Protein</td>
<td>1</td>
</tr>
<tr>
<td>0.5% Phenol Red</td>
<td>1</td>
</tr>
<tr>
<td>Ultrapure water</td>
<td>7</td>
</tr>
<tr>
<td>Total volume</td>
<td>10</td>
</tr>
</tbody>
</table>
**Genomic DNA preparation from zebrafish embryos**

Add 90µl of 50mM Sodium Hydroxide to PCR tubes containing a single embryo (1 to 2 days post-fertilization) (multiple embryos can also be pooled).

Incubate 10 minutes at 98°. Let cool to room temp.

Add 10µl of 1M Tris-HCL pH 7.5

Use 2µl as PCR template.

**Genomic DNA preparation from adult zebrafish**

Adult zebrafish are anesthetized by immersion in 150 mg/mL Tricaine in aquarium water. A portion of the caudal (tail) fin is removed and placed into a PCR tube using surgical scissors and forceps.

Genomic DNA is prepared following the manufacturer’s protocol of the Extract-N-Amp kit (Sigma).

Extracted DNA is diluted 1:10 (e.g. 10ul into 90µl H₂O) and 2µl is used as PCR template.

**PCR to assess CRISPR activity**

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>H₂O</td>
<td>11.8µl</td>
</tr>
<tr>
<td>5x GoTaq Buffer</td>
<td>5µl</td>
</tr>
<tr>
<td>1.25mM dNTPs</td>
<td>4µl</td>
</tr>
<tr>
<td>10uM Forward Primer</td>
<td>1µl</td>
</tr>
<tr>
<td>10uM Reverse Primer</td>
<td>1µl</td>
</tr>
<tr>
<td>Template</td>
<td>2µl</td>
</tr>
<tr>
<td>Promega GoTaq</td>
<td>0.2µl</td>
</tr>
<tr>
<td>Total volume</td>
<td>25µl</td>
</tr>
</tbody>
</table>

Thermocycler conditions: 95° for 1 minute followed by 40 cycles of 95°C for 30 seconds, 54°C for 30 seconds and 72°C for 30 seconds. A final elongation of 72° for 10 min was followed by a hold at 4°C.

PCR products were submitted to Eton Biosciences (Research Triangle Park, NC) directly for Sanger sequencing with the forward primer.
Preliminary results:

Sixteen zebrafish injected zebrafish embryos which survived to adults were found to contain mutations near the target site using primers listed in Table 1. Adults appeared phenotypically normal. A female zebrafish (F₀) with CRISPR-induced mutations was crossed with a wild-type NHGRI to produce F₁ progeny. Eight F₁ embryos at two days post-fertilization were sacrificed and sequenced and a single mutation was found to be heritable (Fig. 2). This mutation causes a premature stop codon in the predicted amino acid sequence and should ablate the function of the NILT protein. Two heterozygous F₁ individuals, when crossed, are expected to produce 25% homozygous knockout F₂ offspring. Future work will focus on establishing a homozygous line of knockout zebrafish from screening the F₂ generation.

Figure 2. CRISPR induced mutations in an activating NILT. CRISPR reagents were designed and injected according to well established methods (Varshney et al. 2015). A. Sequence chromatogram of F₁ CRISPR induced mutation. B. Sequence chromatogram of non-injected, wild-type NHGRI-1. The location where Cas9 causes a double stranded break, proto-adjacent motif (PAM), is outlined by a red box. The predicted amino acid sequence of the modified gene includes a premature stop codon.
References
