

Evolution of Innate Immune Protein Complexes, Toll-like Receptor 4 and Calprotectin, in Early
Vertebrates and Zebrafish

by

Kona Nikole Orlandi

A dissertation accepted and approved in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in Biology

Dissertation Committee:

David Garcia, Chair

Michael Harms, Advisor

Karen Guillemin, Core Member

Laura McKnight, Core Member

Raghuveer Parthasarathy, Institutional Representative

University of Oregon

Spring 2024

© 2024 Kona Nikole Orlandi

DISSERTATION ABSTRACT

Kona Nikole Orlandi

Doctor of Philosophy in Biology

Title: Evolution of Innate Immune Protein Complexes, Toll-like Receptor 4 and Calprotectin, in Early Vertebrates and Zebrafish

The innate immune system is our first line of defense against pathogens as well as our interface with our commensal microbiota. Toll-like receptor 4 (TLR4) and calprotectin are two innate immune proteins that are tightly associated with inflammatory disorders. Zebrafish (*Danio rerio*) has been successfully used to model the human innate immune system, but TLR4 and calprotectin models have not been developed because of their significant divergence in humans and zebrafish. Here, we set out to reveal the evolutionary and functional relationships between human and zebrafish TLR4 and calprotectin. We used phylogenetic analyses to define the evolutionary relationships between homologous proteins and characterized their immune functions in cell-based assays. We found that an antagonist of human TLR4 is a potent agonist for zebrafish TLR4, but when tested in live fish there was no difference in immune stimulation. We further investigated the evolutionary origin of this change in ligand specificity and determine that TLR4 in the cyprinid order of fish likely convergently evolved sensitivity to LPS. Our characterization of zebrafish proteins homologous to human calprotectin also suggest that the zebrafish proteins do not share functional similarities to calprotectin during the immune response. We conclude that although humans and zebrafish share many immune system characteristics, the TLR4 and calprotectin immune responses are not directly comparable.

This dissertation includes previously published and unpublished co-authored material. Supplement includes multiple sequence alignments and phylogenetic trees for TLR4 and MD-2.

CURRICULUM VITAE

NAME OF AUTHOR: Kona Nikole Orlandi

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
University of California, Santa Cruz

DEGREES AWARDED:

Doctor of Philosophy, Biology, 2024, University of Oregon
Bachelor of Science, Biochemistry and Molecular Biology, 2016, University of
California, Santa Cruz

AREAS OF SPECIAL INTEREST:

Protein Evolution
Molecular Biology
Biochemistry
Immunology

PROFESSIONAL EXPERIENCE:

Graduate Student Researcher, University of Oregon, September 2018 – June 2024

Graduate Teaching Assistant, University of Oregon, September 2018 – June 2019

Research Intern, J. Craig Venter Institute for Environmental Genomics, 2018

Post-baccalaureate Researcher, UCSC, 2017

Course Assistant in Eukaryotic Molecular Biology, UCSC, 2017

GRANTS, AWARDS, AND HONORS:

Raymond-Stevens Fellowship, University of Oregon, 2024

National Institutes of Health (NIH) Molecular Biology and Biophysics Training Grant
Appointment T32, University of Oregon, 2019-2022

Institute of Molecular Biology Best Poster Award, University of Oregon, 2019

Promising Scholar Award, University of Oregon, 2018

B.Sc. with Honors in the Major of Biochemistry and Molecular Biology, UCSC, 2016

Blue and Gold Opportunity Plan Scholar, UCSC, 2012-2015

PUBLICATIONS:

Chisholm LO, Orlandi KN, Phillips SR, Shavlik MJ, Harms MJ. “Ancestral Reconstruction and the Evolution of Protein Energy Landscapes” *Annual review of biophysics*, 10.1146/annurev-biophys-030722-125440. 22 Dec. 2023, doi:10.1146/annurev-biophys-030722-125440.

Erin A. Garza, Vincent A. Bielinski, Josh L. Espinoza, Kona Orlandi, Josefa Rivera Alfaro, Tayah M. Bolt, Karen Beeri, Philip D. Weyman, and Christopher L. Dupont. “Validating a Promoter Library for Application in Plasmid-Based Diatom Genetic Engineering.” *ACS Synthetic Biology* 2023 12 (11), 3215-3228. DOI: 10.1021/acssynbio.3c00163

Orlandi KN*, Phillips SR*, Sailer ZR, Harman JL, Harms MJ. “Topiary: Pruning the manual labor from ancestral sequence reconstruction.” *Protein Sci.* 2023 Feb;32(2):e4551. doi: 10.1002/pro.4551. PMID: 36565302; PMCID: PMC9847077.

McKnight, L. E., Crandall, J. G., Bailey, T. B., Banks, O. G., Orlandi, K. N., Truong, V. N., Donovan, D. A., Waddell, G. L., Wiles, E. T., Hansen, S. D., Selker, E. U., McKnight, J. N. Rapid and inexpensive preparation of genome-wide nucleosome footprints from model and non-model organisms. *STAR Protocols* 2: 2 (2021).

Orlandi, K. and McKnight, J. Bulky Histone Modifications May Have an Oversized Role in Nucleosome Dynamics. *BioEssays* 42:1 (2020).

ACKNOWLEDGMENTS

First, I would like to thank my advisor, Dr. Mike Harms, for his unwavering support of my professional growth and his enthusiasm for scientific exploration. Mike's commitment to facilitating student development as scientists and individuals has been an inspiration to me and has allowed me to gain skills and knowledge in several scientific disciplines throughout the course of my graduate work, even in areas that Mike is not an expert. He continually provided both the challenge and support I needed to become a well-rounded scientist.

At the time of joining Mike's lab, there were many reasons for him to feel uncertain about taking me on as a trainee: I had not previously worked in his lab, he had already taken on two students for that year, and it was the peak of a global pandemic. Without Mike and his lab's optimism and faith, I would not have completed my doctorate at the University of Oregon. I will forever be grateful to them for taking me under their wing. I am very lucky to have experienced such a supportive lab community throughout the last three and a half years. Thank you, to all past and present Harms lab members, especially to those whom I worked most closely with, Corinthia Brown, Sophia Phillips, José Sanchez-Borbón, Lauren Chisholm, and Dr. Jon Muyskens.

I want to express my gratitude to my collaborators and the university staff who made it possible for me to accomplish this work. For guidance on zebrafish experiments and interpretation of the results, I greatly appreciate Dr. Karen Guillemin, Dr. Cathy Robinson, Dr. Raghu Parthasarathy, Dr. Julia Ngo, Piyush Amitabh, Jonah Sokoloff, Patrick Horve, Dr. Adam Fries, and Rose Sockol. Thank you so much for teaching me and supporting me on this journey. I also greatly value the discussions and advice I received for experiments in bacteria from Dr. Jarrod Smith, Dr. Cathy Robinson, Dr. Karen Guillemin, and Dr. Melanie Spero. Thank you to Stu Johnson for keeping the institute running and for all of your spot-on song reference sign offs!

And finally, my deepest gratitude goes to my past and present dissertation advisory committee members, Dr. David Garcia, Dr. Alice Barkan, Dr. Eric Selker, Dr. Karen Guillemin, Dr. Raghu Parthasarathy, and Dr. Laura McKnight, for your encouragement, critical questions, and guidance throughout the years.

Special recognition for my success in graduate school is due to the late Dr. Jeff McKnight and members of his lab, Dr. Laura McKnight, Dr. Orion Banks, Dr. Thomas Bailey, Vi Truong, Dr. Drake Donovan, and Abigail Vaaler. I came to the University of Oregon because Jeff believed in me. I had a difficult time adjusting to graduate school, but I found my confidence when I joined their exceptionally compassionate and quirky lab community. In addition to his lessons in yeast genetics, biochemistry, and chromatin remodeling, Jeff taught me how important and effective it is to lift up and create constructive space for others who face discrimination in any form, and especially in academia. Jeff was a true role model to me. I will always miss him and wonder what could have been. Thank you also to Dr. Hinrich Boeger who helped me get started in molecular biology research, connected me with Jeff, and supported my career development into graduate school.

I want to acknowledge Dr. David Garcia and Dr. Alice Barkan for their resolve to support me through the difficult transition when I needed to decide how to continue my career foreseeing Jeff's passing. Without this sincere source of assurance, guidance, and support no matter my decision, I would have been lost.

Through all the highs and lows of my graduate experience, I always received love and reassurance from the friends I made in Eugene. Thank you to Dr. Ethan Shaw, Dr. Julia Ngo, Madelyn Green, Travis Heeren, Dr. Jordan Munroe, Dr. Elizabeth Vargas, Zac Bush, Dr. Michael Shavlik, Dr. Bryce LaFoya, Nan Provant, Zac Provant, and everyone on the Specific Heat soccer

team for being there for me from the beginning and always creating a warm and upbeat atmosphere. The sense of camaraderie among these folks has brought me so much happiness. Our community has grown to include more dear friends who have brought abundant joy. Thank you, Emily Dennis, Max Horrocks, Sophia Phillips, Acadia DiNardo, Molly Shallow, Jared Freedman, William Crow, Sam Horst, Sofia Carlson, Hannah Wilson, and Brenden Campbell. I will miss our community, our soccer games, floating, camping, yard games, tailgating, brewing, skiing, bonfires, surfing, derby day, and even football watch parties!

I could not have come to graduate school without the support of my parents, Arlene and Bruce Orlandi. They have always valued my education and my dreams and made sacrifices to help me continue pursuing my career aspirations, even when that career means I can't pay my own bills until I'm 30. They are my biggest source of encouragement and backing, my first-choice vacation destination, the roots that keep me grounded, and the best parents anyone could ever hope for. I also want to thank my sister, Makayla Orlandi, and my extended family for always being a phone call away.

Finally, I owe very special acknowledgements to my sweet angel pup, Lucy, and my fiancé, Ethan. Lucy came into my life when I was 18 and was by my side through 11 years full of life's transitions. We had a profound bond and a tremendous trust in one another that I did not know was possible. Her companionship throughout graduate school encouraged me to get outside every day, make friends, and remember to appreciate the little things. I wish she could be here now to start the next chapter of life with me, but she did leave me in the capable hands of her favorite person, Ethan. Thank you, Ethan, for caring so much about Lucy and me from the very first moment we met. Your kindness, generosity, and thoughtfulness have always amazed me. I am so grateful that you have been, and will always be there to console me when I am

feeling down, celebrate with me when things are great, and cook food for me when I forget. It has been so special sharing our graduate training together and I can't wait to spend the rest of my life growing and learning with you.

This investigation was supported in part by the National Institute of General Medical Sciences through Molecular Biology and Biophysics Training Grant Appointments, 5T32GM007759-41 & 42, to me and by a grant, R01-GM146114, to Dr. Michael Harms at the University of Oregon. This work was also partially funded by a Raymond-Stevens fellowship awarded to me.

DEDICATION

I dedicate this work to the ones who started this journey with me but aren't here to celebrate its culmination: My Sweet Little Lucy Lady, Auntie Lori Mason, Uncle Greg McMurrough, Grandpa Burl Bradbury, and Prof Jeff McKnight.

TABLE OF CONTENTS

Chapter	Page
I. INTRODUCTION	19
Human microbial communities impart complex influences on our health	20
The innate immune system identifies specific microbes via TLR4	20
Calprotectin mediates TLR4-induced inflammation and fights infections	23
Zebrafish is a powerful model organism for studies of the host-microbe interface.....	24
Evolutionary differences confound the use of model organisms to study human biology	25
Protein evolution is a framework for mapping between model organisms and humans	26
Summary of contributions:	28
Bridge to Chapter II	29
II. TOPIARY: PRUNING THE MANUAL LABOR FROM ANCESTRAL SEQUENCE RECONSTRUCTION.....	31
Abstract.....	32
Introduction.....	33
Overview of Ancestral Sequence Reconstruction.....	35
Define the Problem	35
Construct a Sequence Dataset.....	37
Sequence Alignment	38
Infer a Maximum Likelihood Gene Tree	38

Reconcile the Gene Tree to the Species Tree	40
Reconciliation: The Special Case of Microbial Genes	41
Reconstruct Ancestors	42
Evaluate Results.....	43
The Topiary Pipeline.....	44
Software Design.....	45
Stage 1: Seed to Alignment.....	47
Initial Dataset Construction	47
Redundancy Reduction, Quality Control, and Alignment	49
Alignment	52
Stage 2: Alignment to Ancestors	53
Infer the Evolutionary Model.....	53
Build a Maximum Likelihood Gene Tree	54
Reconcile Gene and Species Tree.....	55
Reconstruct Ancestors	56
Branch Supports.....	57
Output	57
Protocol.....	58
Construct a Seed Dataset	58
Run the Seed-to-Alignment Pipeline	58
Inspect and Edit Alignment	59
Perform the Ancestral Inference	60
Checking Gene/Species-Tree Reconciliation	61

Selecting Ancestors.....	65
On Black Boxes	67
Pipeline Validation.....	68
Conclusion	71
Bridge to Chapter III.....	71
III. TOLL-LIKE RECEPTOR 4 EVOLUTION OF LPS SPECIFICITY IN EARLY VERTEBRATES AND DIVERGENCE IN ZEBRAFISH	72
Abstract.....	73
Introduction.....	74
Results.....	79
Zebrafish TLR4/MD-2 is potently activated by tetra-acyl LPS <i>in vitro</i>	79
A subset of teleost fish evolved a functionally necessary MD-2 C-terminal peptide.....	83
Live zebrafish exhibit reduced immune response to lipid IVa compared to <i>E. coli</i> LPS	87
Zebrafish and human TLR4 evolved from an ancestor with low LPS sensitivity	92
Initial investigations into possible CD14 functional homologs in fish.....	97
Discussion.....	102
Zebrafish TLR4 ohnologs might play important physiological roles.....	103
Possible functional roles of the zebrafish MD-2 C-terminal peptide	103
Is TLR4 used in the zebrafish innate immune response to Gram-negative bacteria?	105

Further probing ancestral complexes will help us understand TLR4 ligand responses	106
Did zebrafish lose CD14 as a mechanism to avoid LPS toxicity?.....	108
Materials and Methods.....	109
Ancestral sequence reconstruction.....	109
Plasmids	110
Cell culture and transfection conditions	111
Oral microgavage of LPS.....	113
Brain tectum microinjection of LPS	113
Bridge to Chapter IV.....	114
IV. ZEBRAFISH DO NOT HAVE CALPROTECTIN	115
Abstract.....	116
Introduction.....	117
Results.....	120
Zebrafish s100a10b is only distantly related to human S100A9 and S100A9	120
Chromosome placement indicates a shared origin but complicated evolution of homologous human and zebrafish S100s	121
Human calprotectin and zebrafish s100 protein sequences have low sequence identity.....	122
Single cell RNA sequencing datasets mining points to candidate zebrafish s100 proteins expressed similar to calprotectin	124
Recombinant zebrafish s100 proteins fold and interact with calcium	126

Zebrafish s100s do not exhibit nutritional immunity like calprotectin	128
Zebrafish s100s do not exhibit proinflammatory activity like S100A9	130
Discussion	132
Materials and Methods	134
Protein purification	134
Far-UV circular dichroism and fluorescence spectroscopy	137
Nutritional immunity assay	138
Proinflammatory activity assay	140
Bridge to Chapter V	142
V. SUMMARY AND CLOSING REMARKS	144
REFERENCES CITED	146

SUPPLEMENTAL FILES

PDF: JOHN WILEY AND SONS LICENSE TERMS AND CONDITIONS

FASTA: TLR4 SEQUENCE ALIGNMENT

PDF: TLR4 PHYLOGENETIC TREE

FASTA: TLR4 ANCESTOR SEQUENCES

FASTA: MD-2 SEQUENCE ALIGNMENT

PDF: MD-2 PHYLOGENETIC TREE

FASTA: MD-2 ANCESTOR SEQUENCES

LIST OF FIGURES

Figure		Page
1.	Figure 2.1. Define the ancestral reconstruction problem	36
2.	Figure 2.2. Ancestral sequence reconstruction has six main steps	39
3.	Figure 2.3. Summarized topiary ASR pipeline	46
4.	Figure 2.4: Topiary redundancy reduction and quality control	51
5.	Figure 2.5. Example trees at each step in the ASR calculation	62
6.	Figure 2.6. Graphs for evaluating ancestor quality	63
7.	Figure 2.7. Validation of the topiary pipeline.....	64
8.	Figure 3.1. Current knowledge of the evolution of TLR4/MD-2 LPS specificity.	75
9.	Figure 3.2. Revealing differences in LPS specificity between human and zebrafish TLR4 complexes	81
10.	Figure 3.3. Zebrafish ohnologs tlr4bb and tlr4al do not respond on their own to LPS variants in vitro	83
11.	Figure 3.4. The C-terminal peptide of fish MD-2 is necessary for zebrafish TLR4 signaling	86
12.	Figure 3.5. Zebrafish response to (L6)-LPS-EK and (L4)-lipid IVa challenge via microgavage and hindbrain injection	90
13.	Figure 3.6. Characterization of TLR4 activity for reconstructed early vertebrate ancestors and modern fish and amphibian sequences.....	93
14.	Figure 3.7. Hybridizing the human TLR4 transmembrane and TIR domains to other species extracellular domains does not reveal function.....	96
15.	Figure 3.8. Zebrafish TLR2, CD180/MD-1, and human transferrin do not	

rescue the zebrafish TLR4 response to lipid IVa in the absence of CD14	99
16. Figure 4.1: Phylogenetic analyses reveal there is no calprotectin ortholog outside of amniotes	123
17. Figure 4.2: Structural comparisons of human and zebrafish S100s.....	129
18. Figure 4.3: Zebrafish s100s do not exhibit nutritional immunity activity like human calprotectin.....	130
19. Figure 4.4: Zebrafish s100s do not exhibit the pro-inflammatory characteristics of human S100A9	132

LIST OF TABLES

Table	Page
1. Table 2.1: Example seed dataset.....	59
2. Table 2.2: Protein families used to validate the topiary pipeline.....	64
3. Table 4.1: Single-cell RNAseq profiles for zebrafish s100s syntenic to human calprotectin.....	126

CHAPTER I

INTRODUCTION

In this dissertation, I will describe my contributions to three bodies of work concerning protein evolution, with emphasis on the divergence between innate immune proteins in humans and zebrafish. Understanding the process of protein evolution is imperative for interpreting discoveries in biology, especially those gained from model organisms.

Chapter II is a published co-authored manuscript describing a bioinformatic phylogenetics tool for ancestral protein sequence reconstruction that I helped develop and make available to the public. Sophia Phillips and I were co-lead authors of this manuscript, Zachary Sailer and Joseph Harman are co-authors, and Michael Harms was the project and software development lead and major writing contributor.

Chapter III describes unpublished findings, including material contributions from José Sánchez-Borbón, Cathy Robinson, and Corinthia Brown, and important insights from Michael Harms, Sophia Phillips and Karen Guillemin. This chapter covers my investigation of the evolutionary history and fluctuating ligand specificity of Toll-like receptor 4 complexes in zebrafish, other modern vertebrates, and their ancestors.

Finally, Chapter IV is a manuscript soon to be submitted evaluating whether zebrafish have convergently evolved a functional homolog of calprotectin, an important biomarker of inflammation severity in human patients. Michael Harms contributed experimental guidance and oversaw the writing of this work. The following introductory sections will provide context and a through line for these studies, with more specific information about each topic in their respective chapters.

Human microbial communities impart complex influences on our health

All animals exist with populations of microbes. At birth, humans are colonized with microorganisms including bacteria, fungi, viruses, archaea, protozoa, and helminths on every surface exposed to the environment¹⁻³ and even within tissues^{4,5}. In the body, human cells are outnumbered by microorganisms, which are most abundant in the gut.⁶⁻⁸ The various interactions between host and microbe may be commensal (beneficial for microbe and no effect on host), mutualistic (both parties benefit), or pathogenic (microbe causes disease). The field describes most symbiotic microbes as ‘commensals’ though they generally provide host benefits, and some can even become opportunistic pathogens.

Symbiotic microbes contribute positively to host health by supporting host physiology, immunological development, metabolism, resistance to infection by pathogenic microorganisms, as well as other essential functions.^{3,9-16} Mutualistic microbiota and their collective genomes (microbiome) provide us with genetic and metabolic capacities we have not been required to evolve on our own,¹⁷ and *vice versa*. However, pathogenic microbes have continuously provoked the evolution of both the human immune system and our commensal microbiota to keep harmful microbes at bay.

The innate immune system identifies specific microbes via TLR4

A major field of study is in the communication between us and our microbiome.¹⁸ How does the host influence microbial community structure? How do microbes influence our development, health, mood, and behaviors? How does our immune system distinguish commensals from pathogens? One of the primary points of contact between these microbial communities and animal physiology is the innate immune system. The innate immune system is

our first line of defense against infection and disease. It consists of barriers like the skin and mucosa, effector cells that destroy pathogens and mediate the immune response, secreted antimicrobial molecules that inhibit pathogen growth, release of proinflammatory or anti-inflammatory signals, and cellular receptors that sense microbial and infectious signals both outside and inside of cells and subsequently activate the immune response.¹⁹

The cellular receptors of the innate immune system are known as pattern recognition receptors (PRRs). They are transmembrane receptors expressed on innate immune cells like macrophages, neutrophils, dendritic cells, natural killer cells, mast cells, basophils, and eosinophil.²⁰ A major function of PRRs is to sense highly conserved microbe-associated molecular patterns (MAMPs) and then transduce this signal across cell membranes to activate immune responses.²¹ One of the most well-studied MAMPs is lipopolysaccharide.

Lipopolysaccharide (LPS) is a major structural feature of Gram-negative bacteria outer membranes and acts as a permeability barrier.²² LPS was first discovered by Richard Pfeiffer in 1892 as the causative agent of sepsis and was coined 'endotoxin' because it was associated with the insoluble part of bacterial cells rather than secreted like other bacterial toxins known at the time.²³ Sepsis is a life-threatening condition that arises when the body's inflammatory response to a Gram-negative bacterial infection causes damage to its own tissues and organs.²⁴

A century after the discovery of endotoxin (LPS), it was determined that Toll-like receptor 4 (TLR4) is the binding partner that discriminates LPS from host lipids and transduces signals across the membrane.²⁵⁻²⁷ Ligand-induced activation of TLR4 triggers signaling cascades that upregulate the expression and secretion of cytokines and other proinflammatory proteins.²⁸⁻³¹ Because of this, TLR4's ability to recognize LPS was identified to be the underpinning of sepsis.²⁵

LPS molecules are present in almost all Gram-negative bacteria and are structurally diverse. The architecture of LPS consists of three domains. 1) The O-antigen is a repeating hydrophilic oligosaccharide that is structurally varied even within a single bacterium and fulfills a range of functions, depending on bacterial lifestyles. 2) The hydrophilic core oligosaccharide linking the other two domains. 3) A hydrophobic lipid A moiety containing a glucosamine disaccharide that can have one or two phosphates and supports 4-8 fatty acid acyl chains of varying lengths.^{29,32,33} The lipid A portion forms the outer leaflet of the outer membrane of Gram-negative bacteria and is therefore highly conserved across species. The lipid A moiety is recognized by TLR4 and thus confers to LPS its proinflammatory characteristics.³²

Structural and functional analyses show that the most proinflammatory form of lipid A has two phosphate groups and six fatty acyl groups with 12-14 carbon chains, which are generally purified from *Escherichia coli* and *Salmonella* strains.³⁴ Lipid A with either more or less fatty acyl chains, longer chains, or a single phosphate group are typically less active and can act as antagonists of toxic LPS.^{34,35}

To recognize LPS, TLR4 forms a complex with accessory protein MD-2. It is well-established that MD-2 forms the LPS-binding pocket in the TLR4/MD-2 complex and confers LPS specificity.^{31,36,37} These hypo-acylated and hypo-phosphorylated LPS variants tightly bind human TLR4/MD-2 in a non-productive fashion, inhibiting other LPS molecules from activating the complex.³⁷⁻⁴² However, slight variations in the LPS-binding pocket of mouse MD-2 and TLR4 permits mouse TLR4 signaling with several of these LPS variants.⁴³⁻⁴⁶ The foundation for this is still being explored.

One way that several commensal bacteria contribute to the flourishing microbiota in the human gut relies on their coevolution with human TLR4/MD-2. Many members of our

microbiota produce LPS, but not all of this LPS is immunogenic.⁴⁷⁻⁵⁰ For example, it has been shown that several members of the order Bacteroidales, which are the dominant Gram-negative bacteria in healthy human gut microbiomes, produce potently antagonistic tetra- and penta-acylated forms of LPS that can silence TLR4 signaling for the entire microbial community.⁴⁷ These findings raise questions of how symbiotic relationships evolve. How does this symbiosis affect our health? Did these bacteria adapt to exploit a blind spot in their host's immune surveillance?⁵¹⁻⁵⁶ Did humans lose the ability to recognize these bacteria due to advantageous selective pressure? Does this symbiosis impact our ability to fight infections from other Gram-negative bacteria? Could we use these LPS variants to suppress TLR4-induced sepsis in human patients?^{57,58}

Calprotectin mediates TLR4-induced inflammation and fights infections

PRRs also recognize endogenously produced danger-associated molecular patterns (DAMPs) including molecules released from dying cells and damaged tissue such as extracellular DNA, RNA, and proteins.⁵⁹ S100A8, S100A9, and their heterodimer state known as 'calprotectin' are all DAMPs recognized by TLR4.^{60,61} S100A8 and S100A9 proteins are multifunctional regulators of the immune response. They exist as homodimers but predominantly form the more stable heterodimeric calprotectin complex.⁶²⁻⁶⁴ S100A8, S100A9, and calprotectin have been shown to play several intracellular roles in calcium-dependent signaling, microtubule reorganization, and arachidonic acid metabolism.⁶⁵⁻⁶⁸ During an immune response, these proteins are released to the extracellular space where they exert several proinflammatory and antibacterial roles.^{61,69-72}

S100A8 and S100A9 homodimers and heterodimers released during the immune response locally activate TLR4 and amplify inflammatory responses.^{73,74} This is essential for successful pathogen-clearing but can be detrimental in excess. One regulatory mode imposed on this proinflammatory activity is the proteolytic sensitivity of homodimer and heterodimer states.⁷⁵⁻⁸⁰ In the presence of calcium, which is expected at a site of inflammation, two heterodimers form a heterotetramer (S100A8/S100A9)₂. This calcium-induced tetramerization inhibits proinflammatory activity, conferring the complex protease resistance.^{76,81,82} This regulatory method is essential for preventing excessive inflammatory amplification that can lead to sepsis, autoimmune disorders, and cancer.⁷⁴

Calprotectin is also antibacterial in both heterodimer and heterotetramer states. The hexahistidine site created at the heterodimer interface can chelate essential transition metal ions like zinc, manganese, and iron: this inhibits bacterial growth during an infection.⁸³⁻⁹⁰ As a stable complex, calprotectin is of particular interest in medicine because it can be used as a non-invasive biomarker for inflammation severity in addition to its roles in immunity.⁹¹⁻⁹⁴

Zebrafish is a powerful model organism for studies of the host-microbe interface

The host-microbe interface is complex and ever evolving. At any given time, each person has a unique population of microbes whose ecological architecture is influenced by interactions between microbes, our genetics, and everything we do like what we eat, who/what we encounter, our hygiene, sudden lifestyle changes, and even the buildings in which we live and work. Because these systems are so complicated, we need to leverage model organisms to facilitate our investigations of host-microbe relations. Indeed, almost everything we know about TLR4 and calprotectin comes from studies in mice.

Danio rerio, the zebrafish, is an outstanding model organism to study vertebrate innate immunity and host-microbe interactions. For the first few weeks of life the zebrafish has a fully functional innate immune system but has not yet developed its adaptive immunity, permitting investigations solely on the innate immune response without genetic modifications.⁹⁵ Larval zebrafish are optically transparent at least up to 5-6 days old which facilitates live imaging of fluorescently tagged proteins and microbes. Also, in the first week of life zebrafish survive using the nutrients in their yolk, and do not need food. This makes it relatively simple to generate gnotobiotic fish (fish with a defined microbiome) by sterilizing the chorion and water the fish develop in, and then introducing only desired microbes.⁹⁶ Moreover, genetically tractable tools, ease of rearing, mating, and maintenance, and large clutch sizes make zebrafish ideal for the research setting.⁹⁷

A major hurdle that all model organism research faces is the confounding variable of evolution. Not all genes, proteins, or physiology have direct matches with other species. This often obscures our ability to learn about human biology when performing studies in zebrafish and other model organisms. Zebrafish exhibit a great amount of homology to the mammalian immune system including a high degree of conservation in inflammatory proteins, effector cell types, and receptors like the Toll-like receptors.⁹⁸ However, humans and zebrafish have experienced diverse pressures necessitating the evolution of modified immune defense techniques unique to each species, many of which are actively being explored.

Evolutionary differences confound the use of model organisms to study human biology

The long and independent evolutionary divergence of humans and zebrafish complicates efforts to compare their biology. The most recent common ancestor of *Homo sapiens* and *Danio*

rerio was a bony fish that lived roughly 430 million years ago.⁹⁹ For context, the species *Homo sapiens* and the mouse *Mus musculus* diverged roughly 87 million years ago from a placental mammal.⁹⁹ Together with the knowledge that life underwater comes with a different set of evolutionary pressures, we can expect great divergence in zebrafish and human proteins, more so than between mouse and human.

Overall, approximately 70% of human genes have at least one obvious zebrafish ortholog.¹⁰⁰ There are three orthologs of human TLR4 in the zebrafish and a single MD-2 gene.¹⁰¹ Zebrafish also share the S100 family with all other vertebrates, but do not possess the proinflammatory calgranulin genes.¹⁰² Zebrafish have a heightened tolerance to LPS challenge compared to mammalian species, requiring much higher doses to be lethal, but exhibit similar inflammatory responses, like immune cell migration and transcriptional changes.^{103–105} It was proposed that this LPS tolerance might be an evolutionary advantage for organisms in intimate contact with microbes, such as that experienced in aqueous environments.¹⁰⁵ Previous work shows that although one of the zebrafish TLR4/MD-2 complexes can activate a low-level immune response to LPS *in vitro*, fish with an MD-2 loss-of-function mutation do not exhibit the classic drastic protection against LPS toxicity observed in mice.¹⁰¹ Their findings suggest that zebrafish have a low-sensitivity TLR4/MD-2 complex that confers LPS responsiveness to a specific set of immune cells, but that there are likely other pathways involved in the zebrafish immune response to LPS. In Chapter III of this dissertation, I revisit this hypothesis.

Protein evolution is a framework for mapping between model organisms and humans

My work has been done under the premise that an explicitly evolutionary lens can allow us to understand how to map studies of the innate immune system between zebrafish and

humans. For example, the low immune response of zebrafish to LPS challenge could reflect the behavior of TLR4/MD-2 in the last common ancestor of humans and zebrafish. If so, results in zebrafish provide a baseline from which to understand human innate immunity: a comparison of the human and zebrafish immune systems would reveal how humans gained high sensitivity.

In contrast, if the last common ancestor of humans and zebrafish had a high response to LPS, it would indicate that the evolutionary change happened on the zebrafish lineage. In this scenario, the low response is not a baseline from which high human activity evolved; rather, low activity is an evolutionary innovation specific to the zebrafish. This would mean a comparison of human and zebrafish immunity reveals how zebrafish lost sensitivity, not how humans gained sensitivity.

Resolving this scenario requires understanding the evolutionary history of the genes encoding innate immune proteins. Key questions include: Which human innate immune genes have orthologs in zebrafish? If not all of these genes are present, was this due to differential gain or loss? Do the genes themselves have the same basic functions?

Answering these questions requires a phylogenetic approach, where we trace the evolutionary history of individual genes. Here, we use computational tools that test whether genes from different organisms are homologous, and whether they are orthologous (arose by speciation), paralogous (arose by gene duplication), or took some more complex path of speciation, duplication, and loss. This allows us to know whether we are comparing the same gene (orthologs) or different genes (paralogs, ohnologs, etc.) between humans and zebrafish. Another powerful evolutionary approach involves tracing how the sequences of proteins have changed over time.¹⁰⁶ This can reveal what sequence changes correlate with what functional

changes, and thus allow us to isolate specific functional transitions and better resolve when and how functional transitions happened.

Summary of contributions

Phylogenetic approaches, while powerful, are also technically difficult. They require multiple software packages, working together in concert, as well as maintaining a database of sequences used for the study. In Chapter II, I describe work I did with members of the Harms lab to develop *topiary*, a software tool for ancestral sequence reconstruction. This tool automates a wide variety of tasks in evolutionary inference, and thus allowed me to study the evolution of TLR4 and MD-2 in bony vertebrates.

In Chapter III, I show an example in which I use our ancestral reconstruction software pipeline, along with careful *in vitro* and *in vivo* characterization, to better understand the evolutionary history of the zebrafish TLR4/MD-2 complex. I found that the zebrafish TLR4/MD-2 complex has a higher specificity for tetra-acylated LPS variants rather than the hexa-acylated variant typically used to assess TLR4 function. This suggested to me that previous zebrafish experiments done with LPS with 6- or 7-acyl chains may have missed important biological insight of TLR4 in the zebrafish immune response. I was also curious to know what structural differences between human and zebrafish TLR4/MD-2 impart this alteration in specificity. Identifying the molecular basis for these functional differences would be quite difficult, though, considering human and zebrafish TLR4 and MD-2 only share 39% and 26% identities, respectively.

With this new information, I investigated the ligand specificity of previously uncharacterized modern species and ancestral TLR4/MD-2 complexes linking zebrafish and

human evolution. I used *topiary* to infer the sequences of ancestral TLR4 and MD-2 proteins and resurrected them in the lab for functional characterization. I also investigated the functional role of a C-terminal peptide unique to a subset of teleost fish MD-2s that appears to be positioned to influence LPS binding and TLR4 activation. Importantly, I tested the hypothesis that zebrafish would exhibit a stronger immune response when challenged with tetra-acylated LPS *in vivo* compared to hexa-acylated LPS. Overall, my results confirm that TLR4 complexes from zebrafish and organisms more closely related to them have low sensitivity for LPS in our *in vitro* system and that the zebrafish *in vivo* response to LPS is not directly comparable to human biology.

In Chapter IV I describe a study that hinges on identifying specific innate immune proteins in zebrafish. I investigated the claim that zebrafish might have convergently evolved a calprotectin-like protein which was previously suggested in the literature.^{107,108} S100A9 and S100A8 do not have orthologs in zebrafish, so I used insights from phylogenetic studies and transcriptional response datasets to identify homologous candidates. I found that the zebrafish protein classified as “calprotectin”, and all other candidate proteins tested, do not exhibit canonical antibacterial or proinflammatory functions of calprotectin. I conclude from this study that when developing zebrafish models of innate immunity, it is necessary and prudent to account for evolutionary divergence and provide functional characterizations of the proteins considered.

Conclusion

In conclusion, my work has revealed that the genes and mechanisms responsible for innate immune recognition and response to pathogenic bacteria have evolved independently

between vertebrate species and have significantly diverged between humans and zebrafish. These results challenge the assumption that innate immune recognition universally relies on specific germline-encoded receptors, but supports that the response to microbial products, like LPS, is an ancestral trait.¹⁰⁹ Although the innate immune roles of TLR4 and calprotectin cannot be directly related between humans and zebrafish, there is much we can still learn about the evolution of protein complexes and innate immunity through further studies to determine how the zebrafish defends itself from infection and injury.

Bridge to Chapter II

Many studies of protein evolution make use of ancestral sequence reconstruction to infer the sequences and structures of proteins from ancestral organisms and compare their function to present day proteins. This technique requires copious amounts of time and stitching together many complex bioinformatic tools, which necessitates expert knowledge of coding, phylogenetics, evolutionary models, and the tools available for your specific application. Previous graduate students in Harms lab, including Dr. Zach Sailer, Dr. Joseph Harman, and Dr. Andrea Loes, had worked with Dr. Harms to develop an ancestral sequence reconstruction pipeline to study the TLR4 complex. When Sophia Phillips and I began our own investigations into the evolution of S100A9 and TLR4, we set out to further develop this pipeline into a widely available and accessible tool. With Dr. Michael Harms as the software development and project administration lead, we created *topiary*: a publicly available ancestral sequence reconstruction pipeline integrating several well-established software packages with handy time-saving scripts to reduce the workload, accompanying explanations of what the software is doing at each step, and guidance on how to tailor and interpret your study.

CHAPTER II

TOPIARY: PRUNING THE MANUAL LABOR FROM ANCESTRAL SEQUENCE

RECONSTRUCTION

*This chapter contains previously published co-authored material. See supplement for copyright terms and conditions.

Orlandi KN, Phillips SR, Sailer ZR, Harman JL, Harms MJ. (2022). Topiary: Pruning the manual labor from ancestral sequence reconstruction. *Protein Science*. 32:e4551.

Author contributions: K. N. O.: Conceptualization (equal); data curation (equal); methodology (equal); software (equal); validation (equal); visualization (equal); writing – original draft (lead); writing – review and editing (lead). S. R. P.: Conceptualization (equal); data curation (equal); investigation (equal); methodology (equal); software (equal); validation (equal); visualization (equal); writing – original draft (lead); writing – review and editing (lead). J. L. H.: Conceptualization (equal); methodology (equal); software (supporting); validation (equal); writing – review and editing (equal). Z. R. S.: Conceptualization (equal); methodology (equal); software (equal); validation (equal); writing – review and editing (supporting). M. J. H.: Conceptualization (equal); funding acquisition (lead); investigation (equal); methodology (equal); project administration (lead); software (lead); visualization (equal); writing – original draft (equal); writing – review and editing (equal).

ABSTRACT

Ancestral sequence reconstruction (ASR) is a powerful tool to study the evolution of proteins and thus gain deep insight into the relationships among protein sequence, structure, and function. A major barrier to its broad use is the complexity of the task: it requires multiple software packages, complex file manipulations, and expert phylogenetic knowledge. Here we introduce *topiary*, a software pipeline that aims to overcome this barrier. To use *topiary*, users prepare a spreadsheet with a handful of sequences. *Topiary* then: (1) Infers the taxonomic scope for the ASR study and finds relevant sequences by BLAST; (2) Does taxonomically informed sequence quality control and redundancy reduction; (3) Constructs a multiple sequence alignment; (4) Generates a maximum-likelihood gene tree; (5) Reconciles the gene tree to the species tree; (6) Reconstructs ancestral amino acid sequences; and (7) Determines branch supports. The pipeline returns annotated evolutionary trees, spreadsheets with sequences, and graphical summaries of ancestor quality. This is achieved by integrating modern phylogenetics software (Muscle5, RAxML-NG, GeneRax, and PastML) with online databases (NCBI and the Open Tree of Life). In this paper, we introduce non-expert readers to the steps required for ASR, describe the specific design choices made in *topiary*, provide a detailed protocol for users, and then validate the pipeline using datasets from a broad collection of protein families. *Topiary* is freely available for download: <https://github.com/harmslab/topiary>.

INTRODUCTION

Since it was first proposed in 1963, ancestral sequence reconstruction (ASR) has become a well-established method to study the evolutionary history of modern-day proteins.^{110,111} Studies of ancestral proteins uniquely reveal sequence features that are important for function and stability that cannot be readily identified from studies on modern-day proteins alone.¹⁰⁶ For example, ASR has been used for crystallographic and kinetic studies on ancestral proteins when their modern-day descendants were not amenable to crystallization¹¹², for bioengineering enzymes that are both thermally stable and catalytically active using ancestral enzymes as templates¹¹³, and in the discovery of an ancestral coagulation factor VIII protein that is now used as a therapeutic for people with hemophilia¹¹⁴. These, and many other studies^{76,111,114–118} have established this technique as an incredibly powerful tool in the protein scientist's toolkit.

Despite its utility, ASR has largely remained a technique for phylogenetics experts. In part, this is due to the complexity of the task. The individual steps of an ASR study—dataset construction, multiple sequence alignment, inference of a phylogenetic tree, and ancestor reconstruction—are usually done using separate software. This means a would-be ASR user must learn and intelligently select the most useful combination of software from a large pool.¹¹¹ The problem is made worse because some often-used software is no longer maintained: for example, PAML4 was last updated in 2007.¹¹⁹ It can also be extraordinarily difficult to organize and convert the outputs from one program into inputs for the next. At best, this is an unproductive use of time; at worst, this can lead to information loss or even errors in the final reconstructed sequences.

Here we introduce topiary, an ASR software pipeline that addresses these problems. Our first goal was to simplify and streamline the tasks necessary for an ASR study, simplifying and

codifying existing best-practice ASR into one convenient package. We hope achieving this goal will make ASR accessible to non-experts. We further hope this will improve reconstruction quality generally by removing monotony and manual file manipulations that can lead to mistakes. Our second goal is to promote and enable high-quality reconstructions. To do so, we built our pipeline around modern software tools and incorporated important-but-sometimes-difficult steps directly into the pipeline: validation of protein identity by reciprocal BLAST, gene-species tree reconciliation, and explicit ancestral character reconstruction of gaps.

There are two design features that set *topiary* apart from many other methods. The first is the use of spreadsheets rather than arcane text formats for inputs and to store the sequence database/alignment through all steps. This makes it much simpler to prepare inputs and track changes over the course of the pipeline. The second design feature is that *topiary* is species-aware through all steps. From the first step onward, it uses the Open Tree of Life synthetic species tree to inform every choice¹²⁰: how to focus initial BLAST queries, how to lower sequence redundancy while preserving taxonomic diversity, and how to construct the best possible evolutionary tree consistent with both the protein and organismal evolutionary signals. This integration greatly simplifies the user experience and ultimately yields rooted, well-resolved phylogenetic trees for ancestral reconstruction.

We have broken our description of the software package into four sections. In the first section, we go through the process of ASR in general, describing the state-of-the-art for such a calculation. Our goal is to familiarize non-specialist readers with the workflow so they can understand what *topiary* does (and why), as well as interpret the output from a *topiary* calculation. In the second section, we describe the specific pipeline and design decisions within the *topiary* package. This section focuses on the automated, software-driven steps in the pipeline.

In the third section, we briefly describe the protocol for running topiary in practical terms for the user, working through an example calculation. Finally, in the fourth section, we describe the work done to validate the pipeline.

OVERVIEW OF ANCESTRAL SEQUENCE RECONSTRUCTION

Define the problem

The most important task in an ASR study is to define the problem. *What ancestors do you want to reconstruct? What feature(s) of those proteins will you measure?* For an evolutionary biochemist or protein engineer, ASR studies often involve tracing the evolution of functions observed in modern proteins. **Figure 2.1** shows this schematically for a hypothetical protein family. Paralog A has some activity (denoted with a star); paralog B does not. (As a reminder, paralogs are homologs that arose by gene duplication; orthologs are homologs that arose by speciation.) If we are interested in the evolution of the star activity, we would likely be interested in reconstructing ancAB and ancA (arrows, Figure 1). Because all A paralogs have the activity, we predict ancA did as well. But because only A paralogs have the activity— and not paralog B or the fish proteins—we predict ancAB was not active. By reconstructing ancA and ancAB, we can isolate and study the key sequence differences between the ancestors that conferred the activity.

The first step in an ASR study is to build up a picture of the functions of modern proteins in the family through pilot studies and literature searches. Specifically, one must know: (1) The biochemical/functional features of interest and, (2) What homologs exist in what organisms. In our example, identifying ancA and ancAB as the ancestors of interest required knowing the distribution of function across modern proteins. If we knew only the function of human paralog

A, but no other proteins in the family, we would be hard-pressed to choose the appropriate scope for the ASR study. Likewise, if we knew that paralog A but not paralog B existed, we would not predict the ancAB to ancA transition. The topiary package uses a list of modern proteins covering the relevant paralogs and species as the starting point for the ASR pipeline (later: the “seed dataset”).

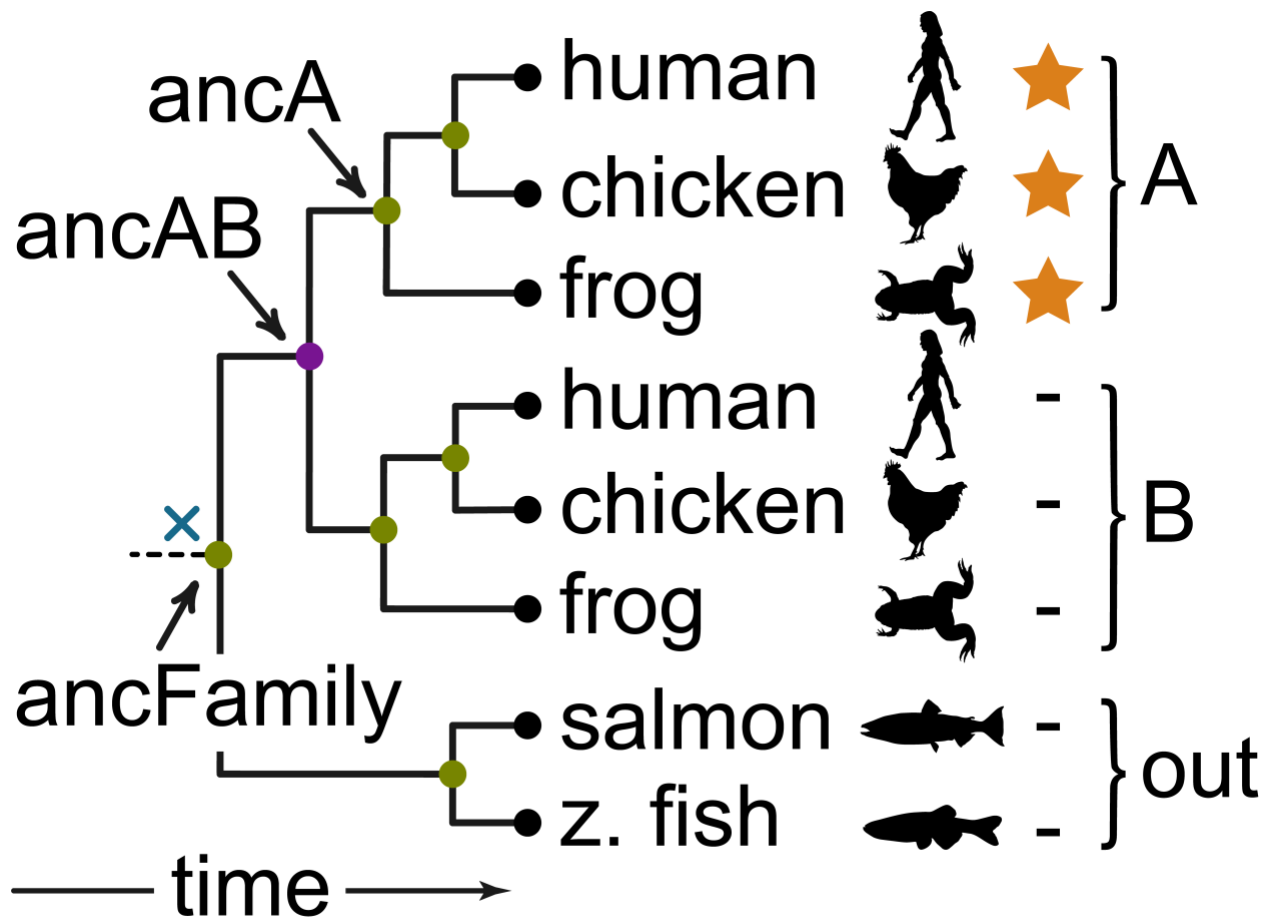


Figure 2.1. Define the ancestral reconstruction problem. The panel shows the evolutionary history of a hypothetical protein family with two paralogs, A and B. The tree is rooted: ancestors are arranged from ancient to recent, left to right. Black circles at the tips of the tree denote modern protein sequences from the indicated species. Colored internal nodes indicate gene duplications (purple) or speciations (green). An ASR study aims to reconstruct the sequences of these ancestral nodes. The node annotated with a blue “x” is not reconstructable (see text). A biological activity of interest is indicated on the tips: active (star), inactive (black dash). The simplest evolutionary scenario would have activity evolving between ancAB and ancA; these would be good candidates for reconstruction.

Note: it is important that the ancestral protein of interest cannot be the root of the phylogenetic tree. To reconstruct an ancestor, one needs input from three branches: the descendants and the previous ancestor. The ancFamily ancestor in **Figure 2.1** at the root of the tree has no sequence information from the ancestral branch (dashed line) thus we cannot reconstruct ancFamily. This contrasts with ancAB, which can be reconstructed because it forms a node at the intersection of three branches: a descendant branch leading to ancA, a descendant branch leading to ancB, and an ancestral branch leading back to the fish proteins (known as the outgroup). This sets a limit on our deepest reconstructable ancestor: our dataset must include an outgroup that diverged one node earlier than our deepest ancestor of interest.

Construct a sequence dataset

Once we have identified the ancestors we would like to reconstruct (**Figure 2.1**), we begin the steps of the ASR pipeline (**Figure 2.2**). The first step is to create a dataset of high-quality sequences spanning the relevant species and protein family members. Continuing our example, we start with a handful of sequences that cover bony vertebrates (humans through fish) and the two paralogs (A and B) (**Figure 2.2a**). We then collect as many sequences from as many species as possible, usually by BLASTing against online databases using our starting sequences as queries (**Figure 2.2b**).

Our confidence in our reconstructed ancestral sequences depends on the quality and diversity of the sequences in the alignment.¹²¹ Because of this, we perform quality control on the resulting sequence dataset. We want to avoid low-quality or partial sequences, keep only one sequence per gene per species, and maintain an even representation of proteins across species. In our example in **Figure 2.1**, the branches leading from ancAB are amniotes

(mammals/birds/reptiles), amphibians, and ray-finned fishes. To maximize reconstruction quality, we should ensure a good representation of protein sequences from these species in our dataset.

Sequence alignment

The next step is to build a multiple sequence alignment (MSA) (**Figure 2.2c**). Alignment quality is critical for a successful reconstruction study.¹²¹ This is because an MSA makes homology statements, asserting that sites within each column arose by evolutionary descent. Incorrect homology statements will lead to poor reconstructions. We use alignment software to generate an MSA, followed by more quality control. Usually, alignment quality ends up being assessed by computational tools^{122,123} and/or by manual evaluation and editing^{124,125}. Generally, we remove difficult-to-align termini, poorly aligned sequences, or whole regions of an alignment that may not be of interest for an ASR study (for example, a disordered and evolutionarily divergent linker region).

Infer a maximum likelihood gene tree

The next step is to construct a phylogenetic tree describing the evolutionary relationships between the sequences in our alignment (**Figure 2.2d**, tree on the right). Most ASR studies do this using probabilistic models of sequence evolution. These are built around substitution matrices that describe the probability of specific amino acid changes over evolutionary time. (For example, aspartic acid to glutamic acid will have a much higher probability than aspartic acid to phenylalanine.) Most models consist of parameters defined in the model as well as parameters

estimated from the input alignment. Selecting the correct substitution matrix is critical to high quality ancestral reconstruction.¹²⁶

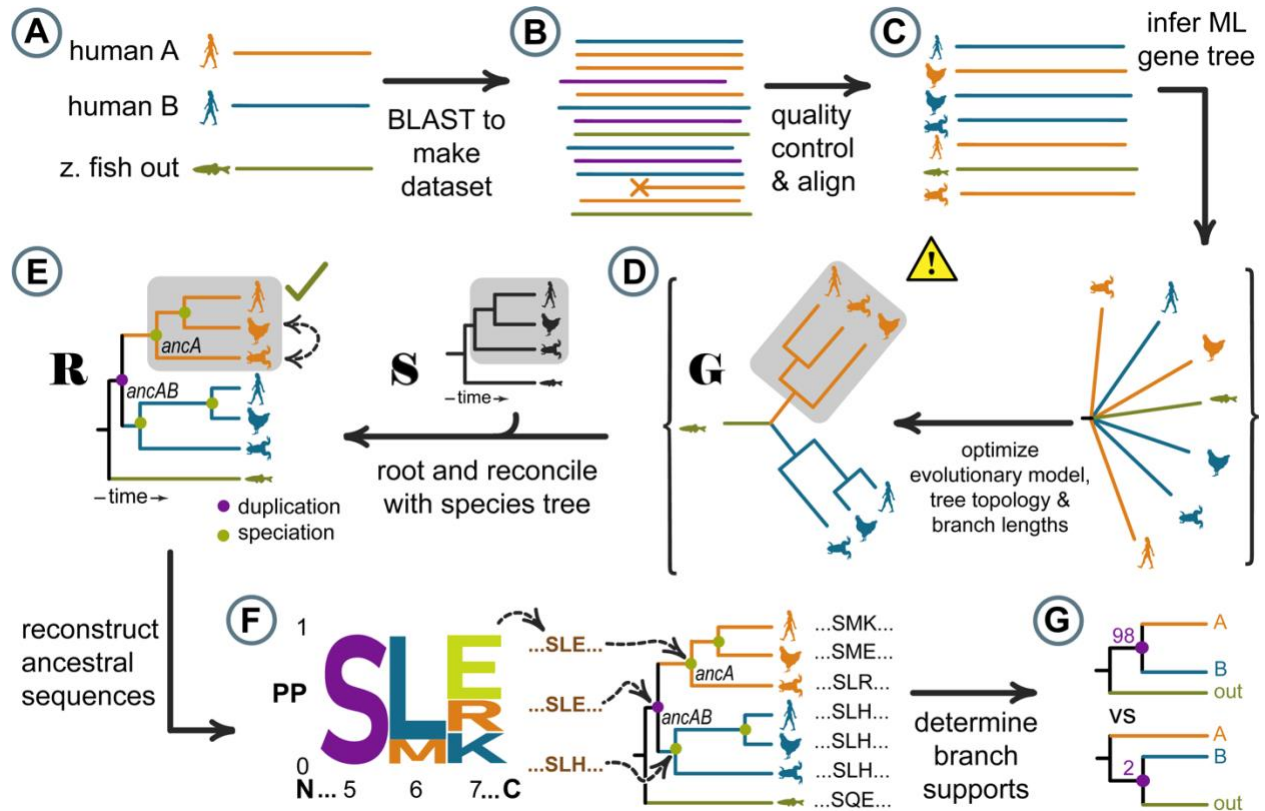


Figure 2.2. Ancestral sequence reconstruction has six main steps. (a) Start with a handful of homologous protein sequences spanning the paralogs of interest and their taxonomic distribution. Throughout the figure, color indicates the identity of the protein (orange: paralog A, blue: paralog B, and green: outgroup); the icon indicates the species (human, chicken, frog, fish). (b) Use these sequences as BLAST queries to construct an initial sequence dataset. Some returned sequences are not homologs of interest (purple); others are low quality (i.e., a partial sequence indicated by 'x'). (c) Select high quality sequences and generate a multiple sequence alignment from that dataset. (d) Infer a maximum likelihood gene tree 'G' for the protein sequences in the alignment. This infers branching relationships but does not orient the tree with respect to time. Poorly reconstructed protein relationships may exist (clade in gray box). (e) Reconcile the gene tree with the species tree 'S', yielding a reconciled gene tree 'R'. This corrects weakly supported protein relationships and roots the tree in time. (f) Reconstruct the sequences of ancestral proteins of interest using the reconciled tree. Sequences are selected by posterior probability (PP). Sequence logo depicts ancestor "ancA" with letter height proportional to amino acid PP. Position 5 is unambiguously "S"; position 6 is likely "L" but could be "M"; position 7 could be "E", "R", or "K". Examples of maximum likelihood ancestral sequences are shown in brown for the specified nodes. (g) Assess confidence in tree topology. Branch supports for two different trees indicate strong support for the top tree (98) and weak support for the bottom tree (2).

Most ASR studies use a maximum likelihood (ML) modeling framework. The goal is to find the substitution model and evolutionary tree that give the highest probability of observing the sequences in the alignment. The maximization process involves selecting a substitution matrix, tuning quantitative model features, inferring the tree topology (i.e., the pattern of branching events that gave rise to the modern sequences), and optimizing the branch lengths (how much evolutionary change occurs between each branching event). This is a complex, many-parameter, optimization problem. For more details, and discussions of alternative approaches including Bayesian methods, see^{111,118,127}.

After this step, one has an ML gene tree with a branching pattern describing the evolutionary relationships between all sequences in the alignment (**Figure 2.2d**, tree G). The inferred tree reveals which sequences group together, but not the order in which these groupings evolved. In technical terms, the tree is unrooted. This is because most probabilistic evolutionary models are time reversible; the probability of the evolutionary branching relationship is independent of where one starts the evolutionary process. In practical terms, it means we cannot determine which ancestors were the most ancient without outside information.

Reconcile the gene tree to the species tree

We now reconcile the inferred gene tree with the species tree to obtain our gene-species reconciled tree (R in **Figure 2.2e**). In this process, we identify nodes in the gene tree that correspond to speciation versus gene duplication events (green and purple nodes on tree R, respectively). Note that reconciliation is not always possible or desirable; however, we will leave that consideration until the next section.

This reconciliation process has two important outcomes. First, it roots the gene tree, allowing us to order the occurrence of ancestors in time. This is because the species tree is rooted; we know which ancestors occurred at what times based on outside information, such as the fossil record. By identifying speciation events in the gene tree, we learn the temporal order of ancestors in the gene tree.

Second, reconciliation resolves ambiguous relationships within the gene tree. This is shown in the gray boxes in **Figure 2.2d,e**. The initial gene tree placed human and frog proteins together to the exclusion of the chicken protein. This does not match known species relationships. One might explain this through a complicated set of gene duplications and losses: maybe, after an early duplication, humans and frogs independently lost one copy of the gene and chickens lost the other. A far simpler explanation is that the gene tree incorrectly placed humans and frogs together. Reconciliation software uses a variety of strategies to determine whether to add evolutionary events or rearrange the tree topology.¹²⁸

For an ASR study, the key takeaway is that gene-species tree reconciliation yields a rooted gene tree that incorporates additional species-level information. This leads to higher quality reconstructed ancestral sequences and allows us to order those ancestors in time.¹²⁹

Reconciliation: The special case of microbial genes

Although reconciliation should, in principle, yield a more accurate picture of the evolutionary history of a protein, in practice, reconciliation is not always possible. Problems are particularly likely for microbial genes. This is because we have relatively low confidence in the microbial species tree. (Indeed, some question the existence of a single microbial species tree, or even the concept of a microbial species¹³⁰) As a result, ASR studies of microbial proteins have

generally relied on unreconciled gene trees.¹³¹ Reflecting this reality, topiary does not reconcile the gene and species trees for datasets consisting of purely microbial genes. Instead, topiary roots the resulting tree using the midpoint approximation method.¹³² Because reconciliation is not performed, topiary does not label nodes with evolutionary events such as duplication or speciation. For the rest of this walk through, we will describe the approach assuming reconciliation is performed as this is a more complex version of the pipeline than the simplified microbial workflow.

Reconstruct ancestors

We can now reconstruct ancestral sequences (**Figure 2.2f**). We traverse the reconciled tree and estimate the sequences of every ancestor.¹³³ For each ancestor, we consider sites individually. We calculate the likelihood of all 20 amino acids at that site given the ML parameters of the probabilistic model and the amino acids observed at that position in the alignment. From these, we determine the posterior probability (PP) for each amino acid. This is the likelihood of a given amino acid relative to the likelihoods of all amino acids. (In mathematical terms, $PP_i = L_i / \text{sum}(L_{aa})$, where L_i is the likelihood of amino acid i and $\text{sum}(L_{aa})$ is the sum of the likelihoods of all amino acids.)

We use these posterior probabilities to construct ML ancestors. For each site, we select the amino acid with the highest posterior probability. For example, at ancA site 5 in **Figure 2.2e**, we select “S” because it has a PP close to 1.0. This is an unambiguous reconstruction. Not all sites are this clear cut. At site 6, two amino acids are possible; we select the amino acid with the higher probability of the two (“L” over “M”). At site 7, there are multiple possibilities; however, we would still select the amino acid with the highest PP. For ancA, the sequence that maximizes

the posterior probability at these positions is “SLE”. (Note: gaps are usually treated separately and reconstructed using maximum parsimony; see Section 3 for details).

Evaluate results

Before synthesizing and characterizing ancestral proteins, we evaluate their quality. We look at two metrics. The first is the average posterior probability for the ML amino acid at all positions in the ancestor. A well reconstructed ancestor would have an average PP of 1.0, meaning the model has high confidence in the sequence at all sites. At the other extreme, a completely ambiguous ancestor would have an average PP of 1/20 (0.05), meaning each site could have any one of the amino acids. Generally, ancestors in published studies have $PP > 0.85$.

To assess the effect of phylogenetic uncertainty on inferences about the functions of ancestors, we synthesize two versions of every ancestor. The first is the ML ancestor, as described above. The second is the so-called altAll ancestor.¹³⁴ For the altAll ancestor, we replace all ambiguous ML amino acids with the next most-probable amino acid. If an ancestor has 10 ambiguous sites, the ML and altAll would differ at all 10 of these sites. By functionally characterizing both the ML and altAll versions of an ancestors, we can determine which features are robust to uncertainty in the reconstruction.^{76,115,135–138}

The second quality metric is the branch support for a given ancestral node. Posterior probabilities measure our confidence in the ancestral sequence given a particular phylogenetic tree, but they do not measure our confidence in the tree itself. (Put another way, we have the sequence of an ancestral node, but how confident are we that the node existed?) Branch supports measure this confidence. We discuss how these are estimated in Section 3; for now, we focus on interpretation.

A branch support measures our confidence that a given group of sequences cluster together, typically on a 0–100 scale. **Figure 2.2g** shows branch supports for two possible arrangements of the tree: placing paralog A with B (orange with blue) or paralog B with the fish outgroup (blue with green). In this example we have high support (98/100) for placing paralogs A and B together, with contrasting low support for separating them (2/100). For an ASR study, we need to have high confidence that an ancestral node existed (typically branch support >85) prior to characterizing the ancestral protein.

THE TOPIARY PIPELINE

The steps above are relatively complex, involving multiple different software packages for dataset construction, sequence quality control, alignment, model selection, gene tree inference, gene-species tree reconciliation, and ancestral reconstruction. Further, there are places where expert phylogenetic knowledge might be required. How does one obtain a species tree? How does one select which species to include when trying to reconstruct a specific ancestor? How does one evaluate whether a given ancestor is well reconstructed? The topiary package aims to streamline this process, simplifying the workflow and helping non-experts make evolutionarily informed decisions.

Only a few steps in ASR require human input: defining the problem, checking the alignment, and characterizing the resulting ancestors. The rest of the steps are computational, with different software packages typically chained together via user manipulation. Given this process, we set out to build software that facilitates the few human-centric steps and then automates the rest of the pipeline (**Figure 2.3**). In this section we walk through the topiary pipeline, describing the design decisions and software used throughout. Here we emphasize the

automated steps; the following Section 4 focuses on the human steps. Both will closely parallel the steps described in general terms in **Figure 2.2**.

Software design

One of our design goals was to use software that is state-of-the-art, up-to-date, and currently maintained. Topiary uses Muscle5 for alignment¹³⁹; RAxML-NG for maximum likelihood gene tree and ancestral sequence inference¹⁴⁰; GeneRax for gene-species tree reconciliation¹²⁸; and PastML for gap reconstruction¹⁴¹. Under the hood, it uses the ETE 3 library for tree manipulations¹⁴²; Biopython to access NCBI BLAST and the NCBI database^{143,144}; python-opentree to interact with the Open Tree of Life taxonomic database^{120,145}; and toytrees for drawing trees¹⁴⁶. This is implemented within a standard Python 3 scientific computing environment built around numpy and pandas.

The pipeline (**Figure 2.3**) is broken into two stages: (1) Construct an MSA from the seed sequences and (2) Construct phylogenetic tree ancestors given the MSA. The first computational stage of the pipeline can be run on a user's personal computer (Linux, macOS, Windows); the second stage is best run using a high-performance computing environment and requires Linux or macOS. Users can run the pipeline via a few command-line programs, or work through each step individually and interactively in a Jupyter notebook. For ease of installation, the software and all dependencies are readily installed using the “conda” software environment. The software is also available for direct download at <https://github.com/harmlab/topiary>. A collection of example datasets and Jupyter notebooks are available at <https://github.com/harmlab/topiary-examples>.

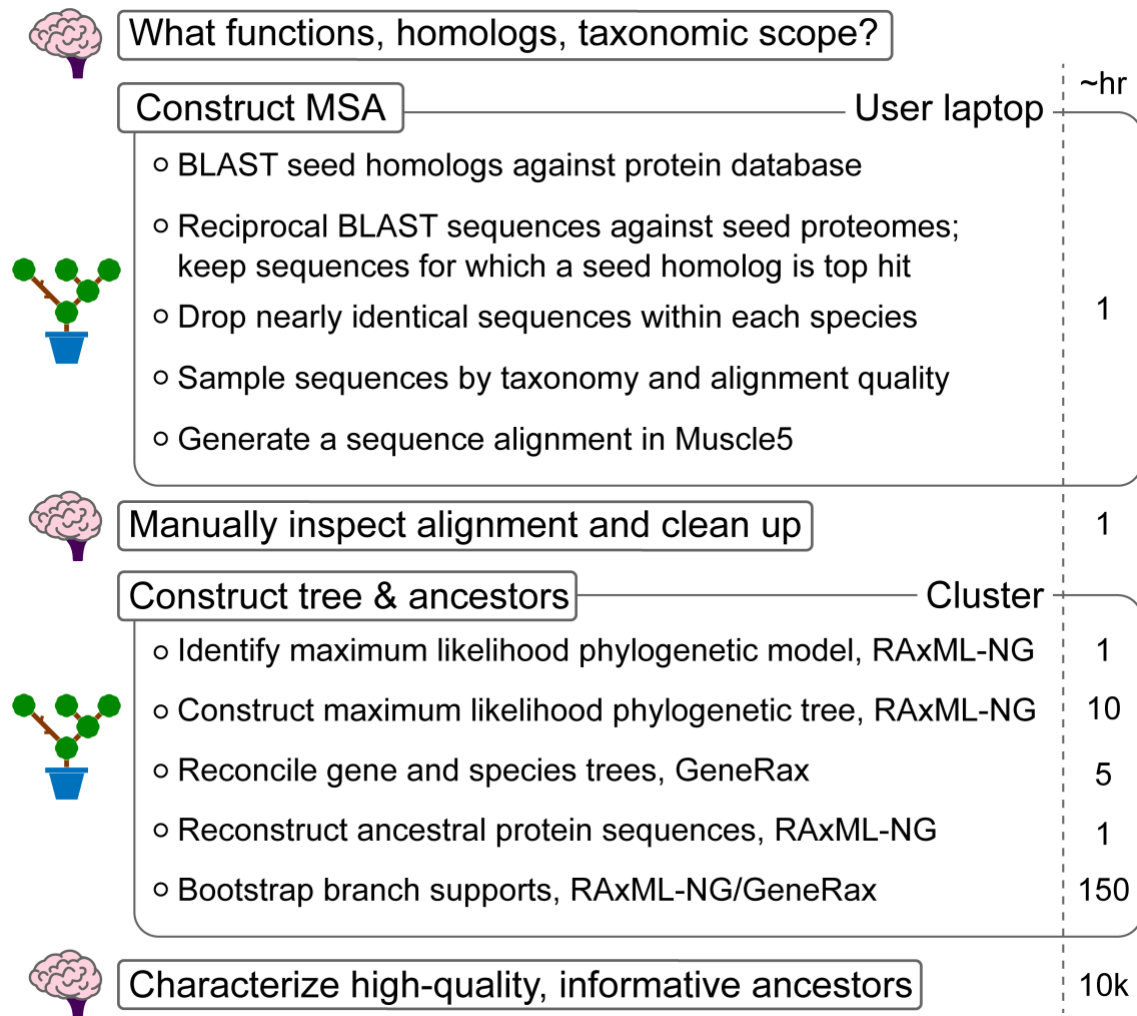


Figure 2.3. Summarized topiary ASR pipeline. The pipeline is a series of human and automatic steps (indicated on the left with brain and topiary icons, respectively). The approximate time, in hours, required for each step is indicated on the right.

Our focus will be on topiary's algorithms and software settings; however, in passing, we want to note several aspects of the software. We refer users to the online documentation (<https://topiary-asr.readthedocs.io/>) for more details.

1. Topiary has a fully documented Application Programming Interface (API), allowing users to run interactive analyses in a Jupyter notebook or write their own python scripts.

2. Topiary is multithreaded, improving the speed of local BLAST queries, redundancy reduction, and NCBI downloads. It also takes full advantage of the parallelization support implemented in Muscle5, RAxML-NG, and GeneRax.

3. Topiary allows users to restart interrupted pipelines without having to start over. This is particularly useful for the second stage, which can take a fair amount of time to run on a computing cluster.

Stage 1: Seed to alignment

As described in the Overview, the starting point for an ASR calculation is defining the problem. Topiary does this in a straightforward way: the user constructs a seed dataset that defines the paralogs of interest and the desired taxonomic distribution for the ASR study. For the example worked through in **Figures 2.1 and 2.2**, the seed might include three sequences: paralogs A and B from humans and a single protein from zebrafish (**Figure 2.2a**). The user prepares the seed dataset as a spreadsheet with four columns: sequence, species, name (e.g., the paralog identity), and aliases (what names this protein has across the various online databases). The species in the seed are used as “key species” in all downstream analyses. We go into further details on how to construct this seed dataset in Section 4 . From this starting point, topiary downloads high-quality homologous protein sequences from public databases and then generates a draft multiple sequence alignment.

Initial dataset construction

Topiary uses the seed sequences to BLAST against the NCBI non-redundant protein sequence database. To maximize the number of productive results, topiary automatically sets the

taxonomic scope of the BLAST search. For non-microbial proteins, the scope is given by the taxonomic rank that encompasses the key species from the seed dataset, plus a user-defined expansion. For the example above—which included humans and zebrafish—the taxonomic rank is Vertebrata. With an expansion of one, the scope would be Craniata; with an expansion of two, the scope would be Chordata (Vertebrata → Craniata → Chordata). Using the default expansion of two, topiary would BLAST each of the seed sequences against the NCBI non-redundant protein database, limiting its results to Chordata. By default, topiary pulls down up to 5000 hits per seed with an intentionally generous e-value cutoff of 0.001. (Users have full control over the BLAST search parameters.) Note that a seed dataset containing only bacterial or archaeal sequences would be assigned a taxonomic scope of “All Bacteria” or “All Archaea.”

In addition to this default method for building a sequence dataset, users can specify other sources of sequences including other NCBI BLAST databases, local BLAST databases, or previously saved BLAST XML files. Users can also manually add sequences by appending them to the initial spreadsheet.

Once the initial dataset is constructed, topiary identifies each hit by reciprocal BLAST. It downloads proteomes for the key species in the seed dataset and constructs a combined local BLAST database. It then uses the hits above as queries against the key species BLAST database, searching the resulting reciprocal hits for text descriptions that match the aliases specified in the seed dataset. (See Section 4 for details about defining aliases.) It weights each hit by $2s/t$ where s is the BLAST bit score, and t is a user-defined parameter (default = 1). Finally, topiary calculates the posterior probability that the sequence is a given paralog by calculating the sum of the weights for all reciprocal hits that match a paralog alias and then dividing by the sum of the weights from all reciprocal hits (Frith, 2019).¹⁴⁷ A sequence is assigned a paralog identity based

on a user-defined stringency cutoff (default = 0.95). Multiple paralogs may be assigned if the sum of their posterior probabilities is above the cutoff.

Redundancy reduction, quality control, and alignment

This BLAST approach typically finds many more sequences than are necessary or practical for a standard phylogenetic analysis. We must therefore select sequences that sample the diversity in the dataset without compromising our ability to infer ancestors (step from **Figure 2.2b,c**). Topiary selects a subset of sequences using a combination of taxonomy, sequence identity, and sequence quality. By default, topiary aims to build an alignment with approximately one sequence per site in the average length of seed sequences. If our seed sequences were 100 amino acids long, topiary would try to build an alignment with 100 sequences. This prevents over-fitting and makes later computational steps faster. (Users can change the target alignment size if desired.)

Topiary uses four strategies to decrease the size of the dataset while maintaining dataset quality. First, sequences defined in the initial seed dataset (**Figure 2.2a**) are kept, regardless of their quality score or redundancy. This means users can pre-specify sequences they need in their final alignment.

Second, for datasets containing non-microbial genes, topiary selects sequences based on their placement on the species tree rather than solely based on their identity. (For microbial datasets, topiary lowers redundancy based on sequence identity alone because microbial species trees are poorly resolved.) When lowering redundancy in a species-aware fashion, topiary takes the desired alignment size and then divides this “budget” across the species seen in the dataset. The algorithm is shown in **Figure 2.4** for a hypothetical dataset with seven orthologous proteins

and a target alignment size of five. Topiary starts by downloading the species tree from the Open Tree of Life for all represented species. It then assigns the deepest ancestral node on the tree a budget of five sequences. Topiary traverses the tree, from ancestor to tips, splitting the sequence budget as evenly as possible among descendant lineages at each step. In the example, it assigns two sequences to the ancestor of bony fishes and three sequences to the ancestor of tetrapods. On the bony fish lineage, it assigns one sequence each to the zebrafish and salmon, meaning these sequences will be kept in the final dataset. On the tetrapod branch, the algorithm continues, assigning one sequence to the frog and two sequences to the ancestor of amniotes. It then gives one sequence to the bird/reptile ancestor (dark green clade) and the other sequence to the mammal ancestor (light green clade).

Because of this explicitly taxonomic strategy, sequences that are taxonomically important are not removed from the dataset, even if their quality is lower than other, taxonomically redundant, sequences. The frog sequence in **Figure 2.4**, for example, has a long lineage-specific insertion. But because it is the only amphibian representative in this (toy) alignment, it is preserved. We leave the decision of whether or not to keep this sequence up to the user when they review the alignment. We also note that, in practice, there is enough sequence and taxonomic diversity in current databases that we rarely need to trade alignment quality for taxonomic diversity.

Third, lowering sequence redundancy, topiary preferentially keeps sequences that align well to the seed sequences. We take this alignment-focused approach because ASR can only reconstruct ancestral states for columns seen in many modern proteins. Lineage-specific insertions and deletions do not contribute to the ancestral inference and, further, may interfere with MSA construction. To calculate alignment quality, topiary aligns clusters of sequences from

closely related organisms to the whole seed sequence dataset using Muscle5. It identifies “dense” columns in which most sequences have non-gap characters (the gray shaded boxes in **Figure 2.4**). It then calculates two quality scores for each sequence. First, it calculates the proportion of dense columns with non-gap characters in the sequence. Lower proportions indicate truncated sequences. Second, it looks for long stretches of non-gap characters that are not in “dense” columns, indicating a lineage-specific insertion. In our example dataset, topiary would select the human and chicken sequences over mouse and lizard, as these have the best alignment scores (**Figure 2.4**).



Figure 2.4: Topiary redundancy reduction and quality control. This analysis starts with seven sequences (taken from seven organisms) with the goal of retaining five for the downstream analysis. The numbers next to the ancestral nodes on the tree are the budget allocated for all descendants: five for all organisms, two for the fishes, three for tetrapods, etc. The “keep” column indicates which sequences are kept for further analysis after the redundancy reduction step. A schematic alignment is shown on the right, with poorly aligned and missing regions labeled. The alignment quality is used to select which sequences to keep within taxonomic blocks (human/mouse and lizard/chicken, in this example).

Fourth and finally, there are a few steps where topiary lowers redundancy based on shared sequence identity. Whenever this is done, topiary chooses the sequence to keep based on its relative quality. It calculates an identity score by performing a pairwise alignment with the

Biopython pairwise2.align.localxx function and dividing the score by the length of the shorter sequence. If this number is above a specified identity cutoff, topiary selects which of the two sequences to discard based on a rank ordered vector of sequence features. These features are: “Sequence length deviates from median sequence length by more than 25%” > “Low quality” > “Partial” > “Predicted” > “Precursor” > “Hypothetical” > “Isoform” > “Structure” > “shorter sequence” > “random choice”. Some of these features are calculated by topiary (i.e., sequence length), others are extracted from NCBI sequence descriptions (i.e., Partial, Hypothetical). This process enriches the final dataset for higher-quality protein sequences.

This protocol yields a relatively clean dataset with 5% more sequences than our target alignment number. We leave these extra sequences in place so we can manually delete the worst aligners upon visual inspection and still have our approximate target number of sequences.

Alignment

Topiary uses Muscle5 with its default parameters to generate the MSA (**Figure 2.2c**).¹³⁹ We selected this algorithm due to its demonstrated high performance, as well as the extremely fast “super5” algorithm that is useful for generating draft alignments for large datasets. Advanced users can set all Muscle5 options via the API.

There are differing views about whether to manually edit alignments or not.^{124,125} The topiary pipeline leaves this decision in the hands of the user. The goal for topiary is to make the task of finalizing an alignment relatively painless by carefully filtering for well-aligned sequences and by using state-of-the-art alignment software: most of the sequences should already be well aligned. Over the years, we have settled on a 5% approach: automate up to the point where the alignment is 95% done, and then finalize the alignment with a human brain. This has

proven much more practical than designing a complicated (and thus fragile and unpredictable) heuristic to completely automate alignment construction.¹²² In practice, most of our manual work consists of deleting a handful of problematic sequences, followed by global realignment in Muscle5. (See Section 4 for details.)

Stage 2: Alignment to ancestors

In stage 2, we go from our alignment to ancestral sequences (**Figures 2.2c–g and 2.3**). We selected RAxML-NG¹⁴⁰ as our primary phylogenetic package. One key reason for this choice was that RAxML-NG integrates well with GeneRax, a clear choice for reconciling gene and species trees. Both GeneRax and RAxMLNG use the same underlying computational phylogenetics library—libpll¹⁴⁸—thus ensuring internally consistent implementations of evolutionary models. Further, GeneRax was explicitly tested with RAxML-NG, making this the most conservative choice of software combinations. Finally, we wanted to calculate branch supports for our species-reconciled gene tree (**Figure 2.2g**). Because GeneRax does not implement any fast-branch support methods, we estimate branch support by non-parametric bootstrap.¹⁴⁹ RAxML-NG can return pseudoreplicate alignments matched to pseudoreplicate trees. This allows us to feed bootstrap pseudoreplicates into GeneRax as separate, parallel calculations and thus conveniently determine branch supports on our species-reconciled gene trees.

Infer the evolutionary model

The first step in a maximum likelihood phylogenetic analysis is determining the maximum likelihood model of sequence evolution. This includes the matrix for amino acid

substitution (i.e., LG, JTT, WAG, etc.), the stationary frequencies for that model, rate variation parameters (Γ distribution, rate categories, etc.), and the proportion of invariant sites. Topiary uses a conventional method to find the best model.¹⁵⁰ It uses RAxMLNG to generate a maximum parsimony tree from the alignment. It then optimizes branch lengths and other parameters using all 360 combinations of these model parameters implemented in the computational library that underlies RAxML-NG and GeneRax. Finally, it ranks these models based on a corrected Akaike Information Criterion, which penalizes models with excess parameters to prevent overfitting.

Although this protocol is done automatically, topiary returns a variety of statistics including AIC (Akaike Information Criterion), AICc (Corrected Akaike Information Criterion), and BIC (Bayesian Information Criterion) to help users who want more control over model selection. Via the API, users can also specify a custom input tree or a subset of the models to test. (Note: as of the current version, topiary excludes the LG4M and LG4X models, as these cause GeneRax to crash during gene-species tree reconciliation.)

Build a maximum likelihood gene tree

Topiary next infers an ML gene tree using the inferred phylogenetic model with the default RAxML-NG settings for the “—search” protocol. This starts the inference from 10 random trees and 10 different parsimony trees. It then optimizes the tree topology using a subtree pruning and regrafting (SPR) subtree cutoff of 1, with an automatically selected fast versus slow SPR radius. Branch lengths are optimized using the NR-FAST algorithm. The tree with the highest likelihood is selected and used for downstream analyses (**Figure 2.2d**, tree G). Advanced users have full access to all RAxML-NG options via the topiary API.

Reconcile gene and species tree

The next step in the pipeline is to reconcile the gene tree with the species tree (**Figure 2.2e**). (Note, this reconciliation step is skipped for datasets containing only microbial genes.) Reconciliation automatically roots the tree and has been shown to improve the quality of reconstructed sequences.¹²⁹ For this purpose, we use GeneRax, a new high-performance program for reconciling gene and species trees. Unlike other, heuristic, methods, GeneRax explicitly models evolutionary events (speciation, duplication, loss, and lateral gene transfer) as well as sequence evolution (e.g., the LG model).¹²⁸ If the gene and species trees are discordant, GeneRax can either rearrange the gene tree to follow the species tree or incorporate an evolutionary event (such as duplication) to account for the discordance. GeneRax finds the maximum likelihood reconciled tree that balances the signal from the aligned sequences against the plausibility of the evolutionary events required to generate that signal.

Topiary uses the ML evolutionary model and ML gene tree inferred previously as inputs to GeneRax. For the rooted species tree, topiary automatically downloads the most recent synthetic tree from the Open Tree of Life (OTL) database.^{120,145} (Previous steps in the pipeline ensure that all sequences that have made it to this step come from species that are present in the OTL database.) Any polytomies in this tree are resolved arbitrarily prior to the reconciliation inference. Topiary runs GeneRax with the default parameters¹²⁸: topology optimization using rounds of SPR with increasing radius (from 1 to 5) using the UndatedDL reconciliation model. The UndatedDL model accounts for duplication and loss events. Topiary users can select the UndatedDTL model, which allows lateral transfer, if they expect lateral gene transfer for their genes of interest.

The resulting tree is a maximum likelihood species-reconciled gene tree with optimized branch lengths and nodes labeled with inferred evolutionary events (speciation, duplication, or transfer). GeneRax returns a variety of other outputs that are made accessible to topiary users, but only the reconciled tree is used further in the pipeline.

Reconstruct ancestors

The next step is to infer sequences of ancestral nodes on the species-reconciled gene tree (**Figure 2.2f**). For this, we use RAxML-NG, which implements a standard marginal ancestral reconstruction method.¹³³ (This differs from previous versions of RAxML, which used a non-standard reconstruction method that was not comparable to other approaches.) RAxML-NG finds the amino acid at each site in each ancestor that maximizes the likelihood of observing the sequence alignment given the tree, branch lengths, and phylogenetic model. This returns a matrix of posterior probabilities for each amino acid at each site in the alignment for each ancestral node. Topiary extracts the sequence of the maximum likelihood ancestor, as well as the so-called altAll version of the ancestor that incorporates alternate reconstructed amino acids at ambiguous positions. It uses a default cutoff of 0.25 to identify ambiguous sites¹³⁴; this can be set by the user.

The evolutionary models used by RAxML-NG do not explicitly treat gaps; therefore, the first draft of the reconstructed ancestor will be ungapped. Topiary assigns gaps by treating them as characters during ancestral character reconstruction. For this purpose, topiary uses the DOWNPASS¹⁵¹ algorithm as implemented by the PastML package¹⁴¹. The final output for this step consists of the gapped sequences of both maximum likelihood and altAll ancestors for each node. These have associated statistical supports: posterior probabilities for each reconstructed

amino acid and support for gaps. Topiary also puts out a variety of summary graphs to help select high quality sequences (see Section 4).

Branch supports

To determine branch supports (**Figure 2.2g**), topiary uses non-parametric bootstrapping.¹⁴⁹ Briefly, RAxML-NG generates pseudoreplicate alignments by sampling columns, with replacement, from the input alignment. RAxML-NG then infers an evolutionary tree for each of these alignments. Topiary generates up to 1000 bootstrap pseudoreplicates, using RAxML-NG's automatic Extended Majority Rules (autoMRE) method with a cutoff of 0.03 to determine the exact number. The output from RAxML-NG is a collection of pseudoreplicate alignments and pseudoreplicate gene trees. Because we are reconstructing ancestors on the reconciled tree, we pass each pseudoreplicate alignment and gene tree into GeneRax for gene-species tree reconciliation, yielding a final collection of pseudoreplicate reconciled trees. Topiary then uses RAxML-NG to map these pseudoreplicate reconciled trees onto the ML reconciled tree as branch supports. Topiary also assesses convergence for the branch support estimate using the “—bsconverge” option.

Output

Topiary generates a single directory containing all ancestors, all trees, and an html file that allows users to browse their results. This directory can be shared with others without requiring the recipient to have installed topiary. The html file can be opened in any web browser and includes information to help users assess the quality of each reconstructed ancestor. In addition to this html output, topiary also writes the output for each step into individual

directories, allowing users to access the intermediate steps and log files from each software package employed in the pipeline.

PROTOCOL

This section complements the previous section, which focused mostly on the computational steps in the pipeline (**Figure 2.3**). We will expand on the steps that require human intervention using the LY86/LY96 protein family to help demonstrate specific considerations and features. More detailed instructions are available in the topiary online documentation (<https://topiary-asr.readthedocs.io>).

Construct a seed dataset

The first step in a topiary ASR calculation is constructing a seed dataset (Figure 2a). This dataset defines protein family members of interest and the distribution of these proteins across species. Topiary uses this seed dataset to automatically find and download sequences to put into the alignment and, ultimately, evolutionary tree. As discussed in the previous sections as well as the documentation, thoughtful consideration goes into selecting proteins of interest for an ASR study and determining the taxonomic distribution of this protein family before key species are chosen for the seed dataset. An example for the LY86/LY96 protein family, a pair of closely related innate immune proteins, is shown in **Table 2.1**.

Run the seed-to-alignment pipeline

At this point the seed dataset is ready to be passed to the topiary-seed-to-alignment script. This script uses BLAST to build a dataset of thousands of protein sequences (**Figure 2.2b**), does

quality control, lowers redundancy, and then generates an alignment of sequences (**Figure 2.2c**). This generally takes less than an hour on a modern laptop. The final output consists of a single spreadsheet and a single FASTA file holding the alignment.

Table 2.1: Example seed dataset.

name	species	sequence	aliases
LY96	Homo sapiens	MLPFLFF...	ESOP1;Myeloid Differentiation Protein-2;MD-2;lymphocyte antigen 96;LY-96
LY96	Danio rerio	MALWCPS.. .	ESOP1;Myeloid Differentiation Protein-2;MD-2;lymphocyte antigen 96;LY-96
LY86	Homo sapiens	MKGFTAT...	Lymphocyte Antigen 86;LY86;Myeloid Differentiation Protein-1;MD-1;RP105-associated 3;MMD-1
LY86	Danio rerio	MKTYFNM.. ..	Lymphocyte Antigen 86;LY86;Myeloid Differentiation Protein-1;MD-1;RP105-associated 3;MMD-1

Inspect and edit alignment

Before reconstructing a phylogenetic tree and ancestors, we strongly recommend inspecting and possibly editing the alignment (**Figure 2.2c**). There are a variety of pieces of software for visualizing alignments, including AliView¹⁵², JALView¹⁵³, and MEGA¹⁵⁴. We generally use AliView because of its balance of utility and simplicity.

There are differing views on whether to manually edit an alignment^{124,125}; the *topiary* package allows a user to manually edit their alignment but does not require it. We generally recommend making a few adjustments to alignments. We describe our approach to editing alignments in detail in the *topiary* documentation (<https://topiary-asr.readthedocs.io/en/latest/protocol.html>). Importantly, if we edit an alignment, we publish the

alignment as supplemental material in the resulting manuscript so others can reproduce our work. Once the alignment is finalized, it can be read back into the topiary spreadsheet with the command line script *topiary-fasta-into-dataframe*.

Perform the ancestral inference

We recommend performing the ancestral inference in a high-performance computing environment. Because of different parallelization requirements, the ancestral inference step uses two scripts run in sequence (*alignment-to-ancestors* and *bootstrap-reconcile*). The first script infers the evolutionary model, builds the ML gene tree, reconciles the gene and species trees, reconstructs ancestors, and generates bootstrap pseudoreplicate gene trees (**Figure 2.2d–g**). It writes out a summary tree at each step (**Figure 2.5a–d**). *Alignment-to-ancestors* should take about a day for a reasonable alignment (~1000 columns, ~500 sequences) running on a reasonable compute node (~30 cores). The second script reconciles each pseudoreplicate gene tree to the species tree and constructs the final branch supports (**Figure 2.5e**). Bootstrap sampling the gene-species reconciliation is computationally intensive but can be readily parallelized. It will likely take approximately a week spread across several cores. As discussed in the next section, if one is using a reconciled gene/species tree it is important to check the validity of the reconciliation before moving onto the *bootstrap-reconcile* step. If the analysis is being done without gene/species tree reconciliation—that is, for microbial genes—only the steps shown in **Figure 2.5a,d** are performed.

Checking gene/species-tree reconciliation

Before selecting ancestors to characterize, it is important to make sure the phylogenetic tree is reasonable. The probabilistic models used in ASR are powerful, but do not capture all possible evolutionary events. One common problem is incomplete lineage sorting (ILS), where a gene duplicates but exists as several variants in a population when speciation occurs.¹⁵⁵ Different duplicates are preserved along the descendant lineages, meaning this cannot be classified as a simple duplication or speciation event. ILS is a general problem with all ASR methods and is specifically noted as being outside the scope of GeneRax.¹²⁸ Another problem is gene fusion, where different parts of a single gene have different evolutionary histories. The methods used by topiary all assume a single genetic history for each protein sequence. If we force such a model to fit a fused alignment, we will likely end up with a nonsensical evolutionary tree and meaningless ancestral sequences.

In the worst case, ILS and gene fusion can lead to nonsensical ancestors that still have high branch supports and high posterior probabilities. Looking at the reconciled tree (**Figure 2.5b**) can help you decide if this might apply to your family. A standard signal for both ILS and gene fusion is high discordance between the inferred gene and species trees. This will manifest as an unexpectedly high number of duplication and/or transfer events in the reconciled tree. If, for example, you are studying a protein family where you expect two paralogs, but you observe 20 duplication events scattered throughout the tree, there is a good chance that the evolutionary models used for ASR are not appropriate for your protein family. Topiary warns users in its summary output if there are an anomalous number of duplication events, suggesting model-violation.

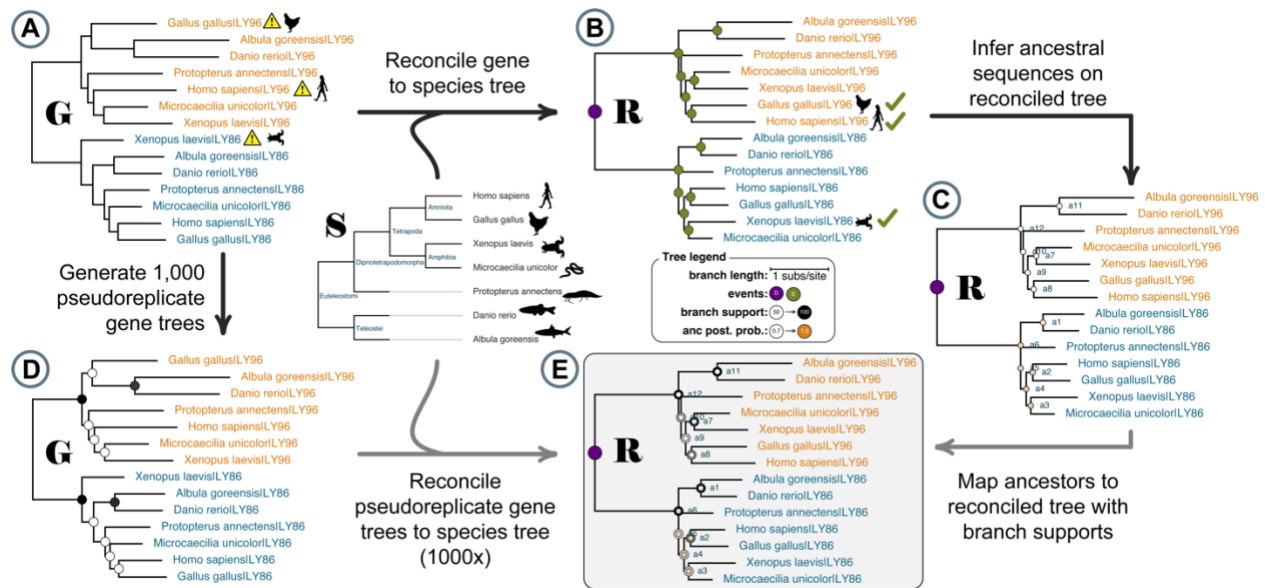


Figure 2.5. Example trees at each step in the ASR calculation. Summary trees from an ASR inference using a toy alignment with seven LY96 sequences (orange) and seven LY86 sequences (blue). Black arrows indicate steps done by the first script (alignment-to-ancestors); gray arrows indicate steps done by the second script (bootstrap-reconcile). G, S, and R indicate gene, species, and species-reconciled gene trees throughout the pipeline, respectively. (a) The ML gene tree inferred by RAxML-NG. Branch lengths are proportional to substitutions/site. This tree has several inferred relationships that are discordant with the species tree (yellow exclamation points). (b) Topiary uses the gene tree from panel A and the Open Tree of Life species tree (S) as inputs to GeneRax, constructing the reconciled tree (R). The discordant species relationships are resolved (green check marks) and each node is now labeled as either a duplication or speciation event (purple and green, respectively). (c) Tree with posterior probabilities for ML ancestors mapped onto nodes as an orange color gradient. (d) Topiary generates 1000 pseudoreplicate gene trees and maps the resulting branch supports onto nodes as a black color gradient. (e) The final output of topiary is the reconciled tree with evolutionary events, ancestor posterior probabilities, and branch supports mapped onto all ancestral nodes. In this figure, the labeled speciation events have been dropped for clarity.

If your protein has more than one domain, one option would be to try to reconstruct each domain independently. If the discordance disappears, it is good evidence for a gene fusion event. If the discordance remains, proceed with extreme caution.

One way forward in the face of discordance is to compare the sequences—and functional characteristics—for any ancestors of interest reconstructed using either the gene tree alone or the reconciled gene tree. (Topiary returns ancestors inferred on both trees.) If the results for

ancestors reconstructed on the two trees differ dramatically, one cannot infer the ancestral sequence with confidence given standard ASR methods. ILS and gene fusion are longstanding problems in phylogenetics; treating them requires expert input.

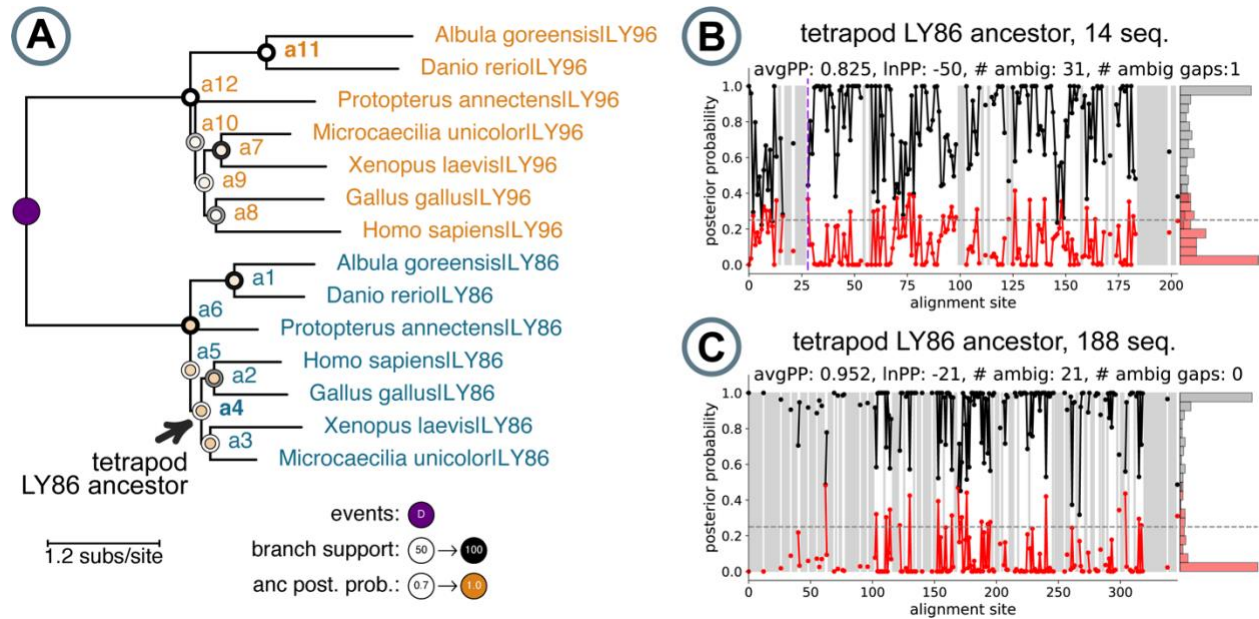


Figure 2.6. Graphs for evaluating ancestor quality. (a) The final bootstrap supported gene-species reconciled tree built from an example set of 14 sequences. Reconstructed ancestral sequences at each node are labeled with a unique name. Duplication events are marked in purple. Each node is labeled with a circle whose inner color represents the sequence's average posterior probability (orange color gradient). The level of branch support from bootstrapping analysis is denoted by the ring around each node circle (black color gradient). Branch lengths represent the average number of amino acid substitutions per site and can be estimated using the scale bar. (b, c): Ancestor summary plots written out by topiary. The black points show the probability of the most likely amino acid at each position. The distribution of these probabilities is given by the histogram on the right. The average posterior probability is the mean of these values. The red points show the probability of the second most likely amino acid at each position, with its distribution on the right. The horizontal dashed line shows the minimum PP cutoff for the altAll reconstruction. Shaded gray regions indicate gaps; vertical purple dashed lines represent ambiguously gapped positions. (b) Summary for anc4 (tetrapod LY86 ancestor) for the 14-sequence alignment (see arrow in a). (c) Summary for the equivalent ancestor from a 188-sequence alignment and phylogenetic tree for LY86/LY96.

Table 2.2: Protein families used to validate the topiary pipeline.

Protein	Taxonomic distribution	Average seed sequence length	Number of seqs in alignment	ML substitution model
Islet Amyloid Polypeptide/ Calcitonin gene-related peptide	Vertebrates	37	39	JTT+G8
S100A5 & S100A6	Amniotes	94	104	JTT+G8
Cytochrome C	All life	109	121	WAG+G8
Ribonuclease HI	Bacterial	163	181	LG+G8
LY86 & LY96	Vertebrates	164	188	VT
Micrococcal nuclease	Bacterial	200	182	LG+G8
Chalcone Synthase	Plants	390	107	DEN+G8
tight junction protein 1	Vertebrates	1705	121	JTT+G8+FO+IO

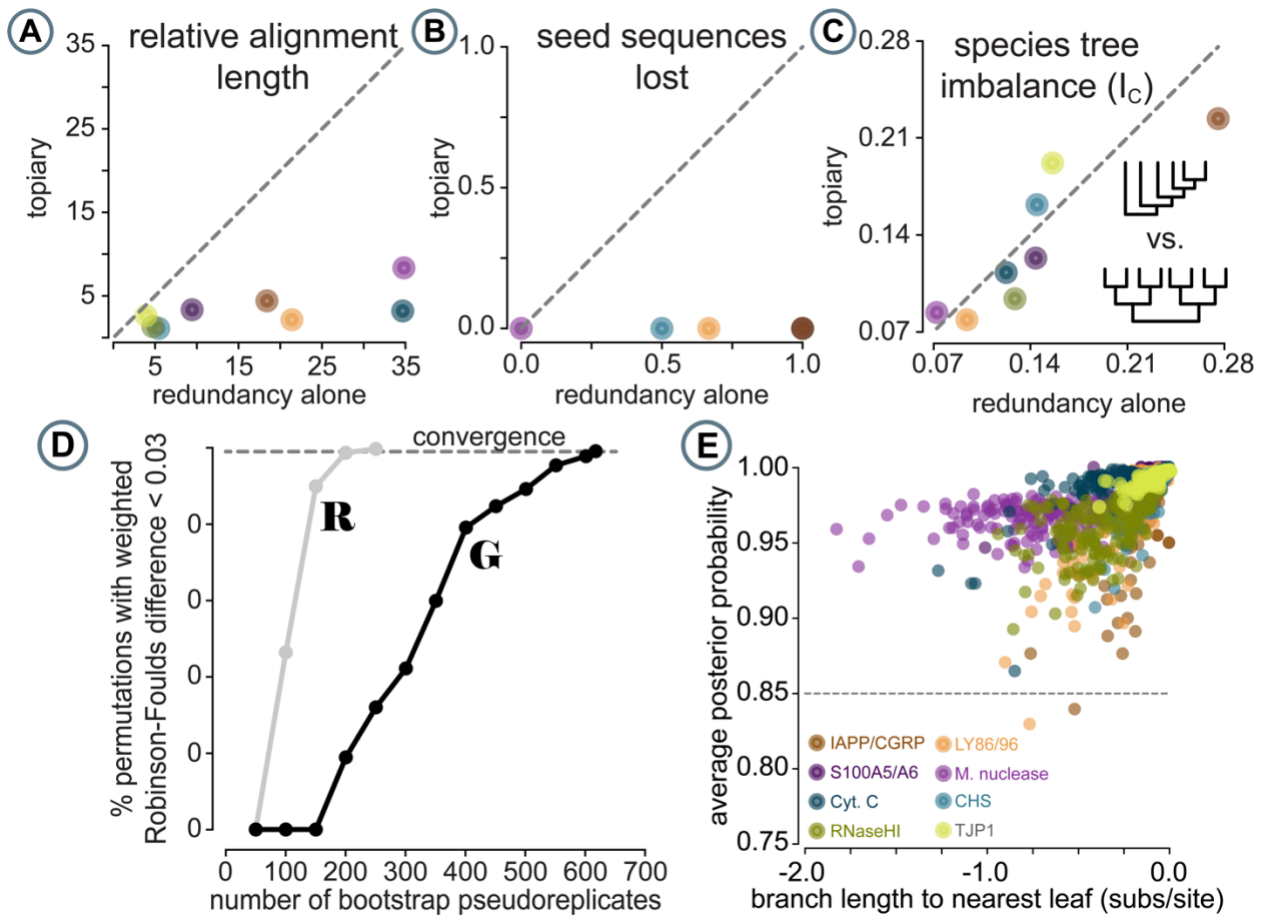


Figure 2.7. Validation of the topiary pipeline. Panels show topiary results generated for the eight protein families from Table 2. Colors indicate the family in question (see panel e for color legend). Panels a–c show topiary alignment quality as measured by three metrics: (a) Relative alignment length (number of columns in alignment divided by the average length of seed

Figure 2.7 (continued) sequences); (b) The fraction of seed sequences lost during redundancy reduction; (c) Species tree imbalance (measured by the Colless Index of the species tree for the sequences in the alignment). (d): Number of pseudoreplicates required for converged branch supports for the gene tree (G) versus the reconciled tree (R) for the LY86/LY96 family. (e) Average posterior probabilities for all ML ancestors plotted against the total branch length between that ancestor and the nearest modern sequence on the tree. More negative values on the x-axis are deeper in the tree. Posterior probability starts at 1.0 near the tips of the tree and decays for more ancient ancestors. The dashed line indicates a “rule of thumb” of 0.85 for usable ancestral sequences.

Selecting ancestors

After checking for a reasonable reconciled tree and running the *bootstrap-reconcile* script, one can identify ancestors that are amenable to reconstruction based on their average posterior probability (**Figure 2.2f**) and branch supports (**Figure 2.2g**). As shown in **Figure 2.6a**, topiary maps these values onto the final tree as color gradients. One typically wants ancestors with average posterior probabilities and branch supports above 0.85 and 85, respectively. Note that the posterior probabilities and branch supports are independent of one another. For example, ancestor 11 has high branch support (dark black circle exterior) but a low ancestral posterior probability (light orange circle interior); ancestor 4, on the other hand, has low branch support but high posterior probability. As noted in the overview section, it is important to select ancestors with both high branch support and high posterior probabilities. (Note that this tree has low supports overall because it was built from a demonstration alignment with only 14 sequences.)

In addition to summary statistics on the tree, topiary provides more detailed information about each ancestor. **Figure 2.6b,c** show minimally modified versions of graphs that topiary automatically writes out for each ancestor. **Figure 2.6b** shows site-specific posterior probabilities for the reconstructed LY86 protein from the ancestor of tetrapods, anc4 (see arrow in **Figure 2.6a**). The average posterior probability (0.825) is the mean of the black points. Some sites have unambiguous reconstructions (black points have PP = 1.0), but many other sites have plausible

alternate reconstructions with similar PP to the ML reconstruction (red). This ancestor has 31 sites that topiary classifies as ambiguous, meaning that there are 31 positions where the alternate reconstruction has a posterior probability above 0.25 (graphically, the number of red points above the dashed horizontal line). Finally, topiary reports sites for which it is ambiguous whether the position should be reconstructed as an amino acid or as a gap (site 27, for example).

We can compare the results in **Figure 2.6b** to the tetrapod LY86 ancestor returned by the pipeline for a 188-sequence alignment of LY86/LY96 sequences without manual MSA edits (**Figure 2.6c**). Upon increasing our number of sequences from 14 to 188 in the alignment, the average posterior probability for this ancestor increases significantly, from 0.825 to 0.952. We also see fewer ambiguous sites and no ambiguous gaps. Overall, this is a much higher-quality ancestor that is likely amenable to experimental characterization.

We note, however, that there are still 21 ambiguous positions with alternate reconstructions whose posterior probabilities are above 0.25. This is real phylogenetic uncertainty that is unlikely to be resolved with the addition of more protein sequences. To account for this uncertainty, we recommend experimentally characterizing both the ML protein and the “altAll” version of the same protein.¹³⁴ Topiary automatically generates both versions of every ancestor. The altAll ancestor reconstruction is made up of the ML sequence with every ambiguous ML amino acid replaced with its next most likely alternate. In other words, it selects the second-most-likely amino acid at every site where the red point is above the horizontal dashed line. For the ancestor shown in **Figure 2.6c**, the ML and altAll versions of the ancestor will differ at 21 positions.

The altAll can be thought of as “worst case” for the reconstruction, allowing one to ask what the consequences would be if the reconstruction got every ambiguous site wrong. The true,

historical ancestral sequence is likely somewhere between the ML and altAll ancestors, but more like the ML than altAll sequence. If, upon synthesis and characterization, both the ML and altAll ancestors have the same measured property, that property is robust to uncertainty in the reconstruction and likely reflects the ancestral state of the protein. In previous experiments, the altAll ancestor has behaved similarly to the ML ancestor.^{76,115,135–138}

On black boxes

Topiary automates much of the drudgery of an ASR study, going from a seed dataset to reconstructed ancestors with minimal input. One of our goals is to make the technique accessible for non-experts. It should not, however, be treated as a black box. To help users better understand what topiary does at each step, we have provided Jupyter notebooks that can either be run locally or via Google Colab that break the topiary pipelines into individual steps (<https://github.com/harmslab/topiaryexamples>). This also provides a framework for users to modify or extend the pipelines to fit their specific needs.

One final note. Generating ancestors is relatively easy, but experimentally characterizing them can take years; it is worth some caution upfront. Specifically, if the species-reconciled gene tree has a huge excess of non-speciation events, pause. Do not trust results from ancestors with low branch supports or low posterior probabilities. And, finally, characterize the robustness of experimental results to phylogenetic uncertainty using altAll versions of ancestors. Following these rules will ensure the quality of your reconstructed ancestors and thus evolutionary conclusions.

PIPELINE VALIDATION

In this final section, we describe how we validated the topiary pipeline itself. Our first level of validation is part of the software package. We developed topiary using a test-driven development framework, meaning we write test code in parallel with our functional code. As of this writing, 87% of the lines in the topiary codebase are automatically tested for correct inputs, outputs, and logic every time we update any part of the code. We paid special attention to core functions in our test development. For example, the module that interfaces with RAxMLNG has 100% test coverage. Such efforts give us confidence that the software should behave as expected.

We also validated that topiary is useful for realistic ASR studies. We solicited seed datasets from scientists studying a wide variety of proteins from different species (**Table 2.2**). This allowed us to test the pipeline on real inputs from different classes of proteins, protein sizes, and taxonomic distributions. We then ran these eight seed datasets through both stages of the pipeline. We did no manual corrections to the alignments, so these represent fully automatic outputs with no human input beyond initial seed dataset construction.

Much of what topiary does is to connect existing pieces of software. Rather than attempting to test each component, we focused our validation on the connections between components. The first step we checked was that of going from BLAST to alignment. Our BLAST/ reciprocal BLAST strategy is standard; however, topiary reduces dataset size in a novel way (**Figure 2.4**). We therefore compared topiary to a strategy that lowered redundancy using sequence identity alone. We performed BLAST/reciprocal BLAST on all eight datasets, reduced redundancy using either topiary or CD-HIT¹⁵⁶, and then aligned the resulting datasets using Muscle5. For each dataset, we selected a CD-HIT redundancy cutoff that yielded the same

number of sequences as the topiary dataset. We then compared the resulting sequence-identity-alone versus topiary datasets with three quality metrics (**Figure 2.7a–c**).

The first metric was alignment length relative to average seed sequence length. A higher value indicates the presence of long, potentially poorly aligned, sequences in the alignment. We found that topiary significantly outperformed a sequence-identity-alone approach using this metric (**Figure 2.7a**). While the sequence-identity-alone approach gave alignments up to 35-times longer than the seed sequence, the longest alignment coming from the topiary pipeline was only 5 times longer than the seed sequences. We next measured retention of key sequences. As expected, topiary never dropped key sequences from the dataset, while the simple redundancy cutoff was highly variable in this metric (**Figure 2.7b**). As a third comparison, we characterized the imbalance of the species tree corresponding to the final sequence dataset using the Colless Index¹⁵⁷ as calculated by DendroPy¹⁵⁸ (**Figure 2.7c**). Because topiary uses a taxonomically informed sampling strategy, we predicted the topiary trees would be more balanced than those from the dataset reduced by simple sequence identity. This was not true; both approaches gave similarly balanced trees for each dataset. This suggests that the tree imbalance reflects the real taxonomic diversity in the sequence databases for these proteins, rather than a problem with how that diversity is sampled to make tractably-sized datasets.

We also validated the reliability of the branch supports generated by topiary. Topiary calculates branch supports by generating pseudoreplicate gene trees in RAxML-NG, then passing them into GeneRax for reconciliation. By default, RAxML-NG generates bootstrap replicates until the supports converge on the gene tree. We wanted to verify that the branch supports on the reconciled tree converged reliably, even though the number of pseudoreplicates required was determined by convergence on the gene tree. To do this, we performed an a posteriori

convergence test on the bootstrap replicate trees generated for either the gene tree alone or the reconciled gene trees. For this, we used the RAxML-NG “—bsconverge” analysis mode with a default cutoff of 0.03.¹⁵⁹ The results for the LY86/ LY96 family are shown in **Figure 2.7d**. The gene tree required over 600 bootstrap replicates for converged branch supports; the reconciled tree required <300. We observed similar results for all eight families, with the gene tree taking more replicates to converge than the reconciled tree. This indicates that the species tree is indeed constraining the gene tree and that the bootstrap supports converge with our standard protocol.

As a final validation of the pipeline, we reconstructed all ML ancestors for the eight protein families (1027 ancestors in total). We then calculated the average posterior probability of each ML ancestor and plotted this against the branch length between that ancestor and the nearest modern protein sequence (**Figure 2.7e**). An ancestor identical to a modern protein would be plotted at zero on the x-axis; a more negative value corresponds to more substitutions per site between that ancestor and the most similar modern protein. In this plot, we observed that ancestral sequences close to the tips of the tree were better reconstructed than earlier ancestors. This is expected: more recent ancestors require less evolutionary extrapolation than more ancient ancestors. Despite the drop in quality for our deepest ancestors, however, we found that most reconstructed sequences are likely usable for reconstruction studies. Only 13 of the 1027 ancestors had average posterior probabilities below 0.90. This demonstrates that the pipeline—even without manual inspection and editing of the sequence alignment—generally yields high quality ancestral sequences.

CONCLUSION

The resources for performing high-quality ancestral sequence reconstruction already exist, but the complexity of the process and the importance of expert knowledge create a barrier to wider adoption; the topiary pipeline overcomes this barrier. It requires only that scientists define an evolutionary question and scope, and then lets computers do the rest, integrating powerful existing software to give users useful output for reconstructing and evaluating ancestral sequences. We hope this will improve the quality of ASR studies by codifying best practices and will increase the accessibility of the technique for protein scientists from a wide variety of backgrounds.

BRIDGE TO CHAPTER III

With this topiary ancestral sequence reconstruction tool in hand, we were able to reconstruct and characterize the bony vertebrate, tetrapod, and teleost ancestral Toll-like receptor 4 complexes used in the next chapter. Being able to resurrect and functionally characterize these key ancestral states was pertinent to revealing the origin of difference in ligand specificities between the human and zebrafish TLR4 complexes.

CHAPTER III

TOLL-LIKE RECEPTOR 4 EVOLUTION OF LPS SPECIFICITY IN EARLY VERTEBRATES AND DIVERGENCE IN ZEBRAFISH

*This chapter contains unpublished co-authored material.

Author contributions: Orlandi KN and Harms MJ designed the study. Orlandi KN designed and performed experiment, analyzed the data, and wrote the text. Harms MJ was the funding acquisition lead and oversaw the experiments and writing. Sánchez-Borbón J constructed the CD14 ancestors used in the study and contributed input on experiments. Brown C performed site-directed mutagenesis on several plasmids and executed one of the experiments included in this text. Robinson C helped to execute the zebrafish experiments. Guillemin K contributed experimental guidance.

ABSTRACT

Toll-Like Receptor 4 (TLR4) plays a pivotal role the innate immune system in humans by activating the inflammatory response to lipopolysaccharide (LPS) from Gram-negative bacteria. Dysregulation of TLR4 activation can cause excessive inflammation resulting in myriad health problems including sepsis, heart disease, chronic arthritis, and other conditions. Much of our understanding about inflammation comes from careful studies of model organisms. One powerful model is the zebrafish, *Danio rerio*, which is often used to study mechanisms of human disease and host-microbe interactions. It was recently discovered that zebrafish express a functional TLR4. This could allow zebrafish to be a valuable, tractable model for the human innate immune response. This would require that we can relate zebrafish and human receptor functions. Here, we explored the function of zebrafish TLR4 *in vivo* and *in vitro*. We discovered that zebrafish TLR4 is activated *in vitro* by a class of LPS molecules that antagonize the human receptor, and that a unique structural feature is necessary for its activity. The mechanism of TLR4-induced inflammation *in vivo* also appears to be different than in humans. To understand the evolutionary context for the disparity between human and zebrafish TLR4 specificity, we used ancestral sequence reconstruction to infer and resurrect ancestral vertebrate TLR4 proteins and functionally compared them to several modern species. The results suggest complicated, species-dependent evolutionary trajectories originating from a low-sensitivity ancestral TLR4. Overall, this work will help guide future investigations using the zebrafish model of innate immunity by providing insight into the divergent functional roles of zebrafish and human TLR4.

INTRODUCTION

Toll-like receptor 4 (TLR4) is known to play a central role in the human immune defense against pathogens. TLR4 is a member of the Toll-like receptor (TLR) family of pattern recognition receptors. TLRs are type I transmembrane proteins expressed on innate immune cells and conserved across vertebrates. They recognize evolutionarily conserved molecules associated with danger and stimulate intracellular signaling cascades that activate the host immune response. TLR4 is responsible for discriminating host lipids from lipopolysaccharide (LPS), a component of Gram-negative bacteria outer.²⁵⁻²⁷ TLR4 was discovered because of its role mediating sepsis; sepsis occurs when the body produces excessive inflammation that damages host tissues in response to an infection. In 2017, sepsis accounted for almost 20% of all global deaths.¹⁶⁰ TLR4 also contributes to the onset and progression of other illnesses such as cancer, atherosclerosis, osteoarthritis, and Alzheimer's disease.^{161,162} Because of its importance to human health, TLR4 is an immune protein of major interest.

TLR4 is only able to sense LPS by forming a heterodimer with a cofactor protein myeloid differentiation factor-2 (MD-2).³⁶ In the absence of LPS, MD-2 forms a stable heterodimer with the extracellular domain of TLR4 and its expression is important for the correct distribution of TLR4 to the cell membrane.¹⁶³ The beta-cup structure of MD-2 creates a hydrophobic pocket which accommodates the hydrophobic fatty acyl chains of LPS in crystal structures.^{41,42,46} MD-2 also forms the interface for dimerization required for TLR4 activation and its positioning of LPS inside the binding pocket is critical^{42,164} (**Fig 3.1A-B**).

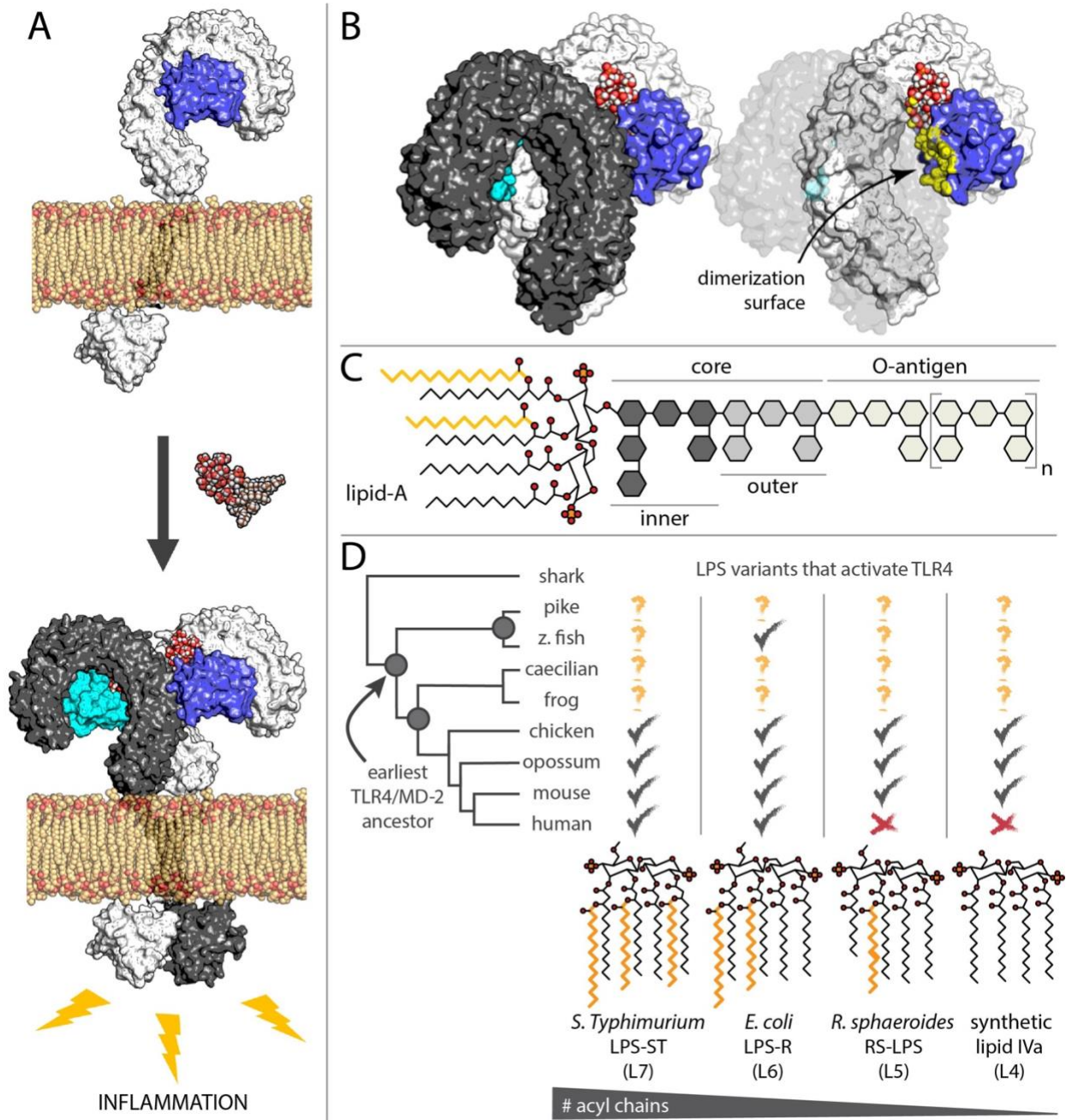


Figure 3.1. Current knowledge of the evolution of TLR4/MD-2 LPS specificity. **A)** TLR4 is shown in gray/white; MD-2 in cyan/blue. LPS (to the right of arrow) induces dimerization of TLR4/MD-2, triggering inflammation. **B)** LPS binds in a deep pocket in MD-2, creating a new dimerization surface (yellow). Right panel makes front TLR4 transparent to reveal interface. **C)** Structure of LPS from *E. coli*. **D)** Phylogenetic tree showing the evolution of TLR4/MD-2 with known activities of extant species. We aimed to characterize agonist specificity of TLR4 from extant species that would provide insight for species-specific TLR4 ligand specificity (sharks have no TLR4) and three ancestors (gray circles).

The transfer of LPS into the binding pocket of MD-2 is catalyzed by the presence of LPS-binding protein (LBP) and cluster of differentiation 14 protein (CD14). LBP is able to bind LPS-rich surfaces like bacterial membranes, somehow altering the membrane, which permits CD14 to bind monomeric LPS.¹⁶⁵⁻¹⁶⁷ CD14 shields the hydrophobic acyl chains as it chaperones the LPS molecule to the binding pocket of MD-2.¹⁶⁸ CD14 seems to be important in TLR4's detection of LPS, but it may not be necessary. CD14-deficient mice are resistant to doses of LPS that are lethal to wild-type mice or would induce cytokine expression, but still respond to high doses of LPS.¹⁶⁹ CD14 likely serves a more complex role in LPS signaling than simply LPS chaperone.^{166,167,170}

Ligand-binding drives the dimerization of two TLR4/MD-2 complexes, which brings the two TLR4 intracellular TIR domains together to act as a scaffold for adaptor proteins involved in MyD88 and TRIF signaling^{170,171} (**Fig 3.1A**). MyD88-mediated TLR4 signaling occurs mainly at the plasma membrane and results in the activation of transcription factor NF- κ B and induction of proinflammatory cytokines like TNF α and IL-6. TRIF-mediated signaling occurs at the endosomal membrane after internalization of TLR4, which further activates IRF3 and the production of type-1 interferons and other IRF3-dependent genes, as well as delayed NF- κ B activation.^{170,172,173} Structural modifications to LPS have been demonstrated to differentially activate these pathways.¹⁷⁴⁻¹⁷⁶

LPS is an essential structural feature of Gram-negative bacteria outer membranes.²² It exhibits a high degree of structural diversity but generally consists of three components: a highly variable O-antigen, a less variable core oligosaccharide, and a highly conserved lipid A (**Fig 3.1C**). The lipid A moiety of LPS is the structural feature recognized by TLR4/MD-2 and therefore accounts for most of the immunostimulatory effects of LPS.¹⁷⁷ Structural and

functional analyses show that the most proinflammatory forms of lipid A, which are generally purified from *Escherichia coli* and *Salmonella* strains, have two phosphate groups and six fatty acyl chains each with 12-14 carbons.³⁴ Lipid A with more or less fatty acyl chains, longer chains, or a single phosphate group are typically less active and can act as antagonists of toxic LPS.^{34,35}

Bacteria have built in pathways to modify their lipid A structures in response to changing environmental conditions. Modifications occur through constitutive and regulated processes in response to external stimuli including changes in growth condition (temperature, nutrient, osmolarity), host detection (conversion of host immune agonist to antagonist), and antimicrobial molecules (modulation of surface exposed negative charge, deacylation of their outer membrane).¹⁷⁸ Several studies of pathogens converge on the theme that alteration to lipid A is a common virulence strategy adopted by bacterial pathogens to evade host innate immune detection.^{49,51-56,170} There are also examples of the order Bacteroidales, a human gut commensal bacteria, that produce tetra- and penta-acylated LPS that can silence TLR4 signaling for the whole microbial community, potentially facilitating host tolerance of a healthy adult microbiome.^{47,48}

To understand the implications of this mutable host-microbe interface and how to develop treatment strategies for infections and disease requires leveraging studies in model systems. There are many inflammatory disease models in zebrafish (*Danio rerio*) which have been widely employed in immune system, host-microbe interaction, and drug discovery studies. Zebrafish are uniquely advantageous for research in these fields because as larvae they are optically transparent, enabling imaging in live organisms, and their microbiota can be easily manipulated. TLR4-induced inflammation has not yet been included in these models. Zebrafish have been shown to have an inflammatory response to LPS. However, the zebrafish response to LPS is

much weaker than in humans or mice. It was widely accepted that this was due to gene loss of the TLR4 cofactors MD-2 and CD14.¹⁷⁹

Recently, the zebrafish MD-2 gene was discovered and shown to be expressed in immune cells.¹⁰¹ Furthermore, the zebrafish TLR4/MD-2 complex can be activated by LPS *in vitro*, although CD14 is required. MD-2 mutant zebrafish exhibited perturbed transcriptional responses to LPS challenge but did not show improved tolerance to LPS-induced death as observed for mice.¹⁰¹ These critical findings suggest that TLR4/MD-2 could play a role in LPS sensing in the zebrafish, but other pathways are likely also involved. This work has paved a way towards developing the zebrafish model of TLR4 inflammation.

Here, we use an evolutionary lens to try to better understand the role of zebrafish TLR4 in innate immunity. It has been shown that mammalian TLR4 is lowly responsive to “foreign” LPS from the deep sea *Moritella* genus of bacteria potentially due to acyl chain length.¹⁸⁰ The authors posit that pattern recognition strategies may be defined by local environment rather than universal threats. We hypothesized, therefore, that differences in evolutionary pressures, like distinctive pathogenic bacteria present in a terrestrial versus tropical freshwater aqueous environment, could have led to divergent LPS specificities in zebrafish and mammals. To test this, we employed a broad range of LPS variants in a functional assay against zebrafish TLR4/MD-2 and discovered that the complex is uniquely sensitive to tetra-acylated lipid A. We also found that the lineage of teleost fish including zebrafish evolved a novel MD-2 C-terminal peptide that is essential for TLR4 signaling in functional assays.

Next, we tested whether the heightened sensitivity to tetra-acyl lipid A we found for zebrafish TLR4/MD-2 *in vitro* would translate to a stronger immune response in the zebrafish immune system *in vivo* compared to previous studies. On the contrary, we find no elevated

immune response to tetra-acyl lipid A relative to the vehicle control or when fish were pre-treated with a TLR4-specific inhibitor. We determine that although zebrafish maintain functional copies of TLR4 and MD-2 with the ability to recognize and respond to various LPS structures, there is significant divergence in the human and zebrafish immune response to LPS. More studies will need to be done to define what role TLR4 might play in zebrafish innate immunity.

We remained curious about the evolutionary origin of the difference in human and zebrafish TLR4/MD-2 specificity observed *in vitro*. Does the zebrafish state represent an ancestral state that was modified along the tetrapod lineage? Or did teleost fish lose high ancestral activity? To explore the specificity of evolutionary intermediates between human and zebrafish, we selected key modern species and reconstructed ancestors to compare in a functional assay (**Fig 3.1D**). We used ancestral sequence reconstruction to infer the ancestral states for TLR4 and MD-2 protein sequences and then resurrected these proteins for our assay. We find that other fish, amphibians, and early vertebrate ancestral complexes exhibit low sensitivity to all ligands tested. We infer that the zebrafish TLR4 lineage evolved heightened LPS sensitivity with unique specificity for tetra-acyl LPS. Interestingly, the tetrapod ancestor is highly active in the absence of ligand and can be stimulated with all ligands tested. This suggests that tetrapods may show increased TLR4 stimulation relative to early branching vertebrates due to sequence changes on the evolutionary trajectory from the ancestor of bony vertebrates to tetrapods.

RESULTS

Zebrafish TLR4/MD-2 is potently activated by tetra-acyl LPS *in vitro*

We started by assessing differences between human and zebrafish TLR4 specificities for structural variations of the lipid A portion of LPS. Depending on the bacterial species, the lipid A

moiety can have between four and eight acyl chains, each with different abilities to activate TLR4/MD-2.¹⁸¹ Loes et al. revealed mild activity when zebrafish TLR4 was challenged with hexa-acylated LPS molecules *in vitro*, and similar results for hexa- and hepta-acylated LPS challenge *in vivo* by immersion or cardiac ventricular injection.¹⁰¹

We challenged human and zebrafish TLR4/MD-2 complexes in an *in vitro* functional assay with commercially available LPS variants. These are generally complex mixtures of LPS structures, so we report our findings based on the most abundant LPS structure stated by the supplier. We used *Salmonella enterica* Typhimurium LPS ((L7)-LPS-ST) to represent our hepta-acyl chain variant, *Escherichia coli* K12 LPS ((L6)-LPS-EK) is our hexa-acylated variant with an O-antigen, *Rhodobacter sphaeroides* LPS ((L5)-LPS-RS) for a penta-acylated structure, and synthetic lipid IVa ((L4)-lipid IVa) to represent tetra-acylated LPS (**Fig 3.2**).

For the TLR4 functional assay, we transfected HEK293T cells with plasmids containing either the human or zebrafish TLR4 complex components under constitutive promoters, as well as a luciferase reporter gene under control of the NF- κ B transcription factor. Since zebrafish do not have a CD14, we included a mouse CD14 plasmid which confers zebrafish TLR4 the greatest sensitivity to LPS.¹⁰¹ HEK293T cells do not endogenously express the TLR4 complex but they do have the capacity to mount an NF- κ B-mediated response to TLR4 activation. The following day, we treated cells with each of the LPS variants described, incubated the cells in treatment media for four hours to allow a robust transcriptional response, and then measured the amount of luciferase enzyme activity associated with each condition. The quantity of luciferase enzyme should be directly proportional to the level of NF- κ B activation initiated by TLR4. To account for differential expression of immune receptors on the surface of cells and activation capacity, we

normalize the observed signal to that of (L6)-LPS-EK for human TLR4 and (L4)-lipid IVa for zebrafish TLR4, unless otherwise noted.

Our *in vitro* experiments (Fig 3.2) show that zebrafish TLR4/MD-2 exhibits a robust response to challenge with tetra-acyl lipid IVa, a low-level response to hexa-acyl LPS-EK, and little to no activity in the presence of hepta- and penta-acylated LPS variants. This specificity contrasts human TLR4, which is inhibited by lipid IVa and strongly activated by both hexa- and hepta-acylated LPS.

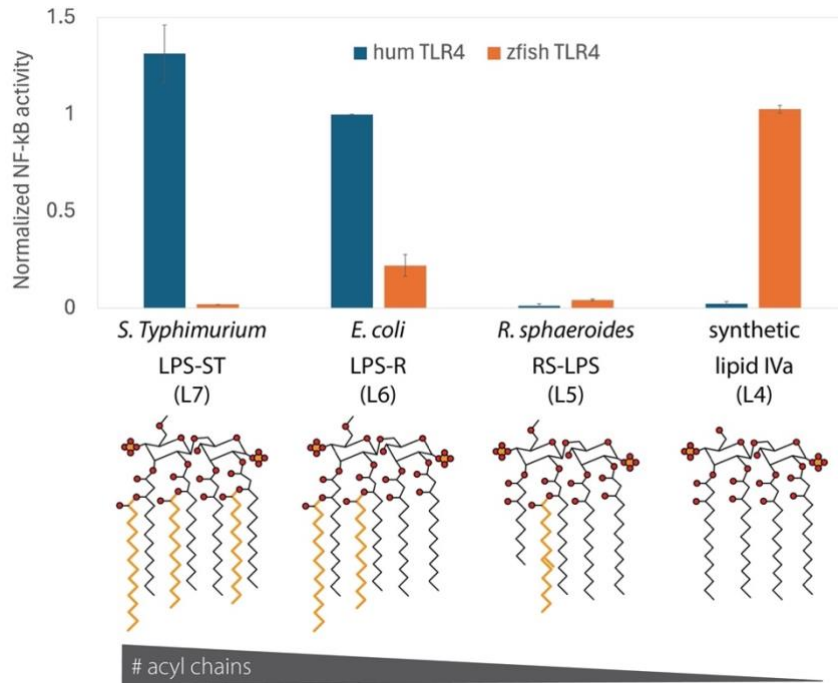


Figure 3.2. Revealing differences in LPS specificity between human and zebrafish TLR4 complexes. In vitro NF-κB activation by human TLR4 (blue) and zebrafish TLR4 (orange) challenged with LPS variants. LPS variants include a gradient of lipid A acyl chain number, from 7 acyl chains (left) to 4 acyl chains (right). NF-κB activity level is shown relative for each species: LPS-EK for human, lipid IVa for zebrafish. These four LPS variants cover a range of chemical features and are readily available commercially. Yellow indicates additional acyl chains relative to lipid IVa. Human TLR4 was treated with 0.1 ng/uL LPS-ST and LPS-EK. All other conditions show treatments at 1 ng/uL LPS. Error bars indicated standard error of the mean across several experiments.

Zebrafish have three TLR4 ohnologs (gene duplicates originating from whole genome duplication): tlr4ba, tlr4bb, and tlr4al which are all expressed in immune cells of larval fish.¹⁰¹ Our previous experiment used the tlr4ba protein because Loes et al. had determined it was the only ohnolog that could respond to (L6)-LPS-EK. Now that we had found a more potent agonist of tlr4ba, we investigated whether the other zebrafish TLR4s or heterocomplexes of the ohnologs could be activated by LPS variants (**Fig 3.3A-B**). Slight variations to transfection conditions and LPS treatment concentrations between the experiments shown in 3.3A and B are noted in the figure caption. Our preliminary experiments did not identify agonists for homocomplexes of tlr4bb or tlr4al. They did, however, suggest that tlr4bb may enhance the signal of tlr4ba when co-expressed *in vitro*. Although this may play a physiological role in the fish, we did not further pursue characterization of tlr4ba/bb complexes because tlr4bb appeared to enhance sensitivity without altering specificity of tlr4ba. Further analysis should be done to determine whether this initial observation is physiologically relevant, or if other heterocomplexes could be informative of zebrafish TLR4 agonist specificity. In the rest of the chapter, we will use zebrafish tlr4ba and TLR4 interchangeably.

Discovering an agonist of zebrafish TLR4 brought up several questions including: What structural features of zebrafish TLR4/MD-2 confer this altered specificity? Is this specificity reminiscent of the vertebrate ancestral TLR4, or did it evolve on the zebrafish lineage? Can we use this agonist to further probe ambiguity in the role of zebrafish TLR4 in innate immunity? Can this difference in specificity explain previous observations of low-level immune responses of zebrafish to challenge with LPS-ST and LPS-EK?¹⁰¹

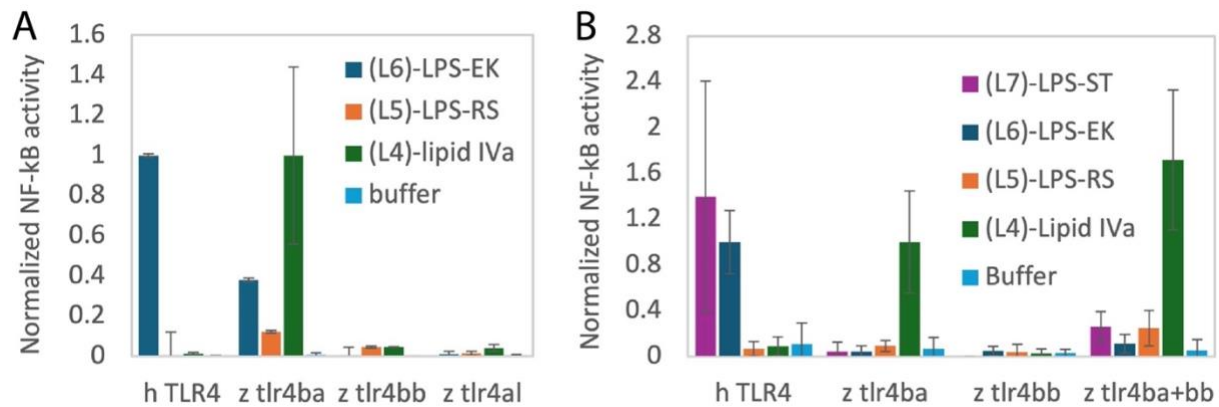


Figure 3.3. Zebrafish ohnologs *tlr4bb* and *tlr4al* do not respond on their own to LPS variants *in vitro*. NF-κB activation of human and zebrafish TLR4 paralogs in response to LPS variant challenge. LPS variants include a gradient of lipid A acyl chain number, from 7- to 4-acyl chains, indicated by colors in the legends. HEK293T cells were transfected with human TLR4 or zebrafish TLR4 ohnologs: *tlr4ba*, *tlr4bb*, or *tlr4al*. **A)** Zebrafish complexes were made by transfecting TLR4:MD-2:mouse CD14 with 25:20:1 ng plasmid/well. All LPS concentrations were 2 ng/μL, except human TLR4 treated with LPS-EK was at 0.2 ng/μL. **B)** Similar to panel A but including a co-transfection with *tlr4ba* and *tlr4bb* (far right) and a 7-acyl chain LPS (purple). For this experiment, zebrafish complexes were made by transfecting TLR4:MD-2:human CD14 with 25:25:1 ng plasmid/well. All LPS variants were used at 0.2 ng/μL. For both panels, NF-κB activity level was buffer-subtracted and is shown normalized by species: human TLR4 signal is normalized to human TLR4 treated with LPS-EK; zebrafish TLR4 signal is normalized to zebrafish *tlr4ba* treated with lipid IVa. Error bars indicate standard deviation for technical triplicates of a single experiment.

A subset of teleost fish evolved a functionally necessary MD-2 C-terminal peptide

We first sought to understand the structural origin of these divergent LPS specificities. We used *topiary*¹⁸², the bioinformatic phylogenetics pipeline discussed in Chapter II, to gather TLR4, MD-2, and their paralog amino acid sequences from the NCBI database, sampling a wide taxonomic spread, and then aligned these sequences. In our multiple sequence alignment generated for MD-2 there was an anomaly in Cyprinidae, a family of ray-finned fish including zebrafish. At the C-terminus, there was an extension of roughly ten amino acids (**Fig 3.4C**).

We used the AlphaFold2 structure prediction tool^{183–185} to predict the structure of the zebrafish TLR4 extracellular domain in complex with MD-2, with and without this C-terminal peptide (**Fig 3.4A**). At least in the absence of LPS, the C-terminal peptide was predicted to form

a flexible linker connected to a small amphipathic alpha helix that fits snugly inside the hydrophobic LPS-binding pocket. The crystal structure of human TLR4/MD-2 bound to *E. coli* LPS-Ra, which is similar to LPS-EK but has no O-antigen, is shown on the left in **Fig 3.4A**. TLR4 is in cyan, MD-2 is yellow, the LPS core oligosaccharide and lipid A diglucosamine group are shown in orange and the hydrophobic acyl chains in pink. Comparing this structure to the predicted zebrafish TLR4/MD-2 structure to the right reveals the peptide in the LPS-binding pocket of MD-2.

The predicted zebrafish TLR4/MD-2 structure is colored by pLDDT to show per residue confidence in the structure, with high confidence in red and low confidence in blue. Most of the TLR4 leucine rich repeat (LRR) domain was predicted with high confidence, which is reasonable given there are many crystal structures of LRR domains. MD-2 was also predicted with high confidence except at the C-terminus which consists of the full C-terminal extension (amino acids GGNKSFFSPQIGRL). We can't be certain that this peptide sits within the binding pocket, but zooming in on the peptide shows that there are nonpolar groups that could associate with the inside of the MD-2 beta-cup while exposing hydrophilic residues to the solvent (**Fig 3.4B**).

Because the C-terminal peptide is at the opening of the MD-2 binding pocket, we hypothesized that it could be a structural feature of zebrafish MD-2 that defines LPS acyl chain number specificity. To test this hypothesis, we made a series of MD-2 C-terminal truncation mutants (**Fig 3.4C**). We chose two cut sites that seemed relevant to other species in the MD-2 alignment and the predicted structure. We made a cut after residue 139 (139 Δ) to mimic the frog sequence in the dataset and eliminate the amino acids associated with low structural confidence. The second truncation was after position 142 (142 Δ) to match the chicken, mouse, and human proteins which have been well characterized. We made three additional distal truncations that

included two more amino acids in each step to evaluate hydrophobic and length requirements of the peptide. Nomenclature refers to the amino acid position after the signal peptide is cleaved.

We transfected cells with zebrafish TLR4 and either full-length (FL:154 amino acids) MD-2 or one of the truncation mutants and challenged them with (L7)-LPS-ST, (L6)-LPS-EK, or (L4)-lipid IVa. We included a wild-type human complex transfection as a positive control for LPS variant treatments. Our results show that the entire zebrafish MD-2 C-terminus is necessary for zebrafish TLR4 activation (**Fig 3.4D**). The signal from zebrafish TLR4/MD-2₁₅₄ (FL) is not rescued by any of the truncation mutants, and we do not see emergence of new specificity. We were curious, then, if this loss-of-function is specific to zebrafish TLR4.

We tested whether human TLR4 would differentially respond to LPS challenge when expressed with either full-length or zebrafish MD-2_{142Δ} mutant (**Fig 3.4E**). We transfected cells with human TLR4, wild-type or truncated zebrafish MD-2, and either human or mouse CD14 and treated with LPS variants (**Fig 3.4E**). The human TLR4/wild-type zebrafish MD-2 complex had not yet been characterized in our assay. In **Figure 3.4E** we show that human TLR4 expressed with full-length zebrafish MD-2 and human CD14 mounts little to no immune response to LPS-EK or lipid IVa. However, in the presence of mouse CD14 there is some signal in response to LPS-EK. Intriguingly, zebrafish MD-2_{142Δ} conferred a slight increase in human TLR4 response in the presence of both human and mouse CD14. This suggests the peptide is not necessary for zebrafish MD-2 to bind LPS for TLR4 signaling but is specifically required for zebrafish TLR4. We infer that cyprinid TLR4 and MD-2 have co-evolved to make this C-terminal extension functionally required for activation of TLR4 in this family of fish. We hope to further investigate the role of the cyprinid C-terminal peptide in TLR4/MD-2/LPS dimerization.

Live zebrafish exhibit reduced immune response to lipid IVa compared to *E. coli* LPS

Zebrafish have been shown to have a high tolerance to challenge with hexa- and hepta-acylated LPS but succumb to high doses.^{101,186} We suspected that this high dose is required either because the hexa- and hepta-acylated LPS molecules are poor zebrafish TLR4 agonists, or because commercially available LPS have low-level lipoprotein and peptidoglycan contaminants that can activate the MyD88-mediated inflammatory response via TLR2. Based on our *in vitro* findings, we hypothesized that if zebrafish TLR4 was playing a major role in the immune response to LPS in these previous experiments then we would see a stronger immune response to challenge with lipid IVa, an ultrapure and potent zebrafish TLR4 agonist *in vitro*.

We used two different LPS delivery routes to probe this question: microgavage into the gut and hindbrain injection into the circulatory system. We did not try delivery by submersion which is often used to test survival to high dose LPS due to the high cost of synthetic lipid IVa. To test the immune response dependence on TLR4, we used the TLR4-specific inhibitor TAK-242 (resatorvid) that has been used in clinical trials.^{187,188} We had also hoped to use both standard purification and ultrapure *E. coli* LPS to test whether we would see a difference in immune stimulation, but for lack of time we have not tried this yet. We used fish at 6 days post-fertilization (dpf) because TLR4 was shown to be upregulated in immune cells starting at 5 dpf.¹⁰¹

We first challenged 6 dpf fish by oral microgavage which involves moving a blunt needle tip into the fish mouth and down its throat to the top of the gut bulb where we dispensed 4.6 nL of 1 mg/mL standard purification (L6)-LPS-0111:B4, (L4)-lipid IVa, or vehicle treatment (**Fig 3.5A**). For this experiment we tracked the immune response by using a fish line with GFP under the TNF α cytokine promoter (*tnfa*:GFP) and mCherry-marked macrophages (*mpeg*:mCherry).

We imaged live fish 6 hours post-gavage using fluorescence stereo microscopy to assess the location and abundance of mCherry and GFP. After imaging, fish were immediately sacrificed and fixed in paraformaldehyde for subsequent staining and quantification of neutrophils associated with gut tissue.

Figure 3.5A shows a representative image of GFP signal at the distal gut region of a fish treated with LPS-0111:B4. For each fish, we quantified the total GFP intensity and the quantity of GFP-positive (GFP+) pixels in the distal gut and divided the total GFP intensity by number of GFP+ pixels. **Figure 3.5B** shows the total GFP intensity averaged across all GFP+ pixels normalized to vehicle treated fish from two experiments. Unexpectedly, there was a significant decrease in *tnfa:GFP* expression for fish gavaged with LPS-0111:B4 relative to vehicle treatment (p value: 0.043). Lipid IVa treated fish did not appear different from control or LPS-0111:B4 treated fish. The mCherry signal was too low for quantification. **Figure 3.5C** shows the neutrophil quantification of dissected gut tissue from one experiment. Because this quantification was subjective, we had two individuals do treatment-blind neutrophil counting. Counter 1 (blue) did not observe any difference between treatments. Counter 2 (orange) observed significant increases for both LPS-0111:B4 and lipid IVa treated fish relative to vehicle treatment (p values: 0.029 and 0.008, respectively). No difference was observed between (L6)-LPS-0111:B4 and (L4)-lipid IVa conditions.

The results of our microgavage experiments indicate that, relative to vehicle treatment, hexa-acylated and tetra-acylated LPS delivered directly to the gut do not increase TNF α expression in the distal gut but may play a role in neutrophil infiltration to gut tissue (**Figure 3.5B-C**). There were no observable differences prompted by different LPS variants. Because

these results did not match previous observations of increased inflammatory responses to LPS-EK, we turned to a different LPS delivery method.

Injection of 5 ng *E. coli* LPS into the brain tectum of 4 dpf larval fish has been shown to result in the systemic distribution of LPS and induction of liver-associated immune responses.¹⁸⁹ Neutrophil and macrophage infiltration into the liver after LPS injection was orchestrated by a MyD88-dependent inflammatory response.¹⁸⁹ An experimental system with systemic LPS distribution through the bloodstream seemed that it would better model the role of TLR4s in sepsis compared to LPS delivery to the gut that is primed with modes for detoxifying LPS.¹⁸⁶ Like before, we hypothesized that if TLR4 plays a role in the zebrafish immune response to LPS, then injecting larval fish with lipid IVa would cause a stronger immune response than when treated with LPS-EK.

For our injection experiment, we chose to test TLR4-specific activity using an orthogonal approach by pre-treating fish with the TLR4-specific inhibitor, TAK-242 (resatorvid).¹⁹⁰ To our knowledge, the effects of TAK-242 on zebrafish or zebrafish TLR4 have not previously been reported. We first confirmed that TAK-242 knocks down zebrafish TLR4 signaling *in vitro* (**Fig 3.5F**). Then we tested the effects of a range of TAK-242 concentrations on larval zebrafish. We did not anticipate that TAK-242 would have detrimental health effects because there are none reported for mouse or human. We found, however, that submersion in 83 μ M or more TAK-242 in embryonic media was toxic to 5 dpf fish. These fish showed signs of yolk sac edema and tissue degradation. Fish treated at 8.3 μ M or less TAK-242 showed no signs of toxicity when assessing heart rate, behavior, and general appearance. Our *in vitro* work showed that short treatments with 1 μ M of TAK-242 is sufficient to block subsequent immune stimulation of

zebrafish TLR4 by lipid IVa. We inferred that 8.3 μM TAK-242 taken up by fish through flask water over 24 hours would be sufficient to block zebrafish TLR4 signaling.

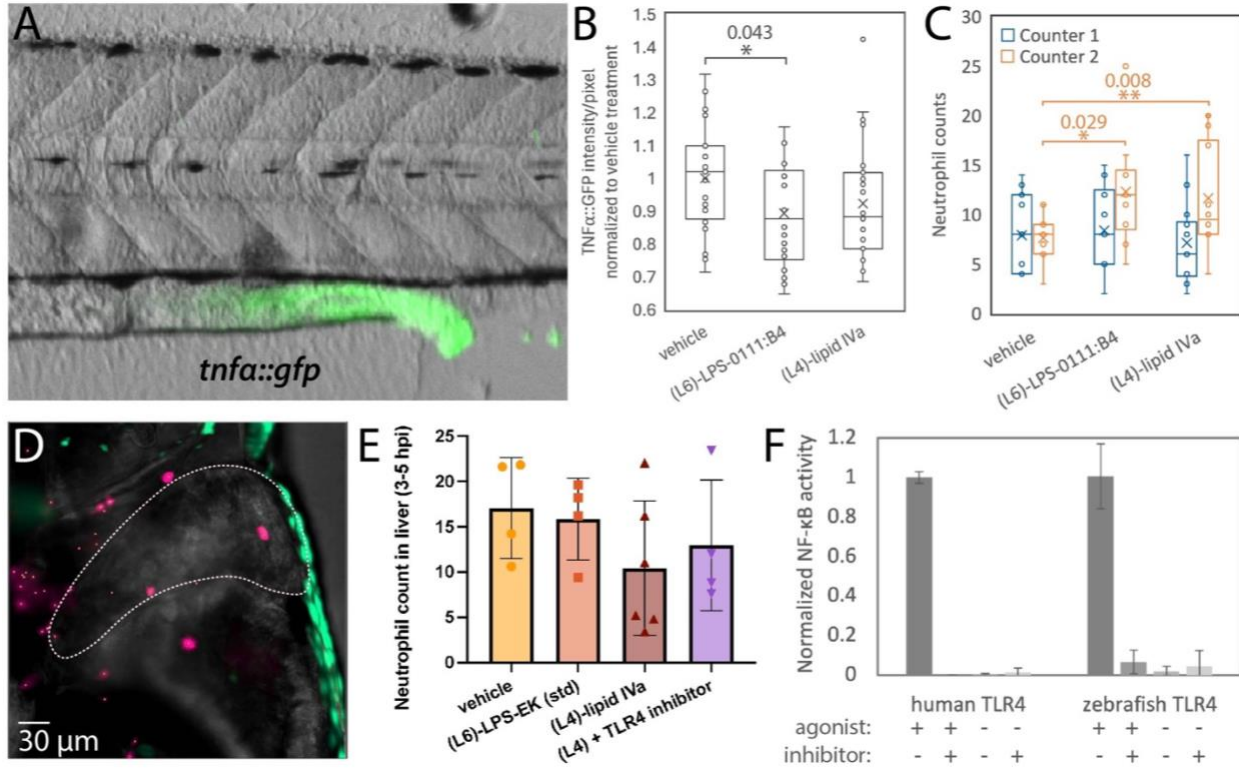


Figure 3.5. Zebrafish response to (L6)-LPS-EK and (L4)-lipid IVa challenge via microgavage and hindbrain injection. **A-C)** Immune response 6 hours after challenge with LPS via microgavage in 6 dpf live fish with transgenes *tnfa::GFP* and *mpeg1::mCherry* (macrophages). **A)** A representative fluorescence microscopy image of GFP signal in the distal gut after challenge with LPS-0111:B4. **B)** Plotting the results of two independent experiments. Each datapoint represents the total GFP signal divided by the quantity of GFP+ pixels in the distal gut region of a single fish and normalized to the average for vehicle-treated controls. **C)** Gut-associated neutrophils counted by two individuals (blue and orange) in a single experiment. An “X” indicates the mean value for each treatment. The asterisk denotes a significant difference determined by a t test; p values indicated on graphs. **D-E)** Hepatic immune response 3-5 hours after challenge with LPS via hindbrain injection in 6 dpf live fish with transgenes *tnfa::GFP* and *mpx::mCherry* (neutrophils). **D)** A representative maximum intensity projection of a mock-treated fish (vehicle) 4 hours post-injection (hpi) with the liver region used for analysis outlined with white dashes and neutrophils labeled with 1 μm spheres. **E)** Neutrophil counts in the liver of fish across two experiments 3-5 hpi with LPS-EK, lipid IVa, or vehicle (PBS). All fish were pre-treated for 24 hours either with TLR4-specific inhibitor (TAK-242) or with volume-matched mock treatment (DMSO). **F)** HEK293T cells expressing either human or zebrafish TLR4 complexes and treated with agonist (LPS-EK for human; lipid IVa for zebrafish) with or without pre-treatment with TLR4-specific inhibitor (TAK-242).

We pretreated 5 dpf fish by submersion in either 8.3 μ M TAK-242 in DMSO or an equal volume of DMSO alone (final concentration of 0.003% DMSO) in embryonic media for 24 hours. The following day, 6 dpf fish were anesthetized then injected with \sim 4.2 nL of 2.17 ng/nL LPS-EK, lipid IVa, or ultrapure PBS control into the brain tectum. After injection, fish were rinsed and recovered in fresh embryonic media until imaging. We imaged fish 3-6 hours post injection (hpi) using fluorescence light sheet microscopy. The fish strain used had transgenes *tnfa:GFP* to report on TNF α cytokine expression and *mpx:mCherry* to mark neutrophils. We collected 5 sets of fluorescent z-stack images over the liver region through the full width of each fish to evaluate neutrophil infiltration of the liver. **Fig 3.5D** shows the maximum intensity projection in GFP and mCherry channels for one set of z-stack images of a mock-treated fish imaged 4 hpi.

We are currently analyzing these data and plan to compare the number of neutrophils (pink) associated with the liver (white dashed line) between conditions. Our preliminary observations in **Fig 3.5E** indicate there are no statistically significant differences between any treatment. From visual observations of data not yet analyzed, there might be increased neutrophil infiltration of the liver for fish treated with standard purification *E. coli* LPS ((L6)-LPS-EK) and decreased levels for fish treated with (L4)-lipid IVa compared to other conditions. If these observations are supported by our analysis, we would be interested in investigating whether TAK-242 inhibits immune stimulation by standard purification *E. coli* LPS to determine if this response is TLR4-dependent. We would also like to test whether ultrapure hexa-acylated LPS, which lacks potential agonists of other immune receptors, is an agonist or antagonist for the zebrafish immune response.¹⁹¹ This would further shed light on our conflicting observations *in vitro* and *in vivo* for lipid IVa-induced TLR4 activation.

These investigations of the *in vivo* role of zebrafish TLR4 support the impression that the human and zebrafish innate immune response to LPS are quite different. However, we find it fascinating that the evolution of zebrafish and other species with low LPS sensitivity would maintain the TLR4 complex genes for hundreds of millions of years if not to use them. If these the lineage leading to zebrafish has not used TLR4 throughout evolution, then it is astonishing to find that the zebrafish TLR4 and MD-2 sequences have not diverged so much to abolish the ability to activate in the presence of LPS. To understand this evolutionary history better, we investigated whether there are evolutionary trends in TLR4 ligand specificity or sensitivity across previously uncharacterized early diverging vertebrate species.

Zebrafish and human TLR4 evolved from an ancestor with low LPS sensitivity

To explore the sensitivity and specificity of TLR4 evolutionary intermediates between human and zebrafish, we selected key modern species and reconstructed ancestors to compare in a functional assay (**Fig 3.6A**). We used *topiary*¹⁸², an ancestral sequence reconstruction pipeline, to infer the most probable amino acid sequences of TLR4, MD-2 and CD14 from the ancestors of tetrapods (ancTetrapod), bony vertebrates (ancBonyVert), and teleost fish (ancTeleost) based on the protein sequences from hundreds of modern species. José Sánchez-Borbón, a fellow graduate student in the Harms lab, has made significant progress in identifying candidate teleost CD14 proteins and was the one to do the CD14 ancestral sequence reconstruction with these protein sequences. I will not elaborate on his findings here, but we plan to co-author a manuscript with our reconstructed ancestral data.

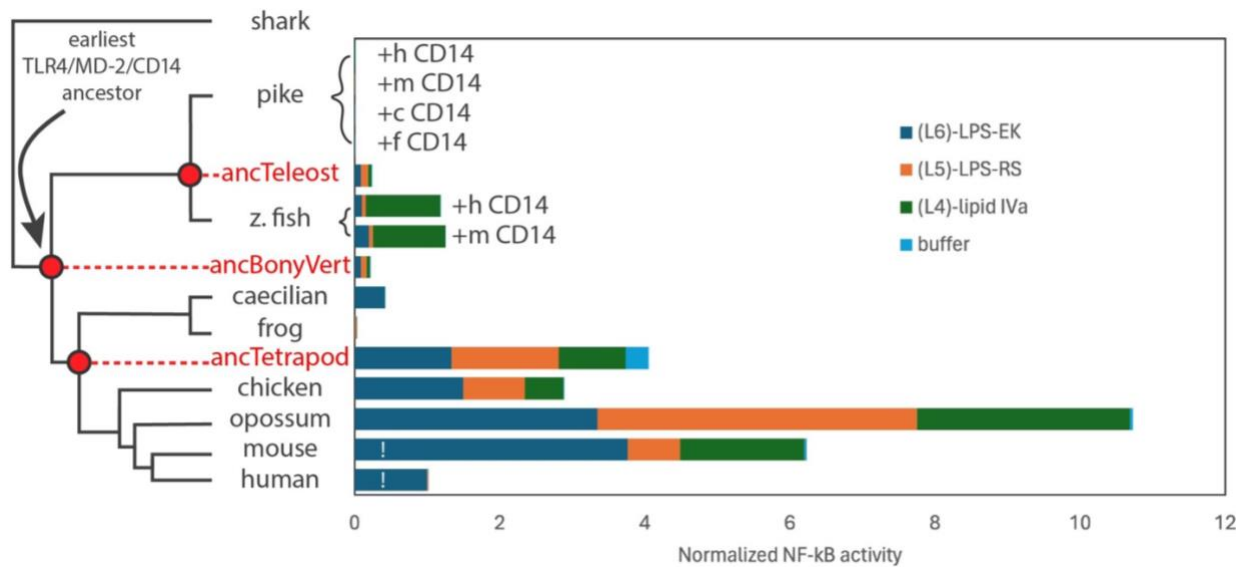


Figure 3.6. Characterization of TLR4 activity for reconstructed early vertebrate ancestors and modern fish and amphibian sequences. The phylogenetic tree on the left demonstrates evolutionary relationships between modern species (black) and ancestors (red) with their TLR4 complex activities compared in the graph to the right. NF- κ B activity measurements of the response to LPS-EK (dark blue), LPS-RS (orange), and lipid IVa (green) are shown stacked for each complex. Because pike and zebrafish are not known to have CD14, we included CD14 from human (h), mouse (m), chicken (c), or frog (f) to assess TLR4/MD-2 function. NF- κ B activities from the teleost fish and the bony vertebrate ancestor complexes are shown relative to zebrafish TLR4/MD-2 with mouse CD14 in the presence of lipid IVa. Activities from all tetrapod complexes are shown normalized to the human TLR4 complex treated with LPS-EK. Human and mouse complexes were treated with 0.1 ng/uL LPS-EK (!). All other treatments were done with 1 ng/uL LPS variant. There was no data collected for caecilian TLR4 challenged with LPS-RS.

We provide a brief overview for our selection of modern species: Zebrafish are a species of teleost within the Otocephala clade. We selected the northern pike (*Esox lucius*) to represent the sister teleost clade, Euteleostei. Teleost fish make up most of the modern ray-finned fish (Actinopterygii). The sister lineage of ray-finned fish is lobe-finned fish (Sarcopterygii), and together they form the clade of bony vertebrates (Euteleostomi). The lobe-finned fish descendants include Tetrapoda, which is comprised of Amphibia and Amniota. We selected two amphibians, the African clawed frog (*Xenopus laevis*) and the two-lined caecilian (*Rhinatrema bivittatum*), to represent two distinct clades of Amphibia: the Salientia and the Gymnophiona,

respectively. We directly compare our findings to the human, mouse, opossum, and chicken complexes as representative amniotes in the functional assay, which have been well studied elsewhere.¹⁹² Sequences for these species were used in the TLR4 and MD-2 ancestral sequence reconstructions. For zebrafish and pike, which do not have CD14, we supplemented these complexes with a CD14 from a different organism in the functional assay.

The functional assay results indicate that low-level LPS sensitivity was present in the bony vertebrate and teleost ancestors and high sensitivity to LPS evolved before the tetrapod ancestor (**Fig 3.6**; red labels). All ancestral complexes show slight specificity for hexa- and penta-acylated LPS over tetra-acylated LPS. The pike TLR4/MD-2 complex did not show activity under any condition. This does not match what we found for zebrafish or ancTeleost, suggesting that the lineage to pike TLR4/MD-2 lost LPS recognition. For the amphibians, the caecilian TLR4/MD-2/CD14 complex responds moderately to (L6)-LPS-EK and not at all to (L4)-lipid IVa. We have not yet tested its response to (L5)-LPS-RS. The frog TLR4/MD-2/CD14 complex does not show activity in the presence of any ligand tested. These data suggest that amphibians have also experienced lineage-specific loss of function in TLR4 complexes.

We wondered if complexes with low signal could alternatively be explained by a problem with our heterologous expression system. In addition to low NF- κ B signal, we noticed that the HEK cells expressing these complexes often had low overall expression as measured by a reporter of constitutive protein expression. We considered that the structural interface of the TIR domain and immune signaling adaptors of early branching vertebrates may have diverged significantly from the human proteins expressed in HEK cells, and this mismatch could result in low immune stimulation as well as altered levels of protein expression. We tested whether creating hybrid TLR4s with human intracellular domains would allow us to better interrogate the

ligand sensitivity and specificity of other species' extracellular domains. We designed a vector with the human TLR4 signaling peptide, transmembrane domain, and TIR domain in which we could scarlessly insert the extracellular domain of any TLR4 between the signaling peptide and transmembrane domain.

This strategy also offered itself to the investigation of the TLR4 and MD-2 paralogs, CD180 and MD-1. CD180 (also known as RP105) does not have a TIR domain but forms a complex with MD-1 that is able to bind lipid A and induce a distinct TLR heterodimer complex.¹⁹³ CD180/MD-1 is thought to regulate LPS-induced TLR4 signaling.^{163,194,195} Exploring the LPS specificity of CD180/MD-1 by providing a measurable output for ligand recognition would give us insight into the early evolution of TLR4/MD-2.

We generated plasmids with the extracellular domains (ectodomains) of two zebrafish TLR4 ohnologs, *tlr4ba* and *tlr4al*, and zebrafish CD180 attached to the transmembrane and TIR domains of human TLR4. We transfected these hybrid plasmids with their corresponding coreceptors, MD-2 for TLR4s and MD-1 for CD180, as well as mouse CD14. We measured the relative NF- κ B activity induced by hybrid zebrafish receptors in response to LPS-EK and lipid IVa (**Fig 3.7A**). Our results show similar LPS specificity of zebrafish *tlr4ba* in the context of zebrafish and human transmembrane and TIR domains, perhaps with slightly decreased sensitivity to LPS-EK when in the human context. Zebrafish *tlr4al* and CD180 hybrid proteins show minimal, if any, activation above background.

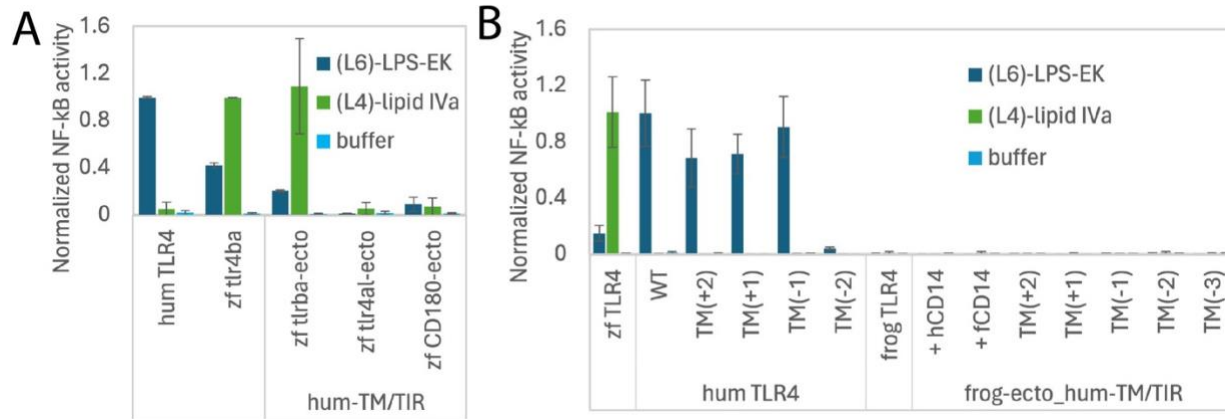


Figure 3.7. Hybridizing the human TLR4 transmembrane and TIR domains to other species extracellular domains does not reveal function. NF-κB activity of hybrid TLR4 and CD180 proteins with the human transmembrane and TIR domains (hum-TM/TIR) in response to LPS-EK and lipid IVa. **A)** On the right, the ectodomains of zebrafish TLR4 ohnologs *tlr4ba* and *tlr4al* (zf *tlr4ba*-ecto and zf *tlr4al*-ecto, respectively) as well as the TLR4 paralog, CD180, from zebrafish (zf CD180-ecto) were expressed with the human transmembrane and TIR domain. **B)** The transmembrane helical register was shifted for human TLR4 and the frog hybrid proteins by adding or removing amino acids from the interface of the transmembrane and ectodomain. These are displayed from left to right as the longest to shortest helix (TM(+/- # amino acids)). The frog helical register mutants were expressed with frog CD14. Wild-type zebrafish, human, and frog TLR4s are included as controls. Error bars indicate standard deviation.

We also created a hybrid frog TLR4 ectodomain/human transmembrane and TIR domain protein and found it did not show activity in the presence of frog or human CD14 (**Fig 3.7B**). To test if our hybrid proteins were nonfunctional because of an improper relative orientation of the ectodomain to TIR domain, Corinthia Brown made mutants of both human and hybrid frog TLR4s to shift the helical register of the transmembrane domains. There are 3.6 amino acid residues per turn of an alpha helix with each additional residue conferring a 100° rotation around the helical axis. We constructed mutants to sample each step of at least one full rotation of the helical transmembrane domain, allowing all possible relative orientations of the extracellular to intracellular domains. When we tested these proteins in the functional assay, we found that human TLR4 can accommodate at least up to a 200° rotation when adding amino acids at the interface between transmembrane and extracellular domain (**Fig 3.7B**). It can only tolerate a 100°

rotation into the membrane, as seen by a loss of function when removing 2 amino acids. Overall, the human TLR4 helical register mutant data indicate that human TLR4 can assume any relative orientation of the extracellular and intracellular domains without breaking function. No mutant of the frog TLR4 hybrid was able to activate activity (**Fig 3.7B**). We conclude that the frog TLR4/MD-2 complex is unable to mount an immune response to LPS.

Overall, these hybrid TLR4 experiments indicated that heterologous expression of TLR4 complexes in HEK cells does not drastically affect NF- κ B signaling. We did note some improvement to the overall level of protein expression in the cells expressing the hybrid forms of TLR4 (data not shown). But this improvement only made us more confident in the measurements of low NF- κ B signal. Therefore, we did not further investigate other low-sensitivity complexes under these conditions.

Initial investigations into possible CD14 functional homologs in fish

We worked to identify a molecule that could serve an LPS-transport role like CD14. CD14 is known to catalyze the delivery of LPS to the TLR4/MD-2 complex in mammals. It retrieves monomeric LPS from the extracellular space, usually with the help of the LPS-binding protein (LBP) that extracts LPS from the outer membranes of Gram-negative bacteria, and then transfers LPS to the TLR4/MD-2 complex. At the time of these experiments, CD14 was believed not to be present in fish; phylogenetic studies suggested CD14 evolved in tetrapods before the divergence of amphibians and amniotes. In this section of the chapter, I will present several preliminary investigations into possible LPS chaperone candidates in fish that I took point on. These experiments paved the way for current research being conducted by José Sánchez-Borbón, a fellow graduate student in the Harms lab. José and I are preparing a manuscript that consists of

findings previously reported in this chapter as well as José's discoveries from digging further into the evolutionary history of fish CD14.

The following preliminary experiments were done while still optimizing our *in vitro* assay for studying the ligand specificity of zebrafish TLR4. Unfortunately, this has made direct comparisons between experiments difficult, but I will highlight important experimental modifications. Critical details that were altered include: 1) *The amount of TLR4, MD-2, and CD14 plasmids transfected into cells.* We noticed this contributed significantly to signal strength. The transfection amount will always be presented as TLR4:MD-2:CD14 with "1" equal to 1 ng of plasmid transfected per well. Transfections with human TLR4 were consistently at 10:0.5:1, as established previously.¹⁰² 2) *The concentration of LPS used to treat cells.* Most often, experiments include a control with human TLR4 treated with 0.2 or 0.1 ng/ μ L (L6)-LPS-EK. There are a few instances that this did not occur. All other treatments were applied with 2 or 1 ng/ μ L LPS. And 3) *the CD14 used to catalyze the zebrafish TLR4 response to LPS.* As I will describe below, a CD14 molecule has not been found in zebrafish but including CD14 drastically improves *in vitro* zebrafish TLR4 signaling.¹⁰¹

Our previous experiments used human or mouse CD14 as these were reported to allow zebrafish *tlr4ba* signaling *in vitro*.¹⁰¹ Now that we had a potent agonist of zebrafish *tlr4ba*, we wondered if we could see activation in the absence of CD14. Our preliminary investigations showed no signal from zebrafish *tlr4ba*/MD-2 in the absence of CD14 (**Fig 3.8A**). This was probably due to using low and variable concentrations of lipid IVa around 0.2 ng/ μ L. Signal amplitude and reproducibility were greatly improved in later experiments by transfecting cells with a plasmid ratio of 10:20:1 (mouse CD14) and treating with 1 ng/ μ L lipid IVa that was ultrasonicated to disrupt micelles prior to the experiment (**Fig 3.8B**). José Sánchez-Borbón first

noticed that zebrafish *tlr4ba*/MD-2 exerts low-level NF- κ B activation independent of CD14 and is currently following up on it.

We considered the possibility that the chaperone role of CD14 could be achieved through some other zebrafish protein. CD14 is a gene duplicate of vertebrate TLR2, which is present in fish. Like TLR4, TLR2 is an innate immune pattern recognition receptor. In mammals, TLR2 binds a wide variety of bacterially associated ligands including acylated lipopeptides and peptidoglycan but does not initiate the inflammatory response to LPS.¹⁹⁶ It can also heterodimerize with other TLRs to initiate immune responses. We hypothesized that CD14's ability to chaperone LPS could stem from an ancestral trait of the TLR2-CD14 ancestor, and perhaps the fish lineage maintained this function in their TLR2. We did not find this to be the case (**Fig 3.8A**). We transfected HEK293T cells with increasing doses of human CD14 or zebrafish TLR2 plasmid and found that all concentrations of CD14 promoted *tlr4ba* activation, but no amount of zebrafish TLR2 plasmid induced LPS-sensitivity. We did not test whether TLR2 was being expressed on the surface of cells, therefore we have not proven here that zebrafish TLR2 lacks an LPS-chaperoning ability, but it seems unlikely given our data.

We next tested whether CD180 with or without MD-1 could act as an LPS chaperone. The CD180/MD-1 complex is able to bind lipid A and is implicated in LPS-induced TLR4 signaling.^{163,194,195} We transfected cells with combinations of human or zebrafish TLR4, MD-2, CD180, MD-1 and CD14 in the presence of LPS-EK and lipid IVa (**Fig 3.8C**). Our data indicate a possible role for human CD180/MD-1 to activate TLR4 in the absence of MD-2, but zebrafish TLR4 is not activated in any condition without both MD-2 and CD14.

Figure 3.8 (next page). Zebrafish TLR2, CD180/MD-1, and human transferrin do not rescue the zebrafish TLR4 response to lipid IVa in the absence of CD14.

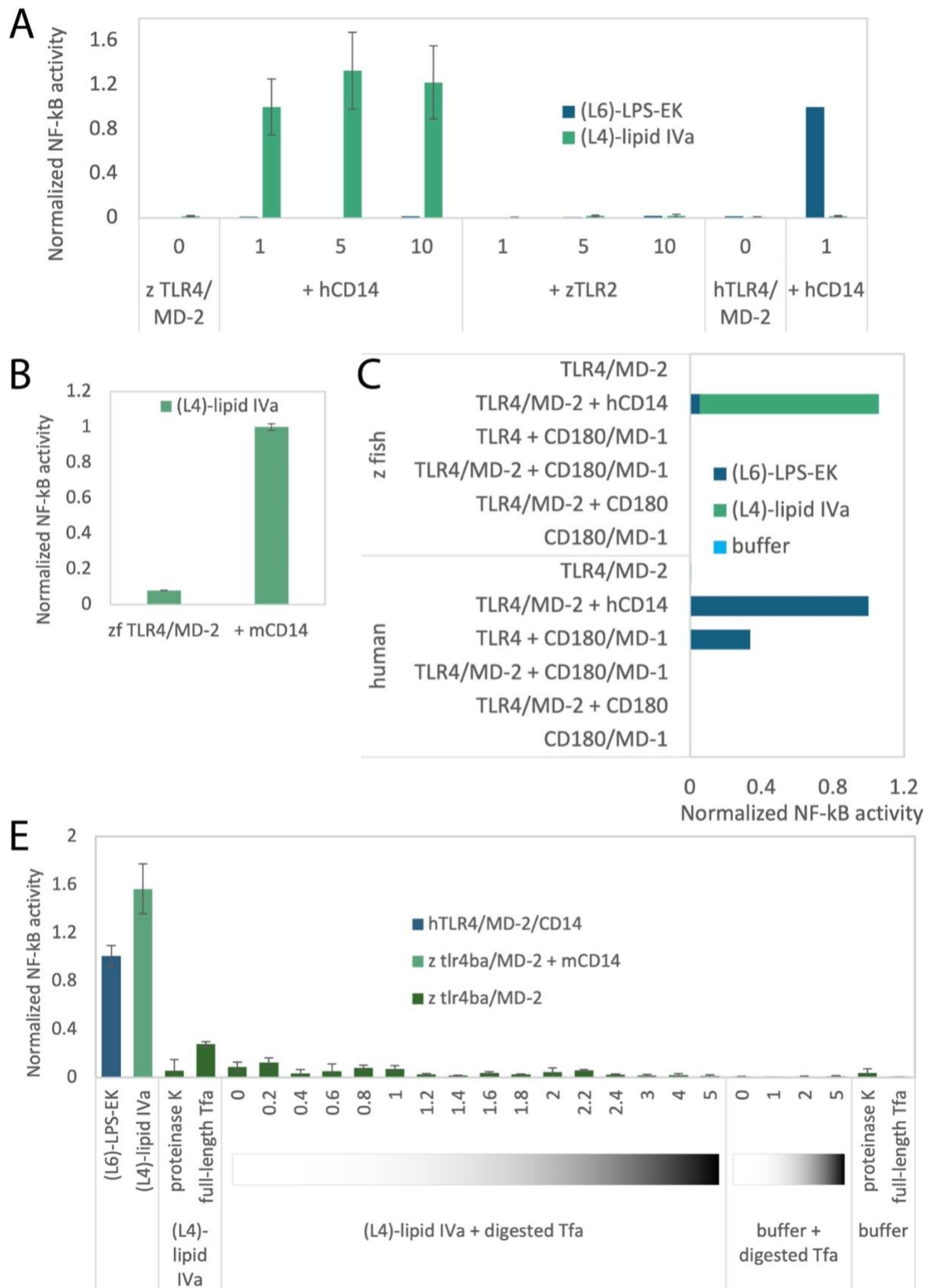


Figure 3.8. Zebrafish TLR2, CD180/MD-1, and human transferrin do not rescue the zebrafish TLR4 response to lipid IVa in the absence of CD14. **A)** NF- κ B activation of zebrafish TLR4 in response to lipid IVa or LPS-EK in the presence of candidate LPS chaperone molecules. We tested zebrafish TLR2's ability to catalyze LPS delivery to zebrafish TLR4. Zebrafish TLR4/MD-2 was transfected at 25:20 ng plasmid/well with human CD14 or zebrafish TLR2 at plasmid concentrations of 0, 1, 5, or 10 ng per well. Cells were challenged with 0.2 ng/ μ L lipid IVa or LPS-EK. For comparison, human TLR4 with and without CD14 is shown on the far right. Human TLR4 was treated with 0.002 ng/ μ L LPS variants. **B)** Zebrafish TLR4/MD-2 challenged with high dose (1 ng/ μ L) lipid IVa in the presence or absence of CD14. In addition to increasing concentration, lipid IVa was ultrasonicated before treatment and cells were transfected with 10:20 TLR4:MD-2 with or without 1 ng mouse CD14 per well. **C)** Testing human and zebrafish CD180/MD-1 for LPS chaperone abilities. Zebrafish complex transfected at 25:20:1, zebrafish CD180:MD-1 transfected at 15:15, and human CD180:MD-1 at 10:10. Human proteins treated were with 0.2 ng/ μ L LPS variants, whereas zebrafish proteins were treated with 2 ng/ μ L LPS. For experiments in panels **A-C**, NF- κ B activity level was buffer-subtracted and is shown normalized by species: human TLR4 signal is relative to human TLR4 treated with LPS-EK; zebrafish TLR4 signal is relative to its signal in the presence of lipid IVa. **D)** Testing if transferrin can act as an LPS chaperone for zebrafish TLR4. We treated zebrafish tlr4ba/MD-2 with 1 ng/ μ L lipid IVa in the presence of full length and proteinase K-digested transferrin peptides (green). Digestion products are displayed in increasing concentration from left to right (gradient). A value of "1" is equal to 71.5 nM of transferrin starting material that was subsequently proteolyzed. For comparison, human TLR4/MD-2/CD14 and zebrafish TLR4/MD-2 with mouse CD14 were treated with LPS-EK and lipid IVa, respectively. For this experiment, the plasmid ratio for zebrafish tlr4ba:MD-2:mouse CD14 was 10:20:1 ng/well. NF- κ B activity level was buffer-subtracted and all data normalized to human TLR4 signal in the presence of 0.1 ng/ μ L LPS-EK.

The final candidate we probed for an LPS delivery role was transferrin. Transferrin is an iron-binding molecule that can bind the lipid A moiety of LPS and is implicated in sequestering iron from microbes. Macrophages from goldfish, teleost fish closely related to zebrafish, express transferrin and transferrin proteases in response to the presence of microbes.¹⁹⁷ Cleavage products of transferrin are shown to enable the activation of goldfish macrophage immune responses in the presence of LPS.^{198,199} One of these cleavage products shares sequence similarity with the LPS-binding site of mouse CD14. In addition, transferrin, likely when bound to the transferrin receptor, colocalizes with TLR4 during LPS-induced TLR4 endocytosis which

is essential for robust immune activation by TLR4.²⁰⁰ The culmination of this evidence suggested that transferrin could serve at least an incidental LPS-transport role that CD14 has become specialized in. We assessed this possibility using full-length or proteinase K-digested human transferrin and tested for CD14-like functions *in vitro*.

Our results show that full-length transferrin might have some capacity to help lipid IVA get to TLR4 (**Fig 3.8D**). This can be seen by comparing the lipid IVA-induced activity of zebrafish TLR4 in the presence of full-length transferrin (second green bar) to the control without transferrin (third green bar: “0”). This is roughly a 3-fold increase (0.09 to 0.28 units). Yet CD14 provides ~5.5 times better enhancement of signaling. While working on this, it seemed that transferrin and/or proteinase K would sometimes cause the cells to form liquid-like droplets on the plate surface. This was more extreme at higher concentrations. Although perhaps interesting, we did not pursue this further.

Overall, we have learned from these experiments that zebrafish TLR2 and transferrin probably do not serve a homologous role to CD14 in zebrafish. However, this work established a basis for José’s future functional characterizations of other candidate fish CD14-like proteins.

DISCUSSION

Extrapolating findings from studies in model organisms to human biology hinges on our ability to define homology between species. We have shown here that the zebrafish innate immune response to LPS is different than in humans. We are optimistic that the root of these differences can be revealed by the functional characterization of homologous proteins *in vitro* and *in vivo*. We have shown that ancestral sequence reconstruction is an excellent tool that can be applied to the interrogation of evolutionary trajectories between zebrafish and human proteins.

We are left with several future directions to explore the zebrafish immune response to LPS, the role of zebrafish TLR4, and the origin of LPS recognition and specificity of TLR4/MD-2/CD14 complexes.

Zebrafish TLR4 ohnologs might play important physiological roles

Our study began with the observation that one of the three zebrafish TLR4 ohnologs, *tlr4ba*, can induce a robust inflammatory response when challenged with tetra-acylated LPS in a human cell-based assay (Fig 3.2 & 3.3). We observed that it may be possible for heterocomplexes of TLR4 ohnologs to amplify the immune response to LPS (Fig 3.3B). Heterocomplexes may play an important role in ligand sensitivity in the zebrafish. The three ohnologs of TLR4 likely arose from unequal gene gain and loss throughout multiple rounds of whole genome duplications in the lineage leading to zebrafish.²⁰¹⁻²⁰² Gene duplications can lead to functional diversification of genes, including gain or loss of function, or subfunctionalization of the original function between the duplicate proteins.²⁰³ These three genes have been maintained in the zebrafish genome for hundreds of millions of years and are all expressed at least in developing larval fish tissues.¹⁰¹ We think it would be an interesting avenue of investigation to uncover what these proteins do for zebrafish physiology and health.²⁰²

Possible functional roles of the zebrafish MD-2 C-terminal peptide

We initially predicted that the zebrafish C-terminal peptide played a role in determining the number of acyl chains that fit into the MD-2 binding pocket. Our data show that it is more complicated than this, rather, the peptide is specifically necessary for zebrafish TLR4 to mount an immune response to LPS (Fig 3.4D-E). We consider that the peptide might stabilize the

dimerization interface of zebrafish TLR4 in a way that is unnecessary for human TLR4. It has been reported that when (L6)-LPS packs its acyl chains into the human MD-2 binding pocket, the lipid A diglucosamine backbone is displaced upwards by $\sim 5\text{\AA}$ relative to lipid IVa-bound structures. This positions the phosphate groups such that they can interact with the positively charged residues of two dimerizing TLR4s (Park et al., 2009).⁴² Perhaps there is also some unique structural feature of the zebrafish TLR4 dimerization interface that requires the C-terminal peptide to correctly position the diglucosamine backbone and phosphates for productive dimerization. This interaction could be between TLR4 and LPS or between TLR4 and the LPS-bound peptide.

A similar but alternative hypothesis is that the peptide serves as part of the hydrophobic core with LPS acyl chains to form the dimerization interface. Ohto et al. proposed from their crystal structures that for productive TLR4 signaling, amino acids at the opening of the MD-2 beta-cup must interact with at least a partial LPS acyl chain to form the dimerization interface between two TLR4/MD-2/LPS complexes.⁴⁵ For dimerizing mouse TLR4/MD-2/lipid IVa complexes, the MD-2 Phe126 side chain shifts more than 4\AA toward the MD-2 cavity where the Phe126 loop interacts with another MD-2 loop and a partially exposed chain of LPS to form the hydrophobic core of the dimerization interface. This core interacts with the hydrophobic patch on the dimerizing TLR4. Perhaps the zebrafish MD-2 peptide has evolved to serve this role at the dimerization interface. The peptide is positioned at the outer edge of the MD-2 beta-cup opening and has two phenylalanines that could potentially be involved in dimerization like the mouse MD-2 Phe126 loop.

Understanding the functional role of the cyprinid MD-2 C-terminal extension would help clarify how this protein has evolved, and potentially why *in vitro* zebrafish TLR4/MD-2 exhibits

contrasting specificity to the human complex. Our next steps would be to gain more detailed structural information about zebrafish and other cyprinid species TLR4/MD-2s. Did other cyprinid species evolve to recognize LPS, and if so, is their function and their ancestral states also dependent on the C-terminal peptide? We could alternatively make mutations to zebrafish TLR4 at sites that are predicted to interface with the peptide during dimerization. This kind of mutational analysis would benefit from structural simulations of dimerization or ligand binding. In conclusion, there are many ways to continue probing the difference between human and zebrafish TLR4 *in vitro*.

Is TLR4 used in the zebrafish innate immune response to Gram-negative bacteria?

Our *in vivo* results suggest that larval 6 dpf fish do not recognize or respond to TLR4 agonists. It is unclear if standard purification LPS-induced immune responses previously reported are mediated by TLR4, TLR2, or a redundant LPS-sensing pathway. RNAseq data of 5 dpf fish showed there are a small subset of immune cells that upregulate expression of zebrafish TLR4 orthologs and MD-2. It is possible, then, that our 6 dpf studies were too early in fish development to assess functional consequences of TLR4 in immunity if this defense mechanism is still maturing. Or perhaps lipid IVa is not a strong agonist of zebrafish TLR4 *in vivo*. Alternatively, maybe zebrafish TLR4 can bind lipid IVa *in vivo* but there are other regulatory mechanisms to dampen the immune response. Zebrafish have been shown to express intestinal alkaline phosphatase which detoxifies LPS in their gut and prevents intestinal inflammation in response to the gut microbiota.¹⁸⁶ We considered that our microgavage experiment may have been affected by the dephosphorylation of LPS in the fish intestine with an intact microbiota.

The inflammasome might serve as an alternate LPS-sensing pathway in zebrafish. Inflammasomes are cytosolic multiprotein complexes that regulate the immune response to intracellular danger signals much like TLRs at the extracellular surface. Caspase-11-deficient mice are resistant to LPS-induced sepsis, implying that caspase-11 participates in host response to LPS.²⁰⁴ Caspase-4 and caspase-5 in humans and the ortholog in mice, caspase-11, are activated by direct binding to intracellular LPS.²⁰⁵ Much like human TLR4/MD-2 activation, underacylated lipid IVa and (L5)-LPS-RS were shown to bind to caspase-4/11 but could not induce oligomerization and activation of the inflammasome.^{204–206} This LPS-binding was mediated by the CARD domain of the caspases.²⁰⁵ Zebrafish have homologs but not direct orthologs of mammalian caspases. One of these homologs, *caspy2* (or *Casb*), when overexpressed in HEK293T cells can directly bind LPS via the N-terminal pyrin death domain, resulting in *caspy2* oligomerization which is necessary for pyroptosis.²⁰⁷ Knockdown of *caspy2* expression protects larvae from lethal sepsis.²⁰⁷ However, LPS needs to be delivered to the cytosol in order for *caspy2* to bind to it. This happens naturally during an infection due to bacterial effector proteins that can inject LPS into host cells. It is possible that this LPS-sensing role of the inflammasome protein *caspy2* has taken over the responsibility of Gram-negative bacterial detection in zebrafish. It would be interesting to investigate whether TLR4 plays a role in priming zebrafish immune cells to upregulate the expression of *caspy2* for LPS recognition, or in regulating the delivery of LPS to the intracellular space.

Further probing ancestral complexes will help us understand TLR4 ligand responses

The reconstructed bony vertebrate and teleost ancestors showed low-level LPS sensitivity. This makes sense given the lack of signal we see for the pike, frog, and zebrafish

tlr4bb and tlr4al complexes, as well as the low sensitivity of zebrafish tlr4ba to LPS variants with more than four acyl chains and the caecilian TLR4 to hexa-acylated LPS. However, a lack of signal is harder to make a case for functional similarity than a positive signal.

Ancestral sequence reconstruction relies on several assumptions of the evolution between sequences in a dataset. We cannot access every sequence change along an evolutionary trajectory and therefore, we rely on complex models to infer the most likely series of mutational events connecting modern sequences. Because of this, it is more than likely that the sequences reconstructed do not reflect the exact ancestral states of the protein. Posterior probability is a useful statistical measurement used to indicate our uncertainty in the amino acid call at each position along a reconstructed protein. If we were completely confident in our reconstruction, meaning there was only one residue at each position with high probability, the average posterior probability across the entire sequence would be 1.0. Our reconstructed sequences for the tetrapod ancestor (ancTetrapod) TLR4, MD-2, and CD14 have average posterior probabilities of 0.881, 0.854, and 0.955, respectively. The average posterior probability for the ancestor of bony vertebrates (ancBonyVert) complex was 0.806, 0.798, and 0.929, respectively. And the teleost ancestor (ancTeleost) yielded average posterior probabilities of 0.885, 0.774, and 0.926 for TLR4, MD-2 and CD14, respectively. Typically, ancestral proteins have been shown to exhibit function when they have posterior probabilities > 0.85 .

It is possible that the ancestral complexes show low activity due to borderline poor-quality reconstructions of TLR4 and MD-2 sequences. However, we argue that there are far more ways to break a protein's function than there are to maintain it. A good next step would be to resurrect and characterize the altAll ancestors—ancestral sequences with every ambiguous amino acid substituted with the next most probable alternate. These proteins would serve as the “worst

case” scenario if the reconstruction chose the wrong amino acid at every ambiguous site. If the maximum likelihood and altAll ancestors have the same function, then the function is robust to uncertainty in the reconstruction and likely reflects the protein’s ancestral state.^{76,115,134–138}

We would also be interested in doing further investigation into the biochemical changes that have occurred throughout these evolutionary intermediates that confer different LPS sensitivity and specificity. We would better understand the mechanism of TLR4/MD-2/CD14 ligand responses if we were to compare these ancestors at a deeper level.

Did zebrafish lose CD14 as a mechanism to avoid LPS toxicity?

A major hindrance to our investigations of zebrafish TLR4 function has been the elusive CD14-like protein. We have investigated several candidate molecules (TLR2, CD180/MD-1, and transferrin) that exist in zebrafish, can bind LPS, and have been shown to be involved in immune responses. However, none of these proteins could catalyze TLR4/MD-2 ligand-induced activity like CD14 in our functional assays. It is possible that zebrafish do not have a CD14-like protein and that could explain their low sensitivity to LPS even if zebrafish TLR4/MD-2 can mount an immune response. Currently, José Sánchez-Borbón is taking point on investigating the evolutionary history of CD14 and has begun to reveal when it became a functional part of the TLR4 complex. José has identified that most fish species either have TLR4/MD-2 or a proto-CD14 molecule, but not all three proteins. If this is true, then perhaps fish lost the ability to sense LPS via TLR4 by selective pressure to evade sepsis-like diseases, or alternatively they have not needed the full complex to deal with infections by Gram-negative bacteria.

In conclusion, we have leveraged ancestral sequence reconstruction, cell-based and organismal functional assays, and mutational analyses to probe the evolutionary divergence

between human and zebrafish TLR4 complex structure, function, and role in the innate immune response to LPS. We find that although human and zebrafish TLR4/MD-2 share many properties, like the ability to bind LPS and stimulate an inflammatory response, there are many dissimilarities after 430 million years of evolutionary separation in the way these homologous proteins are activated by their ligand and how they are involved in the immune response. Much more work remains to really understand the role of zebrafish TLR4, if the zebrafish can be used to model TLR4-induced inflammation, and why the *tlr4ba* ortholog has evolved the unique ability to specifically recognize tetra-acylated LPS.

MATERIALS AND METHODS

Ancestral sequence reconstruction

We reconstructed ancestral sequences using the *topiary* pipeline available on GitHub (<https://github.com/harmslab/topiary>).¹⁸² The multiple sequence alignments generated by the first stage of the *topiary* script were manually edited to remove ambiguous sequences and gene duplicates using AliView software.¹⁵² Sequences from key species were added to the alignment to increase taxonomic sampling when it was lacking. At this point, the signal peptide for every sequence in the alignment was predicted by SignalP – 6.0²⁰⁸ and then removed from the alignment before being fed back in to stage 2 of *topiary* to perform the ancestral inference. TLR4 and MD-2 sequence alignments, reconstructed ancestral sequences, and phylogenetic trees built during ASR are available in the supplement files. The Jones-Taylor-Thornton (JTT) substitution model was used in the MD-2/MD-1 ancestral inference. The maximum likelihood species-reconciled gene tree for MD-2/MD-1 aligned fairly well with the species tree except that it placed two duplication events within the early branches of the MD-1 clade and placed several

amphibians outside of fish. This was probably due to high variability in amphibian MD-1 sequences in the alignment. The MD-2 clade, however, aligns well with the species tree. The MD-2 ancestors we wanted to characterize ancBonyVert (anc88), ancTeleost (anc87), and ancTetrapod (anc76) had average sequence posterior probabilities of 0.798, 0.774, and 0.854, respectively. Bootstrap sampling of the gene-species tree resulted in high branch supports for these nodes. The TLR4/CD180 ancestral inference used the JTT general amino acid exchange rate matrix and discrete Gamma model with 8 rate categories (JTT+G8). The reconciled gene-species tree aligns well with the species tree. There were also duplication events labeled early on in both TLR4 and CD180 clades suggesting fish TLR4 and CD180 have significantly diverged from their tetrapod homologs. The TLR4 ancBonyVert (anc345), ancTeleost (anc57), and ancTetrapod (anc334) reconstructions had posterior probabilities of 0.806, 0.885, and 0.881, respectively. Bootstrap analysis for the TLR4/CD180 tree is not complete, so we do not yet know the branch supports at ancestral nodes. Both maximum likelihood and altAll sequences were reconstructed for every ancestor. We used the maximum likelihood ancestral sequences for functional characterization.

Plasmids

Ancestral gene sequences were human codon optimized and synthesized by GeneWiz (Azenta) in a pcDNA3.1(+) backbone without the T7 promoter. TLR4, MD-2, and CD14 ancestral sequences without native signal peptides were flanked upstream by the CMV enhancer and promoter, a Kozak sequence, the human signal peptide, and a FLAG tag, and then flanked downstream by two stop codons.

Most mammalian expression plasmids were already in house. Pike TLR4 and MD-2 genes were human codon optimized and then synthesized by GenScript in the pcDNA3.1(+) vector with a Kozak sequence upstream of the start codon. Zebrafish *tlr4a1* (Accession No. :NM_001328605.1), *cd180* (Accession No.: NM_001310490.1), *ly86 (md-1)* (Accession No. : NM_001310488.1), and human CD180 (Accession No.: NM_005582.3) and LY86 (MD-1) (Accession No.: NM_004271.4) were cloned by GenScript into the pcDNA3.1(+) expression vector with a Kozak sequence upstream. In house cloning was done using SLIC and KLD mutagenesis kits.

Cell culture and transfection conditions

We followed well-established protocols for transient transfection using the Dual-Glo Luciferase Assay System (Promega).¹⁰² Human embryonic kidney cells (HEK293T /17, American Type Culture Collection CRL-11268) were maintained up to 30 passages in DMEM supplemented with 10% FBS at 37°C with 5% CO₂. For each transfection, a confluent 100 mm plate of cells was treated at room temperature with 0.25% Trypsin-EDTA in HBSS and resuspended with an addition of DMEM + 10% FBS. This was diluted four-fold into fresh medium and 135 µL aliquots of resuspended cells were transferred to a 96-well cell culture treated plate. All transfection mixes were made with 1 ng of Renilla and 20 ng of ELAM-Luc. Transfection mixes for human TLR4 complexes were made with an additional 10 ng of TLR4, 0.5 ng of MD-2, 1 ng of CD14, and 67.5 ng of pcDNA3 per well for a total of 100 ng of DNA. Transfection mixes for other species TLR4 complexes, unless otherwise noted, were made with 10 ng of TLR4, 20 ng of MD-2, 1 ng of CD14, and 48 ng of pcDNA3 per well for a total of 100 ng of DNA. All plasmids contained human codon optimized genes and were in mammalian

expression vectors. Transfection mixes were diluted in OptiMEM to a volume of 10 μL /well. To the DNA mix, 0.5 μL per well of PLUS reagent was added and thoroughly mixed followed by a 10 min incubation at room temperature. Lipofectamine was diluted 0.5 μL into 9.5 μL OptiMEM per well. This was added to the DNA + PLUS mix, mixed well, and incubated at room temperature for 15 min. The transfection mix was diluted to 65 μL /well in OptiMEM and aliquoted onto the cells in the plate. Cells were incubated with transfection mix overnight (20–24 h). For cells that received a pre-treatment with the TLR4 inhibitor TAK-242 (synonyms: Resatorvid; CLI-095) (HY-11109, MedChemExpress), 2 μL of 100 μM TAK-242 in cell culture grade DMSO was applied per well and the plate was incubated for 5 min at 37°C before media was removed and cells were treated as normal with 100 μL of LPS mixtures prepared in 25% PBS, 75% DMEM. *E. coli* K-12 LPS (tlrl-eklps, Invivogen) and *R. sphaeroides* LPS (tlrl-rslps, Invivogen) were dissolved at 5 mg/mL in endotoxin free water, aliquots were stored at -20°C . *S. enterica* serotype typhimurium LPS (L6511, Sigma-Aldrich) was dissolved at 5 mg/mL in endotoxin free water and stored at 4°C . Lipid IVa (CLP-24006-S, Biosynth) was dissolved at 0.1 mg/mL in endotoxin free water, aliquots were stored at -20°C . To avoid freeze-thaw cycles, working stocks of LPS were prepared at 10 μg /mL and stored at 4°C . To disrupt micelle formation and evenly distribute LPS in solution, LPS stocks were placed in a room temperature jewelry ultrasonicator for 15 min prior to use in treatments. Cells were incubated with treatments for 4 hr. The Dual-Glo Luciferase Assay System (Promega) was used to assay Firefly and *Renilla* luciferase activity of individual wells. Each NF- κB induction value shown represents the buffer-subtracted Firefly luciferase activity/vehicle blanked *Renilla* luciferase activity, normalized to LPS-treated transfection controls for each species in order to normalize between plates. For cells treated with transferrin proteolysis products, 100 μL of 100 μM human transferrin (T8158,

Sigma-Aldrich) in endotoxin-free water was digested with 30 μ L of 20 μ M proteinase K overnight at room temperature before being mixed at indicated concentrations with treatment mixes.

Oral microgavage of LPS

Zebrafish experiments were approved by the University of Oregon Institutional Animal Care and Use Committee. Larval 6 dpf fish were anesthetized in 168 mg/ml tricaine methane sulfonate in embryo medium (EM) and microgavaged with 4.6 nL of 1 mg/mL LPS purified from *E. coli* 0111:B4 (L2630; Sigma-Aldrich) dissolved in EM. After gavage, fish were transferred to fresh EM. Fish were imaged 6 h post-gavage using a fluorescence stereo microscope. Fish were anesthetized in 168 mg/ml tricaine methane sulfonate in EM before being mounted on a glass slide and images were taken over the distal gut region in bright field, GFP, and mCherry channels. The fish used in this experiment were *tg(tnfa:GFP; mpeg:mCherry)*. After imaging, fish were immediately sacrificed and fixed in paraformaldehyde for subsequent staining and quantification of neutrophils associated with gut tissue. Fixed zebrafish were dissected to isolate the gut bulb and intestinal tract. Stained neutrophils were counted by two treatment-blind individuals.

Brain tectum microinjection of LPS

We used a modified version of a previously published protocol.²⁰⁹ Larval 5 dpf fish were submersed in either 8.3 μ M TAK-242 (synonyms: Resatorvid; CLI-095) (HY-11109, MedChemExpress) or an equal volume of cell culture grade DMSO (final concentration of 0.003%) in EM for 24 hours. 6 dpf fish were anesthetized in 168 mg/ml tricaine methane

sulfonate in embryo medium (EM) and microinjected by brain tectum injection with ~4.2 nL of 2.17 ng/nL *E. coli* LPS-EK (tlrl-eklps, InvivoGen), lipid IVa (CLP-24006-S, Biosynth), or ultrapure PBS. After injection, fish were rinsed and recovered in fresh EM until imaging. Fish were imaged 3-6 hours post-injection by fluorescence light sheet microscopy. For imaging, fish were anesthetized in 168 mg/mL tricaine methane sulfonate in EM before being mixed with 0.7% low-melt agarose at 40°C in EM and mounted into a capillary tube. We collected 5 sets of fluorescent z-stack images over the liver region through the full width of each fish to track immune cells near the liver. The fish strain used in this experiment was *tg(tnfa:GFP; mpx:mCherry)*. Images were analyzed using Imaris software.

BRIDGE TO CHAPTER IV

There have been many discoveries of host-microbe interactions, drug developments, and disease pathologies using the zebrafish model of innate immunity. In Chapter III, we showed ample evidence of dissimilarities between homologous innate immune proteins in humans and zebrafish. In Chapter IV, we present an investigation of zebrafish proteins that are homologous, but not orthologous to the mammalian innate immune protein, calprotectin. This work was stimulated by a recent publication describing the zebrafish s100a10b protein as the functional homolog of human calprotectin in the zebrafish innate immune response to bacterial infection.

CHAPTER IV
ZEBRAFISH DO NOT HAVE CALPROTECTIN

*This chapter contains unpublished co-authored material.

Author contributions: Orlandi KN and Harms MJ designed the study. Orlandi KN designed and performed experiments, analyzed data, and wrote the manuscript. Harms MJ obtained funding and oversaw the project and writing.

ABSTRACT

The protein heterodimer calprotectin and its component proteins play important antibacterial and proinflammatory roles in the mammalian innate immune response. Calprotectin is also a well-validated, non-invasive biomarker of inflammation. Gaining mechanistic insights into the regulation and biological function of calprotectin will help facilitate patient diagnostics and therapy. Recent literature proposed that the zebrafish S100A10b protein is analogous to human calprotectin based on sequence similarity and genomic context. The field would benefit from expanding the breadth of calprotectin studies into a zebrafish innate immunity model. However, thus far there is no phylogenetic nor functional evidence demonstrating the existence of calprotectin in fish. Here, we evaluate the possibility that a zebrafish S100 protein could have convergently evolved a calprotectin-like role in the zebrafish innate immune response. We show the phylogenetic and syntenic relationships of human and zebrafish S100s. We identify and zebrafish S100s that are expressed in immune cells and upregulated during the immune response. We then recombinantly express and purify four candidate proteins and evaluated them for antimicrobial and proinflammatory characteristics. We find that none of the most promising candidates proved to be functionally orthologous to calprotectin nor its component proteins.

INTRODUCTION

Calprotectin plays critical roles in the innate immune response.²¹⁰ This protein is a complex formed by two calcium binding proteins: S100A8 and S100A9. Calprotectin is found in heterodimeric and heterotetrameric states, both of which play biological roles.^{63,64,82,211} Its individual components, S100A8 and S100A9, are also both found as homodimers with biological functions distinct from the heterocomplexes.

S100A8 and S100A9 are highly expressed in the cytoplasm of immune cells²¹² comprising up to 45% of soluble cytosolic protein in neutrophils⁶². Intracellular S100A8 and S100A9 are implicated in the calcium-dependent microtubule reorganization of phagocytes allowing migration to sites of infection.^{65,66,68} Upon release from cells after damage or during an immune response calprotectin exerts antimicrobial activity by sequestering transition metals essential for microbial growth in the extracellular matrix.^{86-88,213-220} Extracellular S100A8 and S100A9 homodimers can amplify the immune response by activating Toll-like receptor 4 (TLR4) and the Receptor for Advanced Glycation End-products (RAGE) promoting cytokine expression and immune cell migration, respectively.⁷² Several other important functions are associated with S100A8 and S100A9.^{69-71,221} Dysregulated expression of these proteins is linked to ailments such as Alzheimer's disease, Parkinson's disease, cerebral ischemia, obesity and cardiovascular disease.²²² Correlated with its role in the immune response, high levels of calprotectin in tissues, serum, or stool are indicative of inflammation associated with severe infections, cystic fibrosis, digestive tract disorders, autoimmune diseases, rheumatoid arthritis and cancer.⁹¹⁻⁹⁴

Given the importance of calprotectin, there is interest in developing new models to study its function. One attractive model is the zebrafish, which is increasingly being used to understand the molecular mechanisms of immune functions. Recent work has begun to characterize the role

of a calprotectin homolog in the zebrafish response to infection.^{107,108} As vertebrates, zebrafish share much of their physiology and molecular components with humans. They also have exceptional experimental advantages: well-established genetic tools, optically transparent larvae (making it possible to visualize tagged molecules in real-time in live fish), and rapid generation times.²²³ They are particularly useful for studying innate immunity because they survive with only the innate immune responses until 4-6 weeks post-fertilization when their adaptive immune system is morphologically and functionally mature.²²⁴⁻²²⁷

Despite the power of the zebrafish model system, it can be challenging to map zebrafish biology to human biology. Millions of years of evolution have allowed the divergence, emergence, and loss of proteins and protein functions between species, often making the comparison difficult. One of the most important considerations is whether the genes being compared between species are, in fact, the same genes. Are they the result of speciation (orthologs) which often have very similar functions, or did they arise by gene duplication (paralogs) which often have very different functions. Establishing gene orthology is particularly challenging for S100 proteins, as they form the largest subgroup within the superfamily of proteins carrying the Ca²⁺-binding EF-hand motif. Humans have 24 S100 genes^{228,229}; zebrafish have 14.^{230,231}

There is no annotated s100a8 or s100a9 in the zebrafish genome; however, there are several zebrafish S100 genes in a similar genomic location to that of human s100a8 and s100a9. One of the zebrafish genes annotated in this genomic location is s100a10b. When performing a BLAST query against the zebrafish proteome, human S100A8 pulls up zebrafish s100a10b as a top similarity hit. On this basis, zebrafish s100a10b has been classified as “calprotectin.” This was followed by experimental studies reporting its transcriptional response to pathogenic

bacteria.^{107,108} Commercially available “fish calprotectin” ELISA antibodies also imply calprotectin is present in fish. However, these antibodies were raised against the highly conserved N-terminal helix of S100A8 and likely bind several S100 proteins. No rigorous investigations have yet been employed to demonstrate the presence of a calprotectin ortholog, or even a convergently-evolved paralog, in zebrafish.

We set out to find phylogenetic, biochemical, or biological evidence of calprotectin (or calprotectin-like) activity in zebrafish s100 proteins. Through a careful review of existing phylogenetic literature, we confirm that fish do not have a calprotectin ortholog: both S100A8 and S100A9 evolved in mammals 250 million years after the divergence of tetrapods and ray-finned fishes. We support this phylogenetic result through a comparative synteny analysis of s100 genes in zebrafish and human genomes. We also investigate the possibility that fish convergently evolved a calprotectin-like s100 protein using single-cell RNAseq data to identify zebrafish s100 proteins expressed in immune cells. We recombinantly expressed and purified four of these proteins—including zebrafish s100a10b, the protein previously identified as fish calprotectin in the literature—and experimentally tested their antimicrobial and pro-inflammatory activities. None of the proteins give measurable activity.

We conclude that zebrafish have neither a vertically inherited ortholog of calprotectin, nor an obvious candidate protein that convergently evolved similar function. Our results highlight the danger of relying on sequence similarity and genomic placement to identify genes. We demonstrate it is necessary and prudent to use an explicitly evolutionary lens with careful functional analyses when mapping results from model organisms to human biology.

RESULTS

Zebrafish *s100a10b* is only distantly related to human S100A8 and S100A9

We started by looking for phylogenetic evidence that fish have a protein orthologous to mammalian S100A8 or S100A9. Orthologous proteins are ones that arose by speciation and are thus the same gene in the species being compared. Paralogous proteins arose by gene duplication and often exhibit gain or loss of function from the ancestral state²⁰³, establishing themselves as new proteins. **Fig 4.1A** summarizes the evolutionary history of S100s. This tree was built referencing several published phylogenetic analyses of the family, including two from our group^{102,232}. The phylogeny at the top shows the current best estimate of the S100 gene tree; the phylogeny on the left shows the evolutionary history of bony vertebrates. Each circle denotes the S100 gene observed in at least one member of the taxonomic groups on the left.

This evolutionary tree indicates that S100A8 and S100A9 evolved by gene duplication from a single gene in the ancestor of amniotes. Reptiles and birds preserve a single calgranulin protein (MRP-126), while mammals expanded it into three proteins (S100A8, S100A9 and *s100a12*). The closest evolutionary relatives of these proteins are *s100a7*, *s100a7a* and *s100a15*. Like the calgranulins, these arose by duplication of a single gene in the ancestor of amniotes. The reptile/bird protein MRP-126 is the earliest diverging protein known to exhibit nutritional immunity and/or Toll-like receptor 4 activation in functional assays.^{102,233} These observations indicate that calprotectin evolved in amniotes ~320 million years ago.

In contrast, zebrafish *s100a10b* (the putative zebrafish calprotectin) falls into a clade with the proteins S100A10 and S100A11. This is one of the earliest S100 protein subfamilies to evolve, with orthologs present in species ranging from tetrapods to jawless fishes. This group of S100 proteins thus diverged from the lineage that led to mammalian S100A8 and S100A9 at

least 563 million years ago, in the last common ancestor of humans and lampreys. Further, after this speciation event, there were at least two more gene duplications on the lineage leading to S100A8 and S100A9. S100a10b is therefore a different gene than S100A8 or S100A9.

Chromosome placement indicates a shared origin but complicated evolution of homologous human and zebrafish S100s

To cross-validate the lack of evidence for vertical inheritance from published phylogenies, we used syntenic analysis to identify zebrafish s100 genes in a similar genomic location to human S100A8 and S100A9. We used ENSEMBL to identify the zebrafish genomic region most similar to human chromosome Chr 1:152-155M, which encodes 19 of the 24 human S100 proteins, including S100A8 and S100A9. This region corresponded to zebrafish chromosome 16. Specifically, human Chr 1:154.6M-154.7M and zebrafish Chr 16:23.5M-23.7M cover the KCNN3 and ADAR genes adjacent to tandem repeats of S100 genes in both species (**Fig 4.1C**).

The existence of this shared cluster indicates that a handful of s100 genes were in this genomic context at least in the bony vertebrate ancestor ~430 million years ago, as established in previous work.²³⁰ The syntenic relationships, however, also give evidence for extensive evolution after the divergence of bony fishes and tetrapods: the orientation and placement of genes are different, and orthologs to the human S100s are missing from this genomic location but present on other chromosomes. Further, most of the zebrafish s100s in this region appear to be teleost-specific duplicates.²³⁰ This includes ictacalcin (icn), icn2, s100t, s100s, and s100w. The only clear orthologs to human proteins are s100a10b and s100a1 (**Fig 4.1A**).

Human calprotectin and zebrafish s100 protein sequences have low sequence identity

Previous workers identified zebrafish s100a10b as calprotectin using human S100A8 as a query in a BLAST search against the zebrafish proteome.¹⁰⁸ Via ENSEMBL, S100A10b is the top hit; however, the e-value for this hit is only 0.045 and the percent identity is 36.17%. Via NCBI, this hit scores an e-value and percent identity of 8E-17 and 36.36%, respectively. We assessed the quality of the hit by reciprocal BLAST, meaning we used the zebrafish s100a10b protein sequence as a query against the human proteome on NCBI. This yielded human S100A1 (9E-34; 55.9%) as the top hit, not S100A8. In fact, S100A8 (5E-16; 36.4%) was the 13th hit (after S100A1, S100A10, S100Z, S100P, S100B, S100A4, S100A12, S100A5, S100A6, S100A2, S100A4, and S100A9). This is consistent with the previous phylogenetic analyses that place S100A8 and S100A9 as relatively distant paralogs to zebrafish s100a10b (**Fig 4.1A**).

To evaluate sequence similarity and identity, we aligned zebrafish s100 protein sequences from the syntenic region to human S100A8, S100A9, and S100A10 sequences (**Figure 4.1C**). As expected, there is high conservation at sites that form the EF-hand and pseudo-EF-hand calcium-binding domains of the s100 proteins. There is low conservation in the region connecting the two EF-hands and at the termini. We used Clustal Omega to determine the sequence identity shared between zebrafish s100s and human S100A8, S100A9, and S100A10 (**Fig 4.1D**). We find similarly low levels of shared identity between human S100A8 and S100A9 and the zebrafish s100s. Overall, there is no obvious candidate zebrafish s100 that is like calprotectin by sequence similarity or identity.

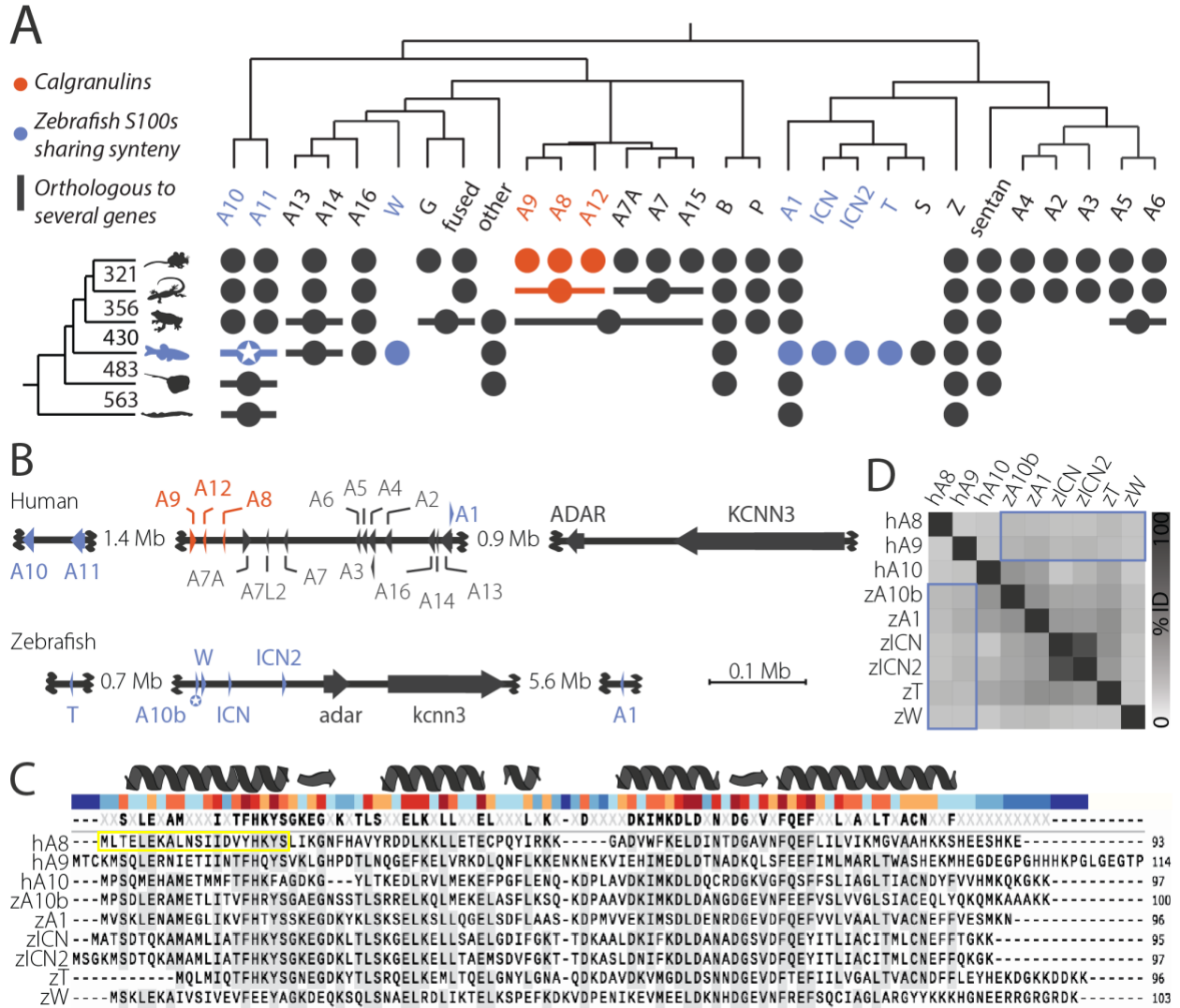


Figure 4.1: Phylogenetic analyses reveal there is no calprotectin ortholog outside of amniotes. **A)** S100 gene tree adapted from Wheeler et al., 2017 shows the evolutionary relationships determined for S100s across vertebrates. The phylogeny on the left shows the relationships between species, with branch point times noted in millions of years ago; the phylogeny on the top shows the estimated S100 gene tree. Circles denote S100 genes from the phylogeny at the top found in at least one member of the taxonomic group from the left. A horizontal line through a circle indicates a single gene that is co-orthologous to multiple S100 genes found in mammals. S100A8, S100A9, and S100A12 form a clade specific to amniotes (orange). Zebrafish genes in the syntenic region shown in panel C are shown in blue. Zebrafish s100a10b, which has been treated as a calprotectin ortholog, is denoted with a white star. **B)** Syntenic regions of human chromosome 1 (top) and zebrafish chromosome 16 (bottom) identified by ENSEMBL. Arrows denote relative gene length and orientation. Human S100A8, S100A9, and S100A12 are shown in orange; zebrafish s100s and their human orthologs are shown in blue. The non-S100 genes *adar*/*ADAR* and *kcnn3*/*KCNN3* are diagnostic for the syntenic region. Not all genes in the region are depicted. **C)** A multiple sequence alignment of the S100 proteins in this dataset displays the amino acid similarity at each position of human

Figure 4.1 (continued) S100A8, S100A9 and S100A10 compared to the zebrafish s100s from a similar genomic context. At the top, secondary structure features of human S100A8 are shown. Under this, the bar is colored by sequence conservation (blue=low, red=high). The consensus sequence from all sequences in the alignment is shown above the individual protein sequences. Amino acids found in at least 50% of the sequences shown are shaded. The antigen for the “Fish Calprotectin” antibody was raised against the peptide boxed in yellow. **D**) Percent identity matrix comparing S100 proteins in this dataset (darker box indicates higher identity). S100 pair identity values range from 25.81-87.37% with a mean value of 39.35% and median value of 35.25%. Identity values comparing human S100A8 and S100A9 to zebrafish S100s are highlighted in the blue boxes.

This alignment also allowed us to ask what zebrafish s100 protein(s) might be recognized by the commercially available “Fish Calprotectin” ELISA Kit from MyBioSource. This kit was made with antibodies raised against a 20 amino acid partial peptide of a human calprotectin-like protein (GenBank: AAB33355.1), which forms the N-terminal helix and beginning of the EF-hand 1 domain of human S100A8 (**Fig 4.1C**, yellow box). An NCBI BLAST search reveals that eight to nine residues in this helix are highly conserved in several zebrafish s100s including s100b, s100a10b, s100a10a, s100a1, s100z, s100s and s100w as well as several other unrelated proteins. These residues are on the inside of the amphipathic helix, involved in stabilizing secondary and tertiary structures, and are potentially used to coordinate metal ions. The N-terminal helix of S100 proteins is at the surface when in dimeric and tetrameric complexes and so the antibody in this kit is likely non-specific.

Single cell RNA sequencing dataset mining points to candidate zebrafish s100 proteins expressed similar to calprotectin

Our bioinformatic analyses revealed that no zebrafish S100 protein is orthologous to human calprotectin; however, it is possible that an S100 protein convergently evolved calprotectin-like activity. To investigate this possibility, we identified zebrafish s100s that share a

similar expression profile to calprotectin. Calprotectin is expressed constitutively in mammalian neutrophils, monocytes, and several epithelial cell types is upregulated upon infection and injury.^{62,212} We queried existing zebrafish single cell RNA sequencing (scRNAseq) datasets for zebrafish s100s expressed in immune cells that are upregulated in response to injury. (We could not find a cell browser for zebrafish infection models.)

We used the UCSC cell browser to visualize a zebrafish development dataset deposited by Farnsworth et al., 2019²³⁴, and assessed constitutive immune cell expression in whole fish 1-, 2-, and 5-days post-fertilization. Of the genes that share a genomic context with s100a9 (**Fig 4.1B**), we found that s100a10b, icn, icn2, s100w and s100a1 are expressed in immune cells of developing zebrafish (**Table 4.1**). s100t, also in this genomic region, shows very low expression in immune cells. We found that five s100 genes from other zebrafish chromosome locations show some expression in immune cells: s100v1, s100v2, s100u, s100z, and s100a11. Finally, the remaining annotated zebrafish s100 genes—s100s, s100b, and s100a10a—appear in very few cells within these clusters.

To evaluate whether these proteins are expressed during the innate immune response to injury, we used the fin clip and tissue regeneration scRNAseq dataset and cell browser provided by Hou et al., 2020.²³⁵ In this dataset, cells were isolated from adult zebrafish caudal fins at 1-, 2-, and 4-days post-amputation. The macrophage marker *mpeg1.1* is highly expressed in hematopoietic cells during the response to fin clip injury and is also expressed in other cell types. The neutrophil marker *mpx* was only detected at very low levels in four basal epithelial cells. We see that s100a10b, icn, icn2, s100w, and s100a1 are expressed in hematopoietic cell clusters, albeit s100a1 to a lesser degree. S100t only appears twice in the hematopoietic cells sequenced but shows more expression in epithelial cells. Zebrafish s100s from other regions of the zebrafish

genome show varying levels of expression in hematopoietic cells. Notably, all s100 genes from outside of zebrafish chromosome 16 were expressed to a lesser degree than s100a10b, icn, icn2 and s100w. The injury model is consistent with the development model, although it is missing data for s100a10a and s100b—perhaps these genes were not detectable in regenerating fin tissues.

Table 4.1: Single-cell RNAseq profiles for zebrafish s100s syntenic to human calprotectin

Gene name	Linkage Group and Synteny	Ortholog Call	Developmental dataset	Injury dataset
s100a1	Chr 16 ~syntenic	Z, A1	Immune cells	Widely upregulated, low-level expression, hematopoietic cells
s100a10b	Chr 16 Syntenic	A10	Immune cells	Widely upregulated, high expression, hematopoietic cells
s100w	Chr 16 Syntenic	A13, A14, A16	Immune cells	Widely upregulated, moderate expression, hematopoietic cells
icn	Chr 16 Syntenic	A1	Immune cells	Widely upregulated, high expression, hematopoietic cells
icn2	Chr 16 Syntenic	A1	Immune cells	Widely upregulated, high expression, hematopoietic cells
s100t	Chr 16 ~syntenic	A1	Low in immune cells	Epithelial cells

Recombinant zebrafish s100 proteins fold and interact with calcium

We chose to functionally characterize a subset of the zebrafish s100s which seemed most promising to behave like calprotectin based on genomic context and gene expression profiles: s100a10b, s100a1, s100 and icn. We left out icn2 because it is very similar to icn by all metrics

including sequence identity (87.37%: they differ by 14 amino acids, 8 of these at the termini; **Fig 4.1**).

We started by structurally characterizing the four selected zebrafish proteins. We used AlphaFold2 to predict structures for all four proteins.^{183–185} Overlaying the predicted structures with the crystal structure of human calprotectin shows high predicted structural similarity (**Fig 4.2A**). High α -helical content is a shared feature of all known S100 proteins, as well as the predicted zebrafish s100 structures. We tested whether this held for zebrafish s100 proteins by heterologously expressing and purifying the proteins from *E. coli* and then measuring their secondary structure content by far-UV circular dichroism (CD). This revealed signal minima at 208 and 222 nm consistent with primarily α helical structures (**Figure 4.2B**).

Most S100 proteins also bind calcium and undergo a conformational change exposing a hydrophobic binding surface.^{236,237} We tested whether this held for the four zebrafish S100 proteins by measuring calcium-induced changes in protein secondary and tertiary structure by far-UV CD and intrinsic fluorescence, respectively. We found that all four recombinantly expressed zebrafish proteins exhibited evidence of calcium-induced conformational change (**Fig 4.2B-C**).

Upon addition of saturating calcium, zebrafish s100a10b and s100a1 exhibited an increase in helical content, while s100w and icn, in contrast, show little change in secondary structure (**Fig 4.2B**). The intrinsic fluorescence of all four proteins, however, responded to calcium (**Fig 4.2C**). Intrinsic fluorescence captures changes in the local chemical environments of tyrosine and tryptophan residues, suggesting that calcium binding induces a change in the tertiary structure of all s100 proteins. This is consistent with the canonical calcium-induced rotation of the third helix relative to the other helices of S100s.²³⁶

Taken together, these results show that these four zebrafish S100 proteins are folded, bind to calcium, and undergo the calcium-induced conformational changes expected for members of the family.

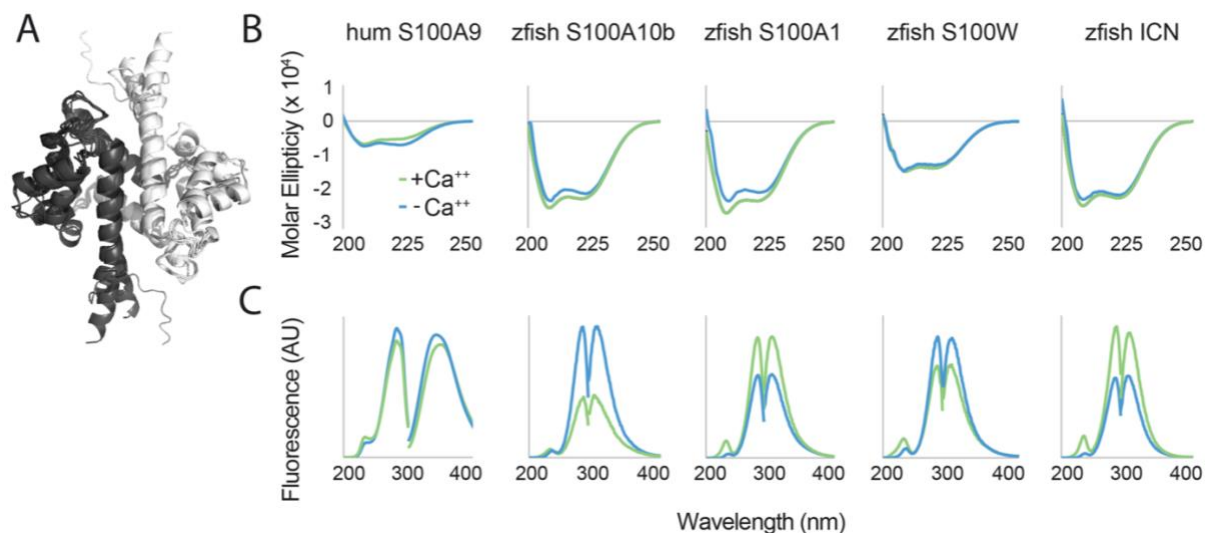


Figure 4.2: Structural comparisons of human and zebrafish S100s. A) Overlaid AlphaFold structure predictions for all S100 homodimers in this dataset, as well as the human S100A8/S100A9 heterodimer (5W1F RCSB ID). Different chains of each homodimer are shown in black or white. B) Far UV circular dichroism spectra for each protein in the presence of 2 mM Ca^{++} (green) and then adding 5mM EDTA (blue). Units are in molar ellipticity ($\text{deg} \times \text{cm}^2/\text{dmol}$) over wavelength (nm). C) Fluorescence excitation and emission spectra for each protein in the presence or absence of calcium (green and blue, respectively). The fluorescence units are arbitrary; the x-axis is wavelength in nanometers. Excitation spectra were collected while observing fluorescence at the maximum emission wavelength; emission spectra were collected while exciting at the maximum excitation wavelength.

Zebrafish s100s do not exhibit nutritional immunity like calprotectin

One of the most important biological functions of human calprotectin is antimicrobial activity via nutritional immunity. We evaluated the antimicrobial abilities of each of the four zebrafish s100 proteins against human-derived *Staphylococcus epidermidis* and zebrafish-derived *Vibrio ZWU0020* and *Aeromonas ZOR001* strains. *S. epidermidis* was previously shown to be susceptible to human calprotectin^{238,239}; the response of the zebrafish-derived strains is unknown. Fig 4.3A shows the dose-dependent antimicrobial activity of human calprotectin

against each strain over 13 hours in nutrient rich media across three biological replicates. For all three strains, increasing amounts of calprotectin (from blue to green) leads to decreased final OD₆₀₀ values (indicated by black arrows). We quantified this response by measuring difference in area under the OD₆₀₀ curve from 0-13 hrs with and without calprotectin (Δ AUC). We then plotted Δ AUC as a function of protein concentration. A negative Δ AUC value indicates growth inhibition at the indicated s100 concentration, while a zero or positive value indicates no antimicrobial activity. This revealed a calprotectin-dependent decrease in growth for all three bacterial strains (**Fig 4.3B**, yellow curves).

To assess the nutritional immunity capacity of the zebrafish s100s, we performed identical experiments using each of the four proteins. We calculated Δ AUC curves for each bacterial strain under increasing concentrations of each s100 (**Fig 4.3B**). Unlike the effect of human calprotectin, none of the four zebrafish s100 proteins exhibited nutritional immunity (**Fig 4.3B**). Human calprotectin was the only protein to exhibit nutritional immunity under these conditions (yellow curve). Bacteria treated with zebrafish s100a1 showed improved growth relative to bacteria in the absence of s100 (brown curve). Zebrafish s100a10b increased growth of *S. epidermidis* and had no effect on growth of zebrafish-derived bacterial strains (orange). Zebrafish s100w (purple) improved *S. epidermidis* growth, showed a possible slight inhibitory effect on *Aeromonas ZOR001* for concentrations at or above 50 μ M, and did not affect *Vibrio ZWU0020* growth at any concentration. Similarly, zebrafish ICN (blue) improved *S. epidermidis* growth, had a slight inhibitory effect on *Aeromonas ZOR001* when used at or above 50 μ M, and increased *Vibrio ZWU0020* growth at all concentrations.

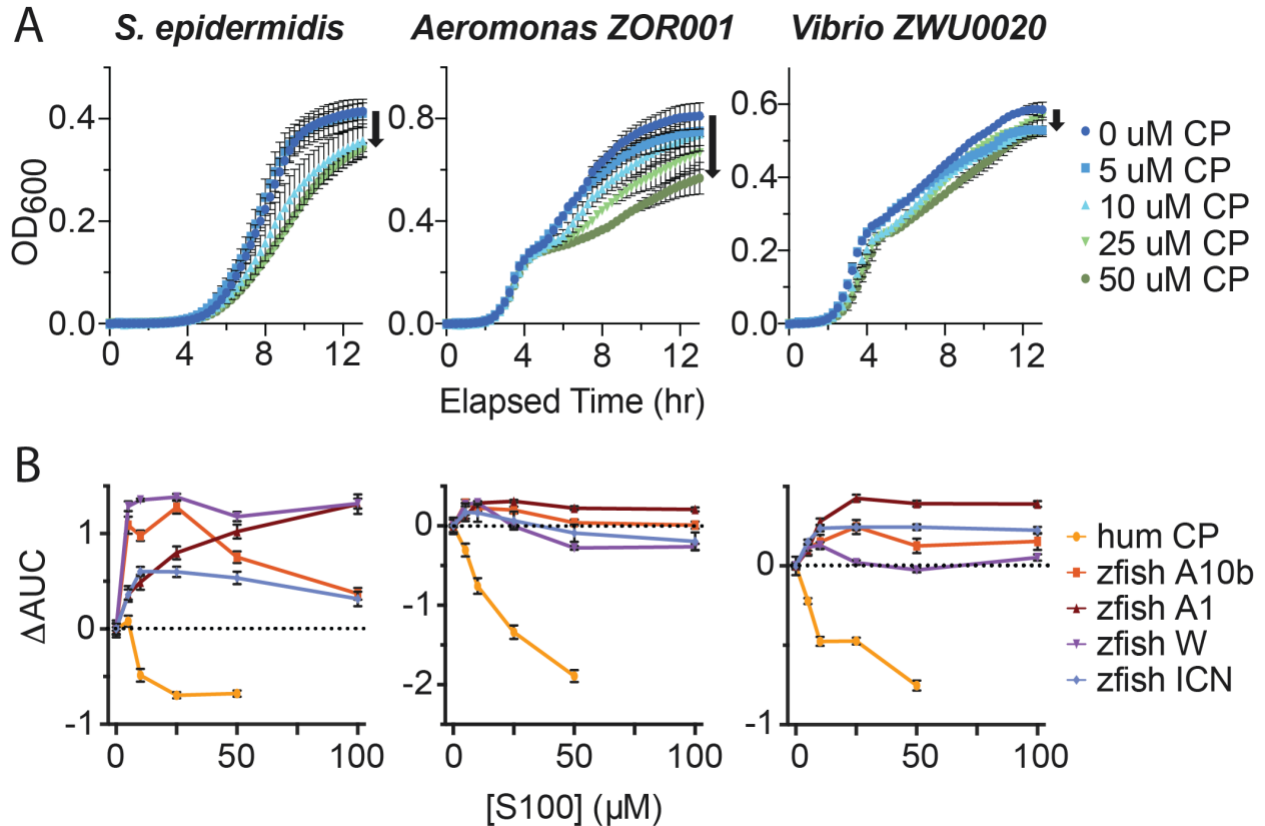


Figure 4.3: Zebrafish s100s do not exhibit nutritional immunity activity like human calprotectin. All data were collected in biological and technical triplicate. Error bars indicate standard error. **A)** Dose dependence of human calprotectin challenge on human and zebrafish commensal bacteria. Each column specifies which bacteria was used for the set of nutritional immunity assays: a human-derived Gram-positive *Staphylococcus epidermidis* and two zebrafish-derived Gram-negative bacteria, *Aeromonas* strain: *ZOR001* and *Vibrio* strain: *ZWU0020*. Bacterial growth was measured by OD₆₀₀ over 13 hours after challenge with increasing doses of human calprotectin noted in the legend on the right. Concentration increases from dark blue to dark green and black arrows indicate how growth is affected as calprotectin increases. **B)** Zebrafish s100 dose effects on human and zebrafish commensal bacterial growth compared to human calprotectin. The dotted line at zero represents no effect of s100 challenge on bacterial growth. Each datapoint shows the change in the area under the curve from the absence of s100 protein to the indicated s100 concentration, measured from growth curves like those shown in panel A.

Zebrafish s100s do not exhibit proinflammatory activity like S100A9

Antimicrobial activity is not the only function of mammalian calprotectin. Calprotectin is a heterodimer of S100A8 and S100A9. The homodimer of mammalian S100A9, for example, can activate an innate immune response through the Toll-like receptor 4 complex (TLR4),

inducing nuclear localization of NF- κ B and transcription of a wide variety of pro-inflammatory proteins. This activity can be reproduced in an *in vitro* functional assay by transfecting HEK293T cells with plasmids encoding the proteins of the TLR4 complex (TLR4, MD-2, and CD14), as well as a plasmid placing luciferase behind an NF- κ B promoter. We can treat the cells with exogenous S100A9 and measure the luciferase response.

We wanted to see if our purified zebrafish s100 proteins could play a similar role; therefore, we tested the ability of these proteins to activate TLR4 in this assay (**Fig 4.4**). Zebrafish have three orthologs of tetrapod TLR4: TLR4ba, TLR4bb, and TLR4al. Zebrafish TLR4ba has been shown to induce inflammation in response to endotoxin: the small molecule lipopolysaccharide (LPS)¹⁰¹ but neither TLR4bb nor TLR4al showed activity. We validated our assay by testing the ability of each complex to activate in response to the canonical agonist for the receptor, endotoxins derived from Gram-negative bacterial outer membranes (green). As expected, the human and TLR4ba complexes responded strongly to endotoxin. TLR4bb and TLR4al did not show signal above vehicle treatment (light blue). We next challenged all four complexes with human S100A9 (yellow). We found that 2 μ M human S100A9 activated human TLR4, as expected, but that none of the zebrafish TLR4 complexes showed signal above background. This suggests that this pro-inflammatory activity is not a conserved function in zebrafish. Finally, we tested the ability of zebrafish s100s to activate each TLR4 complex at 2 μ M: we observed no statistically significant agonist activity for any protein. Although we detected a small amount of signal zebrafish s100a10b (orange) and s100w (brown), this was not statistically significant (Bonferroni-corrected one-sample t-test). We repeated the same experiment with 10 μ M zebrafish s100s and observed no convincing agonist activity.

These zebrafish s100 proteins show no evidence of activity against either human TLR4 or the three zebrafish tlr4 ohnologs. Given the potent response of these receptors to positive controls (endotoxin or human S100A9), this strongly suggests the activity is not present in these proteins. Further, with this assay, false positives are common due to endotoxin contamination from the heterologously expressed proteins. The lack of signal thus gives strong evidence that these zebrafish proteins cannot activate TLR4 in the same fashion as human S100A9.

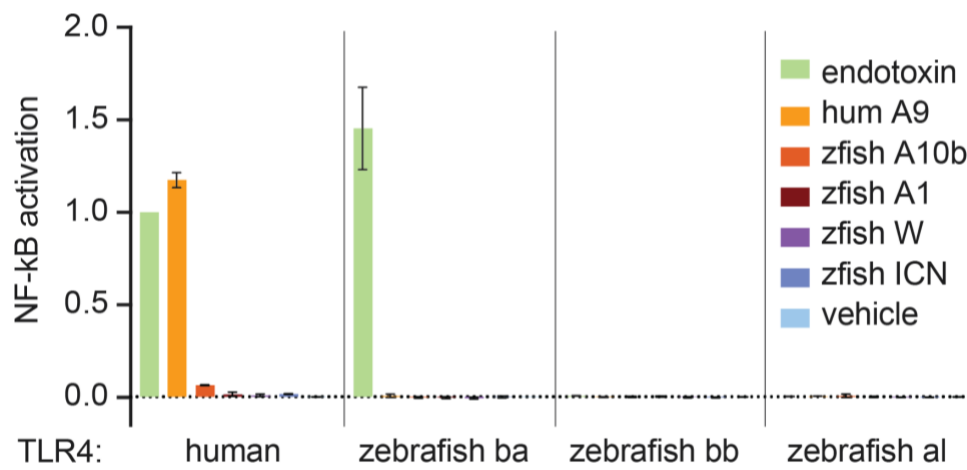


Figure 4.4: Zebrafish s100s do not exhibit the pro-inflammatory characteristics of human S100A9. Activation of human and zebrafish TLR4 complexes in the presence of zebrafish s100 proteins. Bars show the average signal across three biological replicates, with error bars indicating standard error. The positive controls for this experiment included human TLR4 and zebrafish TLR4ba treated with endotoxin (green), and human TLR4 treated with human S100A9 (yellow). There is no known agonist for zebrafish TLR4bb and al complexes. All data was background subtracted and normalized to the signal from human TLR4 treated with endotoxin.

DISCUSSION

Employing zebrafish as a model for studies of innate immunity is a promising field of work. As vertebrates, zebrafish share much of their physiology and immune defense mechanisms with humans, thus enabling mechanistic insight into health, disease, and host-microbe interactions. However, when approaching this research, we must be cognizant of the more than

400 million years since our most recent common ancestor, allowing for species-specific differences evolved to cope with diverse environments and other pressures.

We assert here that recent studies and commercial products have made the incorrect assumption that calprotectin exists in the fish innate immune response. We re-evaluate phylogenetic evidence to look for a homolog of calprotectin in fish and confirm that, although the evolutionary history of S100 proteins is messy, zebrafish do not share an s100 protein within the clade containing mammalian calprotectin.

We also tested the possibility that a fish s100 protein from a similar genomic context to calprotectin might have evolved calprotectin-like innate immune functions. We characterized the nutritional immunity and proinflammatory activity of four zebrafish s100s, including the previously studied zebrafish “calprotectin” s100a10b, in assays which are normally used to test calprotectin function. None of the zebrafish proteins performed like human calprotectin.

We cannot prove a negative: this work does not show that no zebrafish s100 protein does some subset of the functions of human calprotectin. Our work does, however, put the burden of proof on researchers who would claim such functions exist. If a zebrafish s100 has antimicrobial or pro-inflammatory activity, it must have evolved that activity convergently and independently of those activities from mammalian calprotectin. Further, and importantly, such a protein does not shed direct light on mammalian biology. A convergent zebrafish s100 calprotectin-like protein would help us understand zebrafish biology; it would also be intriguing from the perspective of protein evolution. It would almost certainly have a different regulatory scheme and would only have a subset of human calprotectin functions.

We also did not test all possible zebrafish S100 proteins because we focused on those that seemed most promising to be expressed in immune cells during the immune response and shared

genomic origin. Zebrafish s100 proteins represent a largely uncharacterized set of new proteins, many of which are specific to the ray-finned fishes.²³⁰ We are excited to see how investigations of their functional roles continue. We remain intrigued by the idea that convergent evolution may exist between mammalian calprotectin and some other protein(s) in zebrafish. In addition to studying zebrafish s100 proteins individually, one important line of work will be to investigate various s100 heterocomplexes. Mammalian calprotectin is a heterodimer and heterotetramer, but we only performed experiments with homodimeric zebrafish s100s. Studies of other heterodimeric human S100 complexes have been done and prove to have altered functions.²⁴⁰ In the future, this type of analysis could be done with zebrafish proteins to explore whether a heterodimer state confers nutritional immunity or proinflammatory activity. However, there is currently no evidence at this point suggesting this is likely.

We conclude that it is crucial that we use an evolutionary lens and careful biochemical analyses to probe homology between zebrafish and human proteins so that we can make accurate extrapolations of findings from zebrafish models of human biology.

MATERIALS AND METHODS

Protein Purification

We purchased all zebrafish s100 genes from GenScript in the pET-28a(+)-TEV vector with an N-terminal 6x-Histidine tag and TEV protease cleavage site. All genes were codon-optimized for expression in *E. coli*. We expressed human calprotectin (with S100A9 containing the C3S mutation) and human S100A9/C3S in a pET-Duet vector without purification tags. We transformed Rosetta2(DE3)pLysS *E. coli* cells with plasmids. We used transformant glycerol stocks to inoculate cultures in 15 mL Luria broth (LB) with 50 µg/mL kanamycin and 34 µg/mL

chloramphenicol. We incubated cultures overnight at 37 °C, shaking at 250 rpm. The following day, we diluted 15 mL saturated cultures into 1.5 L of LB with antibiotics. When the OD₆₀₀ reached 0.6-1.0, we induced recombinant protein expression with 1 mM IPTG and 0.2% glucose and then grew overnight at 16 °C, shaking at 250 rpm. We pelleted cells at 3,000 rpm for at least 15 minutes in an F6B rotor in a Beckman Coulter preparative centrifuge. We stored pellets at -20 °C for up to one month.

We prepared protein lysates for purification with the following method: We vortexed pellets (6-9 g) in 45 mL buffer from the first chromatography step (see below) until cells were resuspended, added 15 µL each of DNase I and Lysozyme (ThermoFisher Scientific), and incubated at room temperature with gentle shaking for at least 10 minutes. We sonicated the resuspended cells at 55% amplitude with 0.3 second pulse on, 0.7 second pulse off, for 3-5 minutes. We pelleted cell debris by centrifugation at 15,000 rpm at 4 °C for at least 20 minutes in a JA-20 rotor in a Beckman Coulter preparative centrifuge and collected the supernatant. To remove remaining large debris, we filtered lysate supernatant through a 0.2 µm pore syringe filter immediately prior to purification chromatography.

We purified all proteins using an Äkta PrimePlus Fast Protein Liquid Chromatography system using two stacked 5 mL HiTrap columns at each step. We used HisTrap FF columns for Ni-affinity and Q HP columns for anion exchange (GE Health Science). All chromatography was performed at 4 °C. At the end of purification, we confirmed protein purity was >95% by SDS-PAGE. Then, we dialyzed each protein overnight into 4 L of 25 mM Tris, 100 mM NaCl, pH 7.4 at 4 °C. We placed 2 g/L Chelex 100 resin (Bio-Rad) in the dialysis buffer to remove divalent metal ions. We concentrated each protein to roughly 2 mg/mL and syringe-filtered through a 0.22 µm pore filter directly into liquid nitrogen to sterilize and flash freeze before storing at -80 °C.

We purified TEV-cleavable 6xHis-tagged zebrafish S100 proteins with the following scheme. We used 25 mM Tris, 100 mM NaCl, pH 7.4 buffer as the base for all chromatography buffers. We ran our protein lysate over a Ni-affinity column with a 50 mL wash and eluted over a 75 mL gradient from 25-1000 mM imidazole to collect proteins with strong Ni binding capacity. We determined which fractions contained our desired protein by SDS-PAGE and pooled these fractions. To separate our recombinant proteins from their Ni-binding His-tag, we added 5 mM DTT and 6xHis-tagged TEV protease to the pooled fractions and incubated the reaction at room temperature with gentle shaking for at least 5 hours. We then dialyzed the protein solution overnight into 4 L buffer with 25 mM imidazole and 5 mM DTT to allow cleavage to come to completion and to remove excess imidazole from the sample. We performed a second round of Ni-affinity chromatography. Without the His-tag, the zebrafish S100s have low affinity for Ni. Therefore, we isolated pure, non-tagged zebrafish S100 proteins at this step during a 50 mL wash in 25 mM imidazole and then used a step gradient to 1 M imidazole to elute His-tagged and other contaminant proteins that had higher affinity for the Ni column. Purified zebrafish S100s were prepared for storage as described above.

We purified human calprotectin using Ni-affinity chromatography at pH 7.4 and anion exchange at pH 8. When expressing calprotectin, S100A8 and S100A9 homodimers are also expressed and must be removed during chromatography. In the presence of calcium, S100A9 and calprotectin bind divalent metal ions like Ni, but S100A8 and most other lysate proteins do not. We loaded our Ni-affinity column with calprotectin lysate, washed away most A8 and contaminants in a 50 mL wash, and then eluted calprotectin and S100A9 over a 75 mL gradient from 0-1000 mM imidazole and 1-0 mM CaCl₂ in 25 mM Tris, 100 mM NaCl, pH 7.4 buffer. We pooled elution peak fractions containing calprotectin and contaminant S100A9 homodimers, as

determined by SDS-PAGE, and dialyzed overnight in 4 L of 25 mM Tris, 100 mM NaCl at pH 8. We loaded our sample onto an anion exchange chromatography column in 25 mM Tris, 100 mM NaCl at pH 8 with 100 mM NaCl. Because A9 has a lower pI than calprotectin, it binds the anion column more strongly at pH 8. We used a 50 mL wash in 100 mM NaCl to isolate calprotectin and then used a step gradient increasing the salt to 1 M NaCl to remove S100A9 and other contaminants from the column. At this point, calprotectin was pure and prepared for storage as described, and fractions with S100A9 and other contaminants were discarded.

We used a similar protocol to purify human S100A9. We performed Ni-affinity chromatography as described for calprotectin. We then performed anion exchange chromatography using a 50 mL wash and collected fractions over a 70 mL gradient elution from 100-1000 mM NaCl in 25 mM Tris, pH 8 buffer to isolate S100A9 from contaminant proteins that also bind the anion column. We used SDS-PAGE to confirm fractions with S100A9, pooled and dialyzed these fractions overnight into 4 L of 25 mM Tris, 100 mM NaCl at pH 6. As a final step, we loaded the A9 sample onto an anion exchange column in 25 mM Tris, 100 mM NaCl buffer at pH 6. S100A9 binds weakly to the anion column at pH 6. Therefore, we collected S100A9 in a 50 mL wash in 100 mM NaCl, and then removed contaminants from the column with a step elution at 1 M NaCl. Pure S100A9 was prepared for storage as described.

Far-UV Circular Dichroism and Fluorescence Spectroscopy

Prior to biophysical measurements, we thawed and exchanged all proteins into 25 mM Tris, 100 mM NaCl, pH 7.4 via overnight dialysis in 4 L buffer at 4 °C. We determined protein concentrations by Bradford Assay using bovine serum albumin (BSA) standards and the molecular weight of each dimeric structure, then diluted to ~10 μ M in dialysis buffer.

For all spectroscopic measurements, we assessed metal-induced changes to the spectra by measuring the spectrum in the presence of 2 mM CaCl₂ and then adding excess EDTA at 5 mM and re-measuring the spectrum. We collected far-UV circular dichroism data between 200–250 nm using a J-815 CD spectrometer (Jasco) with a 1 mm quartz spectrophotometer cell (Starna Cells, Inc. Catalog No. 1-Q-1). We collected 3 scans for each condition, and then averaged the spectra and subtracted a blank buffer spectrum using the Jasco spectra analysis software suite. We converted raw ellipticity into mean molar ellipticity using the concentration and number of residues in each protein. We collected intrinsic tyrosine and/or tryptophan fluorescence using a J-815 CD spectrometer (Jasco) with an attached model FDT-455 fluorescence detector (Jasco) using a 1 cm quartz cuvette (Starna Cells, Inc.). We collected a single excitation and emission scan at 10 nm/min with a 10 nm bandwidth, 1 nm data pitch, and 1 sec D.I.T. for each condition and then subtracted a blank buffer spectrum using the Jasco spectra analysis software suite. Depending on the sample signal, we set the detector sensitivity to either 630 or 800 Volts. We conducted excitation scans by measuring 305 nm light emitted at for all zebrafish proteins and 345 nm emitted light for human S100A9 for each excitation wavelength from 200-295 nm. For emission scans we used 280 nm light to excite zebrafish proteins and 288 nm for human S100A9, and measured light emitted at all wavelengths from 285-425 nm.

Nutritional Immunity Assay

We measured the antimicrobial activity of zebrafish S100s and human calprotectin against human- and zebrafish-derived bacterial strains using a modified version of a well-established assay^{86,216,219,238} that will be described here. Bacterial strains used in this assay include 1) *Staphylococcus epidermidis*, a human commensal strain previously shown to respond

to calprotectin^{219,238}; 2) *Aeromonas ZOR001*, isolated from zebrafish and not previously characterized for response to calprotectin; and 3) *Vibrio ZWU0020*, isolated from zebrafish and not previously characterized for response to calprotectin but related to human-derived *Vibrio cholerae* shown to respond to calprotectin.¹⁰⁷ We obtained both zebrafish-derived strains from the Guillemin lab at the University of Oregon.

Each week, we plated bacterial strains from glycerol stocks onto antibiotic-free LB agar and grew at 30 °C overnight before storing plates at 4 °C. The day before an experiment, we inoculated a 5 mL culture in liquid LB media with a single colony from each strain and grew overnight at 30 °C with shaking. The following day, we diluted cultures 1:100 in 5 mL LB and grew to an OD₆₀₀ around 0.8 by the time of the experiment. *Aeromonas ZOR001* and *Vibrio ZWU0020* were diluted 2 hours before the experiment. *S. epidermidis* grew more slowly so required dilution 4 hours prior to the experiment.

The day before each experiment, we thawed a single S100 protein from -80 °C, concentrated to at least 200 µM using a Nanosep 3K Omega spin concentrator (Pall Corporation), and dialyzed overnight at 4 °C into 4 L of Experimental Buffer (25 mM Tris, 100 mM NaCl, pH 7.4) with 2 g/L Chelex 100 resin (Bio-Rad) to chelate residual transition metal ions. After dialysis, we filter-sterilized the protein through a Ultrafree-MC-VV centrifugal filter with Durapore PVDF 0.1 µm and kept at 4 °C until time of experiment.

To start the experiment, we made a protein dilution series by mixing a desired amount of protein in sterile Experimental Buffer with the appropriate amount of LB to achieve a ratio of 62:38. We then brought the volume of these protein solutions up to 1.7 mL in Experimental Media (EM). We made EM by diluting 62:38 Experimental Buffer:LB, and filter-sterilized. We distributed each sample in aliquots of 160 µL across ten wells of a clear Falcon 96-Well, Cell

Culture-Treated, Flat-Bottom Microplate. At this time, we diluted each bacterial strain to an estimated OD₆₀₀ of .008 in 5 mL Experimental Media with calcium (EMC). We made EMC by adding 10.2 μM CaCl₂ to EM, and sterile-filtered. Then, we added 40 μL of dilute bacteria or EMC without bacteria (contamination control) to each well, making technical triplicate conditions for bacterial strains. To counteract sample evaporation, the outermost wells of the plate contained 160 μL EM and 40 μL EMC only and we wrapped the plate in a single layer of parafilm.

We measured bacterial growth by OD₆₀₀ every 15 minutes over 13 hours in a Molecular Devices SpectraMax i3. The plate was shaken for 5 seconds before the first read, then for 10 minutes between each subsequent read. We set the plate reader temperature to 25 °C, however, over the course of the overnight growth, the actual temperature reached 37 °C. The final concentration of metals in the media without bacteria was measured using ICP-MS. The measured concentrations were Ni: 45.4 μM, Ca: 107.3 μM, Cu: 157.4 μM, Mg: 160.5 μM, Mn: 216.6 μM, Fe: 1.1 mM, and Zn: 5.9 mM.

For the analysis, we background subtracted each experimental condition using OD₆₀₀ values for the matching concentration of S100 protein concentration in buffer without bacteria added. We used Prism to average the replicates by condition, determine the standard error, and graph the results.

Proinflammatory Activity Assay

We tested the S100A9-like proinflammatory activity of zebrafish S100s using a well-established assay.^{101,102,238} This assay measures relative activation of the TLR4-mediated immune response through NF-κB. For each experiment, we thawed all zebrafish S100 proteins and

human S100A9 from -80 °C, buffer exchanged into endotoxin-free PBS, then treated with endotoxin removal spin columns (ThermoFisher Scientific) to remove LPS residual from the purification process.

We performed each experiment in technical triplicate and followed the Dual-Glo Luciferase Assay System protocol (Promega). We transiently transfected adherent HEK293T cells in a Falcon 96-Well, Cell Culture-Treated, Flat-Bottom Microplate with pcDNA vector plasmids using PLUS and Lipofectamine Reagents (ThermoFisher Scientific). Plasmids contained genes for human or zebrafish TLR4 complex components and *Renilla* luciferase enzyme under constitutively active promoters, and the firefly luciferase gene controlled by an NF- κ B promoter. For human TLR4 complex transfections, we transfected 10 ng human TLR4, 0.5 ng human MD-2, and 1 ng human CD14 plasmids per well. For zebrafish TLR4 complex transfections, we used 10 ng zebrafish TLR4, 20 ng zebrafish MD-2, and 1 ng mouse CD14 plasmids per well, as this ratio gives us the best signal to noise ratio. Zebrafish do not have an annotated CD14, but previous studies have shown zebrafish TLR4ba can be activated in the presence of mouse and human CD14, but more strongly with mouse. We also transfected all wells with 1 ng *Renilla* plasmid, 20 ng elam-Luc (firefly), and brought the total DNA mass per well to 100 ng with empty pcDNA vector in a total media volume of 200 μ L per well.

After 20-24 hours incubation at 37 °C in 5% CO₂, we removed all 200 μ L of transfection mix from each well. We then treated transfected HEK293T cells with 100 μ L of one of the following treatment mixes: 1) 2 μ M S100 protein and 200 ng/ μ L Polymyxin B to bind up LPS in media, 2) 0.2 ng/ μ L LPS-R (tlrl-eklps; Invivogen) as a positive control for human TLR4, or 3) 2 ng/ μ L lipid IVa as a positive control for zebrafish TLR4ba activation. Because there is no known activator of zebrafish TLR4bb and TLR4al complexes (Loes et al., 2021), we treated these

transfections with 2 ng/ μ L lipid IVa for consistency. After incubating again at 37 °C in 5% CO₂ for 3-4 hours, we removed and discarded 60 μ L of treatment mix from each well. We chemically lysed the cells by adding 30 μ L Dual Glo lysis reagent containing firefly luciferin and incubated in the dark for 7 minutes. We then mechanically lysed the cells by scraping the bottom of each well with a pipet tip and transferring 60 μ L of cell solution to an opaque 96-well plate. After a 7-minute incubation in the dark at room temperature, we measured luminescence per well produced by firefly luciferase activity using a Molecular Devices SpectraMax i3. Then we added 30 μ L of Dual-Glo Stop & Glo buffer containing firefly luciferase quencher and *Renilla* luciferase reagent, incubated for 7 more minutes, and measured luminescence.

For the analysis, we took the firefly signal for each experimental condition and background subtracted the averaged firefly signal of wells transfected with the corresponding complex but treated with buffer without agonist. We did the same for the *Renilla* signal, with background signal considered as the averaged signal from wells with same treatment condition but transfected only with vector. We divided the background-subtracted firefly signal for each well by the background-subtracted *Renilla* signal for that same well. To simplify comparisons across experiments, we normalized the firefly/*Renilla* value for each well to the triplicate average of the firefly/*Renilla* values for human TLR4 complex treated with 0.2 ng/ μ L LPS-R.

BRIDGE TO CHAPTER V

We conclude from this chapter that zebrafish do not have an ortholog of human calprotectin and likely do not possess an s100 protein that convergently evolved similar antibacterial and proinflammatory activities. Because S100 proteins generally have the capacity for metal ion binding, further investigation into zebrafish s100 complexes might yield

discoveries of species-specific host-microbe interactions in the fight for metal sequestration. Adding to our conclusions of zebrafish TLR4/MD-2, it is possible that the functions of these important human immune system proteins are paralleled in some way in the zebrafish. But our ~430 million years of unique selective pressures seem to have changed the players and strategies at the host-microbe interface.

CHAPTER V

SUMMARY AND CLOSING REMARKS

Our functional characterizations of homologous human and zebrafish immune proteins have shown us remarkable differences in the way these two species respond to microbes. We find that the zebrafish TLR4 has high specificity for tetra-acylated LPS molecules which inhibit human TLR4 signaling. But even though the zebrafish complex can activate an inflammatory response to tetra-acyl LPS *in vitro*, injecting live fish with this potent TLR4 activator does not stimulate the immune response. This completely contrasts the human and mouse systems, where a stronger TLR4 agonist induces a stronger immune response, which can lead to death caused by an overactive immune system. Resurrected ancestral proteins from early vertebrates suggest that this hyperactive TLR4 response to LPS evolved at some point between the ancestor of bony vertebrates and tetrapods but is not maintained by all tetrapod species. The reconstructed bony vertebrate and teleost ancestor proteins show low-level stimulation by LPS and suggest that zebrafish have evolved a unique sensitivity to tetra-acyl LPS. What could explain this evolution of ligand specificity in the absence of functional consequence? Perhaps we have yet to reveal the true role of zebrafish TLR4.

Similar to our story of TLR4, zebrafish do not appear to have a functional ortholog of human calprotectin. Calprotectin plays several roles in the human defense against pathogens: at sites of inflammation, calprotectin chelates transition metal ions that are essential for microbial growth and so inhibits bacterial growth at wound sites. Calprotectin can also amplify the immune response by activating TLR4 and other damage-sensing immune receptors. But zebrafish do not have an ortholog of calprotectin. Our studies suggest that none of the zebrafish proteins that

share homology with human calprotectin can serve either this antibacterial or proinflammatory role. Have zebrafish evolved alternative compensatory mechanisms to deal with infections? What new immune strategies can we learn from studying the zebrafish immune response in the absence of these proteins that humans rely so heavily on?

Our work demonstrates it is necessary to consider the long evolutionary divergence between human and zebrafish when extrapolating findings from model systems.

REFERENCES CITED

- (1) Bäckhed, F.; Roswall, J.; Peng, Y.; Feng, Q.; Jia, H.; Kovatcheva-Datchary, P.; Li, Y.; Xia, Y.; Xie, H.; Zhong, H.; Khan, M. T.; Zhang, J.; Li, J.; Xiao, L.; Al-Aama, J.; Zhang, D.; Lee, Y. S.; Kotowska, D.; Colding, C.; Tremaroli, V.; Yin, Y.; Bergman, S.; Xu, X.; Madsen, L.; Kristiansen, K.; Dahlgren, J.; Wang, J. Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host & Microbe* **2015**, *17* (5), 690–703. <https://doi.org/10.1016/j.chom.2015.04.004>.
- (2) Houghteling, P. D.; Walker, W. A. Why Is Initial Bacterial Colonization of the Intestine Important to Infants' and Children's Health? *J. pediatr. gastroenterol. nutr.* **2015**, *60* (3), 294–307. <https://doi.org/10.1097/MPG.0000000000000597>.
- (3) Lopez, L. R.; Bleich, R. M.; Arthur, J. C. Microbiota Effects on Carcinogenesis: Initiation, Promotion, and Progression. *Annu. Rev. Med.* **2021**, *72* (1), 243–261. <https://doi.org/10.1146/annurev-med-080719-091604>.
- (4) Michán-Doña, A.; Vázquez-Borrego, M. C.; Michán, C. Are There Any Completely Sterile Organs or Tissues in the Human Body? Is There Any Sacred Place? *Microbial Biotechnology* **2024**, *17* (3), e14442. <https://doi.org/10.1111/1751-7915.14442>.
- (5) Pawelek, J. M.; Low, K. B.; Bermudes, D. Bacteria as Tumour-Targeting Vectors. *The Lancet Oncology* **2003**, *4* (9), 548–556. [https://doi.org/10.1016/S1470-2045\(03\)01194-X](https://doi.org/10.1016/S1470-2045(03)01194-X).
- (6) Limon, J. J.; Skalski, J. H.; Underhill, D. M. Commensal Fungi in Health and Disease. *Cell Host & Microbe* **2017**, *22* (2), 156–165. <https://doi.org/10.1016/j.chom.2017.07.002>.
- (7) Sender, R.; Fuchs, S.; Milo, R. Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. *Cell* **2016**, *164* (3), 337–340. <https://doi.org/10.1016/j.cell.2016.01.013>.
- (8) Luckey, T. D. Introduction to Intestinal Microecology. *The American Journal of Clinical Nutrition* **1972**, *25* (12), 1292–1294. <https://doi.org/10.1093/ajcn/25.12.1292>.
- (9) Ogunrinola, G. A.; Oyewale, J. O.; Oshamika, O. O.; Olasehinde, G. I. The Human Microbiome and Its Impacts on Health. *International Journal of Microbiology* **2020**, *2020*, 1–7. <https://doi.org/10.1155/2020/8045646>.
- (10) McFall-Ngai, M. J. Unseen Forces: The Influence of Bacteria on Animal Development. *Developmental Biology* **2002**, *242* (1), 1–14. <https://doi.org/10.1006/dbio.2001.0522>.
- (11) Bry, L.; Falk, P. G.; Midtvedt, T.; Gordon, J. I. A Model of Host-Microbial Interactions in an Open Mammalian Ecosystem. *Science* **1996**, *273* (5280), 1380–1383. <https://doi.org/10.1126/science.273.5280.1380>.
- (12) Umesaki, Y.; Setoyama, H.; Matsumoto, S.; Imaoka, A.; Itoh, K. Differential Roles of Segmented Filamentous Bacteria and Clostridia in Development of the Intestinal Immune

- System. *Infect Immun* **1999**, 67 (7), 3504–3511. <https://doi.org/10.1128/IAI.67.7.3504-3511.1999>.
- (13) Stappenbeck, T. S.; Hooper, L. V.; Gordon, J. I. Developmental Regulation of Intestinal Angiogenesis by Indigenous Microbes via Paneth Cells. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, 99 (24), 15451–15455. <https://doi.org/10.1073/pnas.202604299>.
- (14) Bates, J. M.; Mittge, E.; Kuhlman, J.; Baden, K. N.; Cheesman, S. E.; Guillemin, K. Distinct Signals from the Microbiota Promote Different Aspects of Zebrafish Gut Differentiation. *Developmental Biology* **2006**, 297 (2), 374–386. <https://doi.org/10.1016/j.ydbio.2006.05.006>.
- (15) Whiteside, S. A.; Razvi, H.; Dave, S.; Reid, G.; Burton, J. P. The Microbiome of the Urinary Tract—a Role beyond Infection. *Nat Rev Urol* **2015**, 12 (2), 81–90. <https://doi.org/10.1038/nrurol.2014.361>.
- (16) Li, C.; Stražar, M.; Mohamed, A. M. T.; Pacheco, J. A.; Walker, R. L.; Lebar, T.; Zhao, S.; Lockart, J.; Dame, A.; Thurimella, K.; Jeanfavre, S.; Brown, E. M.; Ang, Q. Y.; Berdy, B.; Sergio, D.; Invernizzi, R.; Tinoco, A.; Pishchany, G.; Vasani, R. S.; Balskus, E.; Huttenhower, C.; Vlamakis, H.; Clish, C.; Shaw, S. Y.; Plichta, D. R.; Xavier, R. J. Gut Microbiome and Metabolome Profiling in Framingham Heart Study Reveals Cholesterol-Metabolizing Bacteria. *Cell* **2024**, 187 (8), 1834–1852.e19. <https://doi.org/10.1016/j.cell.2024.03.014>.
- (17) Bäckhed, F.; Ley, R. E.; Sonnenburg, J. L.; Peterson, D. A.; Gordon, J. I. Host-Bacterial Mutualism in the Human Intestine. *Science* **2005**, 307 (5717), 1915–1920. <https://doi.org/10.1126/science.1104816>.
- (18) Perry, F.; Arsenault, R. J. The Study of Microbe–Host Two-Way Communication. *Microorganisms* **2022**, 10 (2), 408. <https://doi.org/10.3390/microorganisms10020408>.
- (19) Chaplin, D. D. 1. Overview of the Immune Response. *Journal of Allergy and Clinical Immunology* **2003**, 111 (2), S442–S459. <https://doi.org/10.1067/mai.2003.125>.
- (20) Delneste, Y.; Beauvillain, C.; Jeannin, P. Immunité Naturelle: Structure et Fonction Des Toll-like Receptors. *Med Sci (Paris)* **2007**, 23 (1), 67–74. <https://doi.org/10.1051/medsci/200723167>.
- (21) Akira, S.; Uematsu, S.; Takeuchi, O. Pathogen Recognition and Innate Immunity. *Cell* **2006**, 124 (4), 783–801. <https://doi.org/10.1016/j.cell.2006.02.015>.
- (22) Bertani, B.; Ruiz, N. Function and Biogenesis of Lipopolysaccharides. *EcoSal Plus* **2018**, 8 (1), 10.1128/ecosalplus.ESP-0001–2018. <https://doi.org/10.1128/ecosalplus.esp-0001-2018>.
- (23) Pfeiffer, R. Untersuchungen über das Cholera Gift. *Zeitschr. f. Hygiene.* **1892**, 11 (1), 393–412. <https://doi.org/10.1007/BF02284303>.

- (24) Singer, M.; Deutschman, C. S.; Seymour, C. W.; Shankar-Hari, M.; Annane, D.; Bauer, M.; Bellomo, R.; Bernard, G. R.; Chiche, J.-D.; Coopersmith, C. M.; Hotchkiss, R. S.; Levy, M. M.; Marshall, J. C.; Martin, G. S.; Opal, S. M.; Rubenfeld, G. D.; Van Der Poll, T.; Vincent, J.-L.; Angus, D. C. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* **2016**, *315* (8), 801. <https://doi.org/10.1001/jama.2016.0287>.
- (25) Poltorak, A.; He, X.; Smirnova, I.; Liu, M.-Y.; Huffel, C. V.; Du, X.; Birdwell, D.; Alejos, E.; Silva, M.; Galanos, C.; Freudenberg, M.; Ricciardi-Castagnoli, P.; Layton, B.; Beutler, B. Defective LPS Signaling in C3H/HeJ and C57BL/10ScCr Mice: Mutations in *Tlr4* Gene. *Science* **1998**, *282* (5396), 2085–2088. <https://doi.org/10.1126/science.282.5396.2085>.
- (26) Qureshi, S. T.; Larivière, L.; Leveque, G.; Clermont, S.; Moore, K. J.; Gros, P.; Malo, D. Endotoxin-Tolerant Mice Have Mutations in Toll-like Receptor 4 (*Tlr4*). *The Journal of Experimental Medicine* **1999**, *189* (4), 615–625. <https://doi.org/10.1084/jem.189.4.615>.
- (27) Wright, S. D. Toll, A New Piece in the Puzzle of Innate Immunity. *The Journal of Experimental Medicine* **1999**, *189* (4), 605–609. <https://doi.org/10.1084/jem.189.4.605>.
- (28) Chow, J. C.; Young, D. W.; Golenbock, D. T.; Christ, W. J.; Gusovsky, F. Toll-like Receptor-4 Mediates Lipopolysaccharide-Induced Signal Transduction. *Journal of Biological Chemistry* **1999**, *274* (16), 10689–10692. <https://doi.org/10.1074/jbc.274.16.10689>.
- (29) Raetz, C. R. H.; Whitfield, C. Lipopolysaccharide Endotoxins. *Annu. Rev. Biochem.* **2002**, *71* (1), 635–700. <https://doi.org/10.1146/annurev.biochem.71.110601.135414>.
- (30) Cohen, J. The Immunopathogenesis of Sepsis. *Nature* **2002**, *420* (6917), 885–891. <https://doi.org/10.1038/nature01326>.
- (31) Park, B. S.; Lee, J.-O. Recognition of Lipopolysaccharide Pattern by TLR4 Complexes. *Exp Mol Med* **2013**, *45* (12), e66–e66. <https://doi.org/10.1038/emm.2013.97>.
- (32) Luderitz, O.; Galanos, C.; Lehmann, V.; Nurminen, M.; Rietschel, E. T.; Rosenfelder, G.; Simon, M.; Westphal, O. Lipid A: Chemical Structure and Biological Activity. *Journal of Infectious Diseases* **1973**, *128* (Supplement 1), S17–S29. https://doi.org/10.1093/infdis/128.Supplement_1.S17.
- (33) Erridge, C.; Bennett-Guerrero, E.; Poxton, I. R. Structure and Function of Lipopolysaccharides. *Microbes and Infection* **2002**, *4* (8), 837–851. [https://doi.org/10.1016/S1286-4579\(02\)01604-0](https://doi.org/10.1016/S1286-4579(02)01604-0).
- (34) Takayama, K.; Qureshi, N.; Ribic, E.; Cantrell, J. L. Separation and Characterization of Toxic and Nontoxic Forms of Lipid A. *Clinical Infectious Diseases* **1984**, *6* (4), 439–443. <https://doi.org/10.1093/clinids/6.4.439>.

- (35) Qureshi, N.; Takayama, K.; Kurtz, R. Diphosphoryl Lipid A Obtained from the Nontoxic Lipopolysaccharide of *Rhodopseudomonas Sphaeroides* Is an Endotoxin Antagonist in Mice. *Infect Immun* **1991**, *59* (1), 441–444. <https://doi.org/10.1128/iai.59.1.441-444.1991>.
- (36) Shimazu, R.; Akashi, S.; Ogata, H.; Nagai, Y.; Fukudome, K.; Miyake, K.; Kimoto, M. MD-2, a Molecule That Confers Lipopolysaccharide Responsiveness on Toll-like Receptor 4. *The Journal of Experimental Medicine* **1999**, *189* (11), 1777–1782. <https://doi.org/10.1084/jem.189.11.1777>.
- (37) Visintin, A.; Halmen, K. A.; Latz, E.; Monks, B. G.; Golenbock, D. T. Pharmacological Inhibition of Endotoxin Responses Is Achieved by Targeting the TLR4 Coreceptor, MD-2. *The Journal of Immunology* **2005**, *175* (10), 6465–6472. <https://doi.org/10.4049/jimmunol.175.10.6465>.
- (38) Coats, S. R.; Pham, T.-T. T.; Bainbridge, B. W.; Reife, R. A.; Darveau, R. P. MD-2 Mediates the Ability of Tetra-Acylated and Penta-Acylated Lipopolysaccharides to Antagonize *Escherichia Coli* Lipopolysaccharide at the TLR4 Signaling Complex. *The Journal of Immunology* **2005**, *175* (7), 4490–4498. <https://doi.org/10.4049/jimmunol.175.7.4490>.
- (39) Teghanemt, A.; Zhang, D.; Levis, E. N.; Weiss, J. P.; Gioannini, T. L. Molecular Basis of Reduced Potency of Underacylated Endotoxins. *The Journal of Immunology* **2005**, *175* (7), 4669–4676. <https://doi.org/10.4049/jimmunol.175.7.4669>.
- (40) Saitoh, S. -i. Lipid A Antagonist, Lipid IVa, Is Distinct from Lipid A in Interaction with Toll-like Receptor 4 (TLR4)-MD-2 and Ligand-Induced TLR4 Oligomerization. *International Immunology* **2004**, *16* (7), 961–969. <https://doi.org/10.1093/intimm/dxh097>.
- (41) Ohto, U.; Fukase, K.; Miyake, K.; Satow, Y. Crystal Structures of Human MD-2 and Its Complex with Antiendotoxic Lipid IVa. *Science* **2007**, *316* (5831), 1632–1634. <https://doi.org/10.1126/science.1139111>.
- (42) Park, B. S.; Song, D. H.; Kim, H. M.; Choi, B.-S.; Lee, H.; Lee, J.-O. The Structural Basis of Lipopolysaccharide Recognition by the TLR4–MD-2 Complex. *Nature* **2009**, *458* (7242), 1191–1195. <https://doi.org/10.1038/nature07830>.
- (43) Anderson, J. A.; Loes, A. N.; Waddell, G. L.; Harms, M. J. Tracing the Evolution of Novel Features of Human Toll-like Receptor 4. *Protein Science* **2019**, pro.3644. <https://doi.org/10.1002/pro.3644>.
- (44) Oblak, A.; Jerala, R. The Molecular Mechanism of Species-Specific Recognition of Lipopolysaccharides by the MD-2/TLR4 Receptor Complex. *Molecular Immunology* **2015**, *63* (2), 134–142. <https://doi.org/10.1016/j.molimm.2014.06.034>.
- (45) Ohto, U.; Fukase, K.; Miyake, K.; Shimizu, T. Structural Basis of Species-Specific Endotoxin Sensing by Innate Immune Receptor TLR4/MD-2. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109* (19), 7421–7426. <https://doi.org/10.1073/pnas.1201193109>.

- (46) Kim, H. M.; Park, B. S.; Kim, J.-I.; Kim, S. E.; Lee, J.; Oh, S. C.; Enkhbayar, P.; Matsushima, N.; Lee, H.; Yoo, O. J.; Lee, J.-O. Crystal Structure of the TLR4-MD-2 Complex with Bound Endotoxin Antagonist Eritoran. *Cell* **2007**, *130* (5), 906–917. <https://doi.org/10.1016/j.cell.2007.08.002>.
- (47) d’Hennezel, E.; Abubucker, S.; Murphy, L. O.; Cullen, T. W. Total Lipopolysaccharide from the Human Gut Microbiome Silences Toll-Like Receptor Signaling. *mSystems* **2017**, *2* (6). <https://doi.org/10.1128/mSystems.00046-17>.
- (48) Vatanen, T.; Kostic, A. D.; d’Hennezel, E.; Siljander, H.; Franzosa, E. A.; Yassour, M.; Kolde, R.; Vlamakis, H.; Arthur, T. D.; Hämäläinen, A.-M.; Peet, A.; Tillmann, V.; Uibo, R.; Mokurov, S.; Dorshakova, N.; Ilonen, J.; Virtanen, S. M.; Szabo, S. J.; Porter, J. A.; Lähdesmäki, H.; Huttenhower, C.; Gevers, D.; Cullen, T. W.; Knip, M.; Xavier, R. J. Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans. *Cell* **2016**, *165* (4), 842–853. <https://doi.org/10.1016/j.cell.2016.04.007>.
- (49) Curtis, M. A.; Percival, R. S.; Devine, D.; Darveau, R. P.; Coats, S. R.; Rangarajan, M.; Tarelli, E.; Marsh, P. D. Temperature-Dependent Modulation of *Porphyromonas Gingivalis* Lipid A Structure and Interaction with the Innate Host Defenses. *Infect Immun* **2011**, *79* (3), 1187–1193. <https://doi.org/10.1128/IAI.00900-10>.
- (50) Tan, Y.; Zanoni, I.; Cullen, T. W.; Goodman, A. L.; Kagan, J. C. Mechanisms of Toll-like Receptor 4 Endocytosis Reveal a Common Immune-Evasion Strategy Used by Pathogenic and Commensal Bacteria. *Immunity* **2015**, *43* (5), 909–922. <https://doi.org/10.1016/j.immuni.2015.10.008>.
- (51) Montminy, S. W.; Khan, N.; McGrath, S.; Walkowicz, M. J.; Sharp, F.; Conlon, J. E.; Fukase, K.; Kusumoto, S.; Sweet, C.; Miyake, K.; Akira, S.; Cotter, R. J.; Goguen, J. D.; Lien, E. Virulence Factors of *Yersinia Pestis* Are Overcome by a Strong Lipopolysaccharide Response. *Nat Immunol* **2006**, *7* (10), 1066–1073. <https://doi.org/10.1038/ni1386>.
- (52) Coats, S. R.; Jones, J. W.; Do, C. T.; Braham, P. H.; Bainbridge, B. W.; To, T. T.; Goodlett, D. R.; Ernst, R. K.; Darveau, R. P. Human Toll-like Receptor 4 Responses to *P. Gingivalis* Are Regulated by Lipid A 1- and 4'-Phosphatase Activities. *Cellular Microbiology* **2009**, *11* (11), 1587–1599. <https://doi.org/10.1111/j.1462-5822.2009.01349.x>.
- (53) Rangarajan, M.; Aduse-Opoku, J.; Paramonov, N.; Hashim, A.; Bostanci, N.; Fraser, O. P.; Tarelli, E.; Curtis, M. A. Identification of a Second Lipopolysaccharide in *Porphyromonas Gingivalis* W50. *J Bacteriol* **2008**, *190* (8), 2920–2932. <https://doi.org/10.1128/JB.01868-07>.
- (54) Moran, A. P.; Lindner, B.; Walsh, E. J. Structural Characterization of the Lipid A Component of *Helicobacter Pylori* Rough- and Smooth-Form Lipopolysaccharides. *J Bacteriol* **1997**, *179* (20), 6453–6463. <https://doi.org/10.1128/jb.179.20.6453-6463.1997>.

- (55) Guo, L.; Lim, K. B.; Gunn, J. S.; Bainbridge, B.; Darveau, R. P.; Hackett, M.; Miller, S. I. Regulation of Lipid A Modifications by *Salmonella Typhimurium* Virulence Genes *phoP-phoQ*. *Science* **1997**, 276 (5310), 250–253. <https://doi.org/10.1126/science.276.5310.250>.
- (56) Paciello, I.; Silipo, A.; Lembo-Fazio, L.; Curcurù, L.; Zumsteg, A.; Noël, G.; Ciancarella, V.; Sturiale, L.; Molinaro, A.; Bernardini, M. L. Intracellular *Shigella* Remodels Its LPS to Dampen the Innate Immune Recognition and Evade Inflammasome Activation. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, 110 (46). <https://doi.org/10.1073/pnas.1303641110>.
- (57) Chen, F.; Zou, L.; Williams, B.; Chao, W. Targeting Toll-Like Receptors in Sepsis: From Bench to Clinical Trials. *Antioxidants & Redox Signaling* **2021**, 35 (15), 1324–1339. <https://doi.org/10.1089/ars.2021.0005>.
- (58) Opal, S. M.; Laterre, P.-F.; Francois, B.; LaRosa, S. P.; Angus, D. C.; Mira, J.-P.; Wittebole, X.; Dugernier, T.; Perrotin, D.; Tidswell, M.; Jauregui, L.; Krell, K.; Pachel, J.; Takahashi, T.; Peckelsen, C.; Cordasco, E.; Chang, C.-S.; Oeyen, S.; Aikawa, N.; Maruyama, T.; Schein, R.; Kalil, A. C.; Van Nuffelen, M.; Lynn, M.; Rossignol, D. P.; Gogate, J.; Roberts, M. B.; Wheeler, J. L.; Vincent, J.-L.; Access Study Group, F. T. Effect of Eritoran, an Antagonist of MD2-TLR4, on Mortality in Patients With Severe Sepsis: The ACCESS Randomized Trial. *JAMA* **2013**, 309 (11), 1154. <https://doi.org/10.1001/jama.2013.2194>.
- (59) Amarante-Mendes, G. P.; Adjemian, S.; Branco, L. M.; Zanetti, L. C.; Weinlich, R.; Bortoluci, K. R. Pattern Recognition Receptors and the Host Cell Death Molecular Machinery. *Front. Immunol.* **2018**, 9, 2379. <https://doi.org/10.3389/fimmu.2018.02379>.
- (60) Foell, D.; Wittkowski, H.; Vogl, T.; Roth, J. S100 Proteins Expressed in Phagocytes: A Novel Group of Damage-Associated Molecular Pattern Molecules. *Journal of Leukocyte Biology* **2007**, 81 (1), 28–37. <https://doi.org/10.1189/jlb.0306170>.
- (61) Björk, P.; Björk, A.; Vogl, T.; Stenström, M.; Liberg, D.; Olsson, A.; Roth, J.; Ivars, F.; Leanderson, T. Identification of Human S100A9 as a Novel Target for Treatment of Autoimmune Disease via Binding to Quinoline-3-Carboxamides. *PLoS Biol* **2009**, 7 (4), e1000097. <https://doi.org/10.1371/journal.pbio.1000097>.
- (62) Edgeworth, J.; Gorman, M.; Bennett, R.; Freemont, P.; Hogg, N. Identification of P8,14 as a Highly Abundant Heterodimeric Calcium Binding Protein Complex of Myeloid Cells. *Journal of Biological Chemistry* **1991**, 266 (12), 7706–7713. [https://doi.org/10.1016/S0021-9258\(20\)89506-4](https://doi.org/10.1016/S0021-9258(20)89506-4).
- (63) Vog, T.; Roth, J.; Sorg, C.; Hillenkamp, F.; Strupat, K. Calcium-Induced Noncovalently Linked Tetramers of MRP8 and MRP14 Detected by Ultraviolet Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **1999**, 10 (11), 1124–1130. [https://doi.org/10.1016/S1044-0305\(99\)00085-9](https://doi.org/10.1016/S1044-0305(99)00085-9).
- (64) Strupat, K.; Rogniaux, H.; Van Dorsselaer, A.; Roth, J.; Vogl, T. Calcium-Induced Noncovalently Linked Tetramers of MRP8 and MRP14 Are Confirmed by Electrospray

- Ionization-Mass Analysis. *J. Am. Soc. Mass Spectrom.* **2000**, *11* (9), 780–788.
[https://doi.org/10.1016/S1044-0305\(00\)00150-1](https://doi.org/10.1016/S1044-0305(00)00150-1).
- (65) Roth, J.; Burwinkel, F.; Van Den Bos, C.; Goebeler, M.; Vollmer, E.; Sorg, C. MRP8 and MRP14, S-100-like Proteins Associated with Myeloid Differentiation, Are Translocated to Plasma Membrane and Intermediate Filaments in a Calcium-Dependent Manner. *Blood* **1993**, *82* (6), 1875–1883. <https://doi.org/10.1182/blood.V82.6.1875.1875>.
- (66) Goebeler, M.; Roth, J.; Van Den Bos, C.; Ader, G.; Sorg, C. Increase of Calcium Levels in Epithelial Cells Induces Translocation of Calcium-Binding Proteins Migration Inhibitory Factor-Related Protein 8 (MRP8) and MRP14 to Keratin Intermediate Filaments. *Biochemical Journal* **1995**, *309* (2), 419–424. <https://doi.org/10.1042/bj3090419>.
- (67) Kerkhoff, C.; Klempt, M.; Kaefer, V.; Sorg, C. The Two Calcium-Binding Proteins, S100A8 and S100A9, Are Involved in the Metabolism of Arachidonic Acid in Human Neutrophils. *Journal of Biological Chemistry* **1999**, *274* (46), 32672–32679. <https://doi.org/10.1074/jbc.274.46.32672>.
- (68) Vogl, T.; Ludwig, S.; Goebeler, M.; Strey, A.; Thorey, I. S.; Reichelt, R.; Foell, D.; Gerke, V.; Manitz, M. P.; Nacken, W.; Werner, S.; Sorg, C.; Roth, J. MRP8 and MRP14 Control Microtubule Reorganization during Transendothelial Migration of Phagocytes. *Blood* **2004**, *104* (13), 4260–4268. <https://doi.org/10.1182/blood-2004-02-0446>.
- (69) Lackmann, M.; Cornish, C. J.; Simpson, R. J.; Moritz, R. L.; Geczy, C. L. Purification and Structural Analysis of a Murine Chemotactic Cytokine (CP-10) with Sequence Homology to S100 Proteins. *Journal of Biological Chemistry* **1992**, *267* (11), 7499–7504. [https://doi.org/10.1016/S0021-9258\(18\)42545-8](https://doi.org/10.1016/S0021-9258(18)42545-8).
- (70) Passey, R. J.; Williams, E.; Lichanska, A. M.; Wells, C.; Hu, S.; Geczy, C. L.; Little, M. H.; Hume, D. A. A Null Mutation in the Inflammation-Associated S100 Protein S100A8 Causes Early Resorption of the Mouse Embryo. *The Journal of Immunology* **1999**, *163* (4), 2209–2216. <https://doi.org/10.4049/jimmunol.163.4.2209>.
- (71) Ryckman, C.; Vandal, K.; Rouleau, P.; Talbot, M.; Tessier, P. A. Proinflammatory Activities of S100: Proteins S100A8, S100A9, and S100A8/A9 Induce Neutrophil Chemotaxis and Adhesion. *The Journal of Immunology* **2003**, *170* (6), 3233–3242. <https://doi.org/10.4049/jimmunol.170.6.3233>.
- (72) Chen, B.; Miller, A. L.; Rebelatto, M.; Brewah, Y.; Rowe, D. C.; Clarke, L.; Czapiga, M.; Rosenthal, K.; Imamichi, T.; Chen, Y.; Chang, C.-S.; Chowdhury, P. S.; Naiman, B.; Wang, Y.; Yang, D.; Humbles, A. A.; Herbst, R.; Sims, G. P. S100A9 Induced Inflammatory Responses Are Mediated by Distinct Damage Associated Molecular Patterns (DAMP) Receptors In Vitro and In Vivo. *PLoS ONE* **2015**, *10* (2), e0115828. <https://doi.org/10.1371/journal.pone.0115828>.
- (73) Riva, M.; Källberg, E.; Björk, P.; Hancz, D.; Vogl, T.; Roth, J.; Ivars, F.; Leanderson, T. Induction of Nuclear Factor- κ B Responses by the S 100 A 9 Protein Is Toll-like

- Receptor-4-dependent. *Immunology* **2012**, *137* (2), 172–182. <https://doi.org/10.1111/j.1365-2567.2012.03619.x>.
- (74) Ehrchen, J. M.; Sunderkötter, C.; Foell, D.; Vogl, T.; Roth, J. The Endogenous Toll-like Receptor 4 Agonist S100A8/S100A9 (Calprotectin) as Innate Amplifier of Infection, Autoimmunity, and Cancer. *Journal of Leukocyte Biology* **2009**, *86* (3), 557–566. <https://doi.org/10.1189/jlb.1008647>.
- (75) Nacken, W.; Kerkhoff, C. The Hetero-oligomeric Complex of the S100A8/S100A9 Protein Is Extremely Protease Resistant. *FEBS Letters* **2007**, *581* (26), 5127–5130. <https://doi.org/10.1016/j.febslet.2007.09.060>.
- (76) Harman, J. L.; Loes, A. N.; Warren, G. D.; Heaphy, M. C.; Lampi, K. J.; Harms, M. J. Evolution of Multifunctionality through a Pleiotropic Substitution in the Innate Immune Protein S100A9. *eLife* **2020**, *9*, e54100. <https://doi.org/10.7554/eLife.54100>.
- (77) Kessenbrock, K.; Dau, T.; Jenne, D. E. Tailor-Made Inflammation: How Neutrophil Serine Proteases Modulate the Inflammatory Response. *J Mol Med* **2011**, *89* (1), 23–28. <https://doi.org/10.1007/s00109-010-0677-3>.
- (78) Heutinck, K. M.; Ten Berge, I. J. M.; Hack, C. E.; Hamann, J.; Rowshani, A. T. Serine Proteases of the Human Immune System in Health and Disease. *Molecular Immunology* **2010**, *47* (11–12), 1943–1955. <https://doi.org/10.1016/j.molimm.2010.04.020>.
- (79) Janoff, A. Neutrophil Proteases in Inflammation. *Annu. Rev. Med.* **1972**, *23* (1), 177–190. <https://doi.org/10.1146/annurev.me.23.020172.001141>.
- (80) Jerke, U.; Hernandez, D. P.; Beaudette, P.; Korkmaz, B.; Dittmar, G.; Kettritz, R. Neutrophil Serine Proteases Exert Proteolytic Activity on Endothelial Cells. *Kidney International* **2015**, *88* (4), 764–775. <https://doi.org/10.1038/ki.2015.159>.
- (81) Vogl, T.; Stratis, A.; Wixler, V.; Völler, T.; Thurainayagam, S.; Jorch, S. K.; Zenker, S.; Dreiling, A.; Chakraborty, D.; Fröhling, M.; Paruzel, P.; Wehmeyer, C.; Hermann, S.; Papantonopoulou, O.; Geyer, C.; Loser, K.; Schäfers, M.; Ludwig, S.; Stoll, M.; Leanderson, T.; Schultze, J. L.; König, S.; Pap, T.; Roth, J. Autoinhibitory Regulation of S100A8/S100A9 Alarmin Activity Locally Restricts Sterile Inflammation. *Journal of Clinical Investigation* **2018**, *128* (5), 1852–1866. <https://doi.org/10.1172/JCI89867>.
- (82) Stephan, J. R.; Nolan, E. M. Calcium-Induced Tetramerization and Zinc Chelation Shield Human Calprotectin from Degradation by Host and Bacterial Extracellular Proteases. *Chem. Sci.* **2016**, *7* (3), 1962–1975. <https://doi.org/10.1039/C5SC03287C>.
- (83) Steinbakk, M.; Naess-Andresen, C.-F.; Fagerhol, M. K.; Lingaas, E.; Dale, I.; Brandtzaeg, P. Antimicrobial Actions of Calcium Binding Leucocyte L1 Protein, Calprotectin. *The Lancet* **1990**, *336* (8718), 763–765. [https://doi.org/10.1016/0140-6736\(90\)93237-J](https://doi.org/10.1016/0140-6736(90)93237-J).
- (84) Corbin, B. D.; Seeley, E. H.; Raab, A.; Feldmann, J.; Miller, M. R.; Torres, V. J.; Anderson, K. L.; Dattilo, B. M.; Dunman, P. M.; Gerads, R.; Caprioli, R. M.; Nacken, W.;

- Chazin, W. J.; Skaar, E. P. Metal Chelation and Inhibition of Bacterial Growth in Tissue Abscesses. *Science* **2008**, *319* (5865), 962–965. <https://doi.org/10.1126/science.1152449>.
- (85) Kehl-Fie, T. E.; Chitayat, S.; Hood, M. I.; Damo, S.; Restrepo, N.; Garcia, C.; Munro, K. A.; Chazin, W. J.; Skaar, E. P. Nutrient Metal Sequestration by Calprotectin Inhibits Bacterial Superoxide Defense, Enhancing Neutrophil Killing of *Staphylococcus Aureus*. *Cell Host & Microbe* **2011**, *10* (2), 158–164. <https://doi.org/10.1016/j.chom.2011.07.004>.
- (86) Brophy, M. B.; Hayden, J. A.; Nolan, E. M. Calcium Ion Gradients Modulate the Zinc Affinity and Antibacterial Activity of Human Calprotectin. *J. Am. Chem. Soc.* **2012**, *134* (43), 18089–18100. <https://doi.org/10.1021/ja307974e>.
- (87) Damo, S. M.; Kehl-Fie, T. E.; Sugitani, N.; Holt, M. E.; Rathi, S.; Murphy, W. J.; Zhang, Y.; Betz, C.; Hench, L.; Fritz, G.; Skaar, E. P.; Chazin, W. J. Molecular Basis for Manganese Sequestration by Calprotectin and Roles in the Innate Immune Response to Invading Bacterial Pathogens. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110* (10), 3841–3846. <https://doi.org/10.1073/pnas.1220341110>.
- (88) Nakashige, T. G.; Zhang, B.; Krebs, C.; Nolan, E. M. Human Calprotectin Is an Iron-Sequestering Host-Defense Protein. *Nat Chem Biol* **2015**, *11* (10), 765–771. <https://doi.org/10.1038/nchembio.1891>.
- (89) Zygiel, E. M.; Nolan, E. M. Transition Metal Sequestration by the Host-Defense Protein Calprotectin. *Annu. Rev. Biochem.* **2018**, *87* (1), 621–643. <https://doi.org/10.1146/annurev-biochem-062917-012312>.
- (90) Rosen, T.; Nolan, E. M. Metal Sequestration and Antimicrobial Activity of Human Calprotectin Are pH-Dependent. *Biochemistry* **2020**, *59* (26), 2468–2478. <https://doi.org/10.1021/acs.biochem.0c00359>.
- (91) Carnazzo, V.; Redi, S.; Basile, V.; Natali, P.; Gulli, F.; Equitani, F.; Marino, M.; Basile, U. Calprotectin: Two Sides of the Same Coin. *Rheumatology* **2024**, *63* (1), 26–33. <https://doi.org/10.1093/rheumatology/kead405>.
- (92) Romand, X.; Bernardy, C.; Nguyen, M. V. C.; Courtier, A.; Trocme, C.; Clapasson, M.; Paclet, M.-H.; Toussaint, B.; Gaudin, P.; Baillet, A. Systemic Calprotectin and Chronic Inflammatory Rheumatic Diseases. *Joint Bone Spine* **2019**, *86* (6), 691–698. <https://doi.org/10.1016/j.jbspin.2019.01.003>.
- (93) Shabani, F.; Farasat, A.; Mahdavi, M.; Gheibi, N. Calprotectin (S100A8/S100A9): A Key Protein between Inflammation and Cancer. *Inflamm. Res.* **2018**, *67* (10), 801–812. <https://doi.org/10.1007/s00011-018-1173-4>.
- (94) Ometto, F.; Friso, L.; Astorri, D.; Botsios, C.; Raffener, B.; Punzi, L.; Doria, A. Calprotectin in Rheumatic Diseases. *Exp Biol Med (Maywood)* **2017**, *242* (8), 859–873. <https://doi.org/10.1177/1535370216681551>.

- (95) Lam, S. H.; Chua, H. L.; Gong, Z.; Lam, T. J.; Sin, Y. M. Development and Maturation of the Immune System in Zebrafish, *Danio Rerio*: A Gene Expression Profiling, in Situ Hybridization and Immunological Study. *Developmental & Comparative Immunology* **2004**, *28* (1), 9–28. [https://doi.org/10.1016/S0145-305X\(03\)00103-4](https://doi.org/10.1016/S0145-305X(03)00103-4).
- (96) Melancon, E.; Gomez De La Torre Canny, S.; Sichel, S.; Kelly, M.; Wiles, T. J.; Rawls, J. F.; Eisen, J. S.; Guillemin, K. Best Practices for Germ-Free Derivation and Gnotobiotic Zebrafish Husbandry. In *Methods in Cell Biology*; Elsevier, 2017; Vol. 138, pp 61–100. <https://doi.org/10.1016/bs.mcb.2016.11.005>.
- (97) William Detrich, H.; Westerfield, M.; Zon, L. I. Preface. In *Methods in Cell Biology*; Elsevier, 2017; Vol. 138, pp xxiii–xxiv. [https://doi.org/10.1016/S0091-679X\(17\)30010-9](https://doi.org/10.1016/S0091-679X(17)30010-9).
- (98) Meeker, N. D.; Trede, N. S. Immunology and Zebrafish: Spawning New Models of Human Disease. *Developmental & Comparative Immunology* **2008**, *32* (7), 745–757. <https://doi.org/10.1016/j.dci.2007.11.011>.
- (99) Kumar, S.; Suleski, M.; Craig, J. M.; Kasprowitz, A. E.; Sanderford, M.; Li, M.; Stecher, G.; Hedges, S. B. TimeTree 5: An Expanded Resource for Species Divergence Times. *Molecular Biology and Evolution* **2022**, *39* (8), msac174. <https://doi.org/10.1093/molbev/msac174>.
- (100) Howe, K.; Clark, M. D.; Torroja, C. F.; Torrance, J.; Berthelot, C.; Muffato, M.; Collins, J. E.; Humphray, S.; McLaren, K.; Matthews, L.; McLaren, S.; Sealy, I.; Caccamo, M.; Churcher, C.; Scott, C.; Barrett, J. C.; Koch, R.; Rauch, G.-J.; White, S.; Chow, W.; Kilian, B.; Quintais, L. T.; Guerra-Assunção, J. A.; Zhou, Y.; Gu, Y.; Yen, J.; Vogel, J.-H.; Eyre, T.; Redmond, S.; Banerjee, R.; Chi, J.; Fu, B.; Langley, E.; Maguire, S. F.; Laird, G. K.; Lloyd, D.; Kenyon, E.; Donaldson, S.; Sehra, H.; Almeida-King, J.; Loveland, J.; Trevanion, S.; Jones, M.; Quail, M.; Willey, D.; Hunt, A.; Burton, J.; Sims, S.; McLay, K.; Plumb, B.; Davis, J.; Clee, C.; Oliver, K.; Clark, R.; Riddle, C.; Elliott, D.; Threadgold, G.; Harden, G.; Ware, D.; Begum, S.; Mortimore, B.; Kerry, G.; Heath, P.; Phillimore, B.; Tracey, A.; Corby, N.; Dunn, M.; Johnson, C.; Wood, J.; Clark, S.; Pelan, S.; Griffiths, G.; Smith, M.; Glithero, R.; Howden, P.; Barker, N.; Lloyd, C.; Stevens, C.; Harley, J.; Holt, K.; Panagiotidis, G.; Lovell, J.; Beasley, H.; Henderson, C.; Gordon, D.; Auger, K.; Wright, D.; Collins, J.; Raisen, C.; Dyer, L.; Leung, K.; Robertson, L.; Ambridge, K.; Leongamornlert, D.; McGuire, S.; Gilderthorp, R.; Griffiths, C.; Manthravadi, D.; Nichol, S.; Barker, G.; Whitehead, S.; Kay, M.; Brown, J.; Murnane, C.; Gray, E.; Humphries, M.; Sycamore, N.; Barker, D.; Saunders, D.; Wallis, J.; Babbage, A.; Hammond, S.; Mashreghi-Mohammadi, M.; Barr, L.; Martin, S.; Wray, P.; Ellington, A.; Matthews, N.; Ellwood, M.; Woodmansey, R.; Clark, G.; Cooper, J. D.; Tromans, A.; Grafham, D.; Skuce, C.; Pandian, R.; Andrews, R.; Harrison, E.; Kimberley, A.; Garnett, J.; Fosker, N.; Hall, R.; Garner, P.; Kelly, D.; Bird, C.; Palmer, S.; Gehring, I.; Berger, A.; Dooley, C. M.; Ersan-Ürün, Z.; Eser, C.; Geiger, H.; Geisler, M.; Karotki, L.; Kirn, A.; Konantz, J.; Konantz, M.; Oberländer, M.; Rudolph-Geiger, S.; Teucke, M.; Lanz, C.; Raddatz, G.; Osoegawa, K.; Zhu, B.; Rapp, A.; Widaa, S.; Langford, C.; Yang, F.; Schuster, S. C.; Carter, N. P.; Harrow, J.; Ning, Z.; Herrero, J.; Searle, S. M. J.; Enright, A.; Geisler, R.; Plasterk, R. H. A.; Lee, C.; Westerfield, M.; De Jong, P. J.; Zon, L. I.;

- Postlethwait, J. H.; Nüsslein-Volhard, C.; Hubbard, T. J. P.; Crollius, H. R.; Rogers, J.; Stemple, D. L. The Zebrafish Reference Genome Sequence and Its Relationship to the Human Genome. *Nature* **2013**, *496* (7446), 498–503. <https://doi.org/10.1038/nature12111>.
- (101) Loes, A. N.; Hinman, M. N.; Farnsworth, D. R.; Miller, A. C.; Guillemin, K.; Harms, M. J. Identification and Characterization of Zebrafish Tlr4 Coreceptor Md-2. *The Journal of Immunology* **2021**, *206* (5), 1046–1057. <https://doi.org/10.4049/jimmunol.1901288>.
- (102) Loes, A. N.; Bridgham, J. T.; Harms, M. J. Coevolution of the Toll-Like Receptor 4 Complex with Calgranulins and Lipopolysaccharide. *Front Immunol* **2018**, *9*, 304. <https://doi.org/10.3389/fimmu.2018.00304>.
- (103) Yang, L.-L.; Wang, G.-Q.; Yang, L.-M.; Huang, Z.-B.; Zhang, W.-Q.; Yu, L.-Z. Endotoxin Molecule Lipopolysaccharide-Induced Zebrafish Inflammation Model: A Novel Screening Method for Anti-Inflammatory Drugs. *Molecules* **2014**, *19* (2), 2390–2409. <https://doi.org/10.3390/molecules19022390>.
- (104) Watzke, J.; Schirmer, K.; Scholz, S. Bacterial Lipopolysaccharides Induce Genes Involved in the Innate Immune Response in Embryos of the Zebrafish (*Danio Rerio*). *Fish & Shellfish Immunology* **2007**, *23* (4), 901–905. <https://doi.org/10.1016/j.fsi.2007.03.004>.
- (105) Novoa, B.; Bowman, T. V.; Zon, L.; Figueras, A. LPS Response and Tolerance in the Zebrafish (*Danio Rerio*). *Fish & Shellfish Immunology* **2009**, *26* (2), 326–331. <https://doi.org/10.1016/j.fsi.2008.12.004>.
- (106) Harms, M. J.; Thornton, J. W. Analyzing Protein Structure and Function Using Ancestral Gene Reconstruction. *Current Opinion in Structural Biology* **2010**, *20* (3), 360–366. <https://doi.org/10.1016/j.sbi.2010.03.005>.
- (107) Farr, D.; Nag, D.; Chazin, W. J.; Harrison, S.; Thummel, R.; Luo, X.; Raychaudhuri, S.; Withey, J. H. Neutrophil-Associated Responses to *Vibrio Cholerae* Infection in a Natural Host Model. *Infect Immun* **2022**, *90* (3), e00466-21. <https://doi.org/10.1128/iai.00466-21>.
- (108) Nag, D.; Farr, D.; Raychaudhuri, S.; Withey, J. H. An Adult Zebrafish Model for Adherent-Invasive *Escherichia Coli* Indicates Protection from AIEC Infection by Probiotic *E. Coli* Nissle. *iScience* **2022**, *25* (7), 104572. <https://doi.org/10.1016/j.isci.2022.104572>.
- (109) Janeway, C. A.; Medzhitov, R. Innate Immune Recognition. *Annu. Rev. Immunol.* **2002**, *20* (1), 197–216. <https://doi.org/10.1146/annurev.immunol.20.083001.084359>.
- (110) Pauling, L.; Zuckerkandl, E.; Henriksen, T.; Löfstad, R. Chemical Paleogenetics. Molecular “Restoration Studies” of Extinct Forms of Life. *Acta Chem. Scand.* **1963**, *17* *supl.*, 9–16. <https://doi.org/10.3891/acta.chem.scand.17s-0009>.
- (111) Spence, M. A.; Kaczmarek, J. A.; Saunders, J. W.; Jackson, C. J. Ancestral Sequence Reconstruction for Protein Engineers. *Current Opinion in Structural Biology* **2021**, *69*, 131–141. <https://doi.org/10.1016/j.sbi.2021.04.001>.

- (112) Nicoll, C. R.; Bailleul, G.; Fiorentini, F.; Mascotti, M. L.; Fraaije, M. W.; Mattevi, A. Ancestral-Sequence Reconstruction Unveils the Structural Basis of Function in Mammalian FMOs. *Nat Struct Mol Biol* **2020**, *27* (1), 14–24. <https://doi.org/10.1038/s41594-019-0347-2>.
- (113) Furukawa, R.; Toma, W.; Yamazaki, K.; Akanuma, S. Ancestral Sequence Reconstruction Produces Thermally Stable Enzymes with Mesophilic Enzyme-like Catalytic Properties. *Sci Rep* **2020**, *10* (1), 15493. <https://doi.org/10.1038/s41598-020-72418-4>.
- (114) Zakas, P. M.; Brown, H. C.; Knight, K.; Meeks, S. L.; Spencer, H. T.; Gaucher, E. A.; Doering, C. B. Enhancing the Pharmaceutical Properties of Protein Drugs by Ancestral Sequence Reconstruction. *Nat Biotechnol* **2017**, *35* (1), 35–37. <https://doi.org/10.1038/nbt.3677>.
- (115) Anderson, D. P.; Whitney, D. S.; Hanson-Smith, V.; Woznica, A.; Campodonico-Burnett, W.; Volkman, B. F.; King, N.; Thornton, J. W.; Prehoda, K. E. Evolution of an Ancient Protein Function Involved in Organized Multicellularity in Animals. *eLife* **2016**, *5*, e10147. <https://doi.org/10.7554/eLife.10147>.
- (116) Diez-Hermano, S.; Ganfornina, M. D.; Skerra, A.; Gutiérrez, G.; Sanchez, D. An Evolutionary Perspective of the Lipocalin Protein Family. *Front. Physiol.* **2021**, *12*, 718983. <https://doi.org/10.3389/fphys.2021.718983>.
- (117) Mascotti, M. L. Resurrecting Enzymes by Ancestral Sequence Reconstruction. In *Enzyme Engineering*; Magnani, F., Marabelli, C., Paradisi, F., Eds.; Methods in Molecular Biology; Springer US: New York, NY, 2022; Vol. 2397, pp 111–136. https://doi.org/10.1007/978-1-0716-1826-4_7.
- (118) Merkl, R.; Sterner, R. Ancestral Protein Reconstruction: Techniques and Applications. *Biological Chemistry* **2016**, *397* (1), 1–21. <https://doi.org/10.1515/hsz-2015-0158>.
- (119) Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* **2007**, *24* (8), 1586–1591. <https://doi.org/10.1093/molbev/msm088>.
- (120) Rees, J.; Cranston, K. Automated Assembly of a Reference Taxonomy for Phylogenetic Data Synthesis. *BDJ* **2017**, *5*, e12581. <https://doi.org/10.3897/BDJ.5.e12581>.
- (121) Vialle, R. A.; Tamuri, A. U.; Goldman, N. Alignment Modulates Ancestral Sequence Reconstruction Accuracy. *Molecular Biology and Evolution* **2018**, *35* (7), 1783–1797. <https://doi.org/10.1093/molbev/msy055>.
- (122) Tan, G.; Muffato, M.; Ledergerber, C.; Herrero, J.; Goldman, N.; Gil, M.; Dessimoz, C. Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference. *Syst Biol* **2015**, *64* (5), 778–791. <https://doi.org/10.1093/sysbio/syv033>.

- (123) Tumescheit, C.; Firth, A. E.; Brown, K. CIAlign: A Highly Customisable Command Line Tool to Clean, Interpret and Visualise Multiple Sequence Alignments. *PeerJ* **2022**, *10*, e12983. <https://doi.org/10.7717/peerj.12983>.
- (124) Catanach, T. A.; Sweet, A. D.; Nguyen, N. D.; Peery, R. M.; Debevec, A. H.; Thomer, A. K.; Owings, A. C.; Boyd, B. M.; Katz, A. D.; Soto-Adames, F. N.; Allen, J. M. Fully Automated Sequence Alignment Methods Are Comparable to, and Much Faster than, Traditional Methods in Large Data Sets: An Example with Hepatitis B Virus. *PeerJ* **2019**, *7*, e6142. <https://doi.org/10.7717/peerj.6142>.
- (125) Morrison, D. A. Multiple Sequence Alignment for Phylogenetic Purposes. *Aust. Systematic Bot.* **2006**, *19* (6), 479. <https://doi.org/10.1071/SB06020>.
- (126) Del Amparo, R.; Arenas, M. Consequences of Substitution Model Selection on Protein Ancestral Sequence Reconstruction. *Molecular Biology and Evolution* **2022**, *39* (7), msac144. <https://doi.org/10.1093/molbev/msac144>.
- (127) Joy, J. B.; Liang, R. H.; McCloskey, R. M.; Nguyen, T.; Poon, A. F. Y. Ancestral Reconstruction. *PLoS Comput Biol* **2016**, *12* (7), e1004763. <https://doi.org/10.1371/journal.pcbi.1004763>.
- (128) Morel, B.; Kozlov, A. M.; Stamatakis, A.; Szöllösi, G. J. GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss. *Molecular Biology and Evolution* **2020**, *37* (9), 2763–2774. <https://doi.org/10.1093/molbev/msaa141>.
- (129) Groussin, M.; Hobbs, J. K.; Szöllösi, G. J.; Gribaldo, S.; Arcus, V. L.; Gouy, M. Toward More Accurate Ancestral Protein Genotype–Phenotype Reconstructions with the Use of Species Tree-Aware Gene Trees. *Molecular Biology and Evolution* **2015**, *32* (1), 13–22. <https://doi.org/10.1093/molbev/msu305>.
- (130) Gogarten, J. P.; Doolittle, W. F.; Lawrence, J. G. Prokaryotic Evolution in Light of Gene Transfer. *Molecular Biology and Evolution* **2002**, *19* (12), 2226–2238. <https://doi.org/10.1093/oxfordjournals.molbev.a004046>.
- (131) Parks, D. H.; Chuvpochina, M.; Waite, D. W.; Rinke, C.; Skarszewski, A.; Chaumeil, P.-A.; Hugenholtz, P. A Standardized Bacterial Taxonomy Based on Genome Phylogeny Substantially Revises the Tree of Life. *Nat Biotechnol* **2018**, *36* (10), 996–1004. <https://doi.org/10.1038/nbt.4229>.
- (132) Kinene, T.; Wainaina, J.; Maina, S.; Boykin, L. M. Rooting Trees, Methods For. In *Encyclopedia of Evolutionary Biology*; Elsevier, 2016; pp 489–493. <https://doi.org/10.1016/B978-0-12-800049-6.00215-8>.
- (133) Yang, Z.; Kumar, S.; Nei, M. A New Method of Inference of Ancestral Nucleotide and Amino Acid Sequences. *Genetics* **1995**, *141* (4), 1641–1650. <https://doi.org/10.1093/genetics/141.4.1641>.

- (134) Eick, G. N.; Bridgham, J. T.; Anderson, D. P.; Harms, M. J.; Thornton, J. W. Robustness of Reconstructed Ancestral Protein Functions to Statistical Uncertainty. *Mol Biol Evol* **2016**, msw223. <https://doi.org/10.1093/molbev/msw223>.
- (135) Akanuma, S.; Nakajima, Y.; Yokobori, S.; Kimura, M.; Nemoto, N.; Mase, T.; Miyazono, K.; Tanokura, M.; Yamagishi, A. Experimental Evidence for the Thermophilicity of Ancestral Life. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110* (27), 11067–11072. <https://doi.org/10.1073/pnas.1308215110>.
- (136) Bridgham, J. T.; Keay, J.; Ortlund, E. A.; Thornton, J. W. Vestigialization of an Allosteric Switch: Genetic and Structural Mechanisms for the Evolution of Constitutive Activity in a Steroid Hormone Receptor. *PLoS Genet* **2014**, *10* (1), e1004058. <https://doi.org/10.1371/journal.pgen.1004058>.
- (137) McKeown, A. N.; Bridgham, J. T.; Anderson, D. W.; Murphy, M. N.; Ortlund, E. A.; Thornton, J. W. Evolution of DNA Specificity in a Transcription Factor Family Produced a New Gene Regulatory Module. *Cell* **2014**, *159* (1), 58–68. <https://doi.org/10.1016/j.cell.2014.09.003>.
- (138) Wheeler, L. C.; Anderson, J. A.; Morrison, A. J.; Wong, C. E.; Harms, M. J. Conservation of Specificity in Two Low-Specificity Proteins. *Biochemistry* **2018**, *57* (5), 684–695. <https://doi.org/10.1021/acs.biochem.7b01086>.
- (139) Edgar, R. C. Muscle5: High-Accuracy Alignment Ensembles Enable Unbiased Assessments of Sequence Homology and Phylogeny. *Nat Commun* **2022**, *13* (1), 6968. <https://doi.org/10.1038/s41467-022-34630-w>.
- (140) Kozlov, A. M.; Darriba, D.; Flouri, T.; Morel, B.; Stamatakis, A. RAxML-NG: A Fast, Scalable and User-Friendly Tool for Maximum Likelihood Phylogenetic Inference. *Bioinformatics* **2019**, *35* (21), 4453–4455. <https://doi.org/10.1093/bioinformatics/btz305>.
- (141) Ishikawa, S. A.; Zhukova, A.; Iwasaki, W.; Gascuel, O. A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios. *Molecular Biology and Evolution* **2019**, *36* (9), 2069–2085. <https://doi.org/10.1093/molbev/msz131>.
- (142) Huerta-Cepas, J.; Serra, F.; Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol Biol Evol* **2016**, *33* (6), 1635–1638. <https://doi.org/10.1093/molbev/msw046>.
- (143) Altschul, S. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Research* **1997**, *25* (17), 3389–3402. <https://doi.org/10.1093/nar/25.17.3389>.
- (144) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25* (11), 1422–1423. <https://doi.org/10.1093/bioinformatics/btp163>.

- (145) Mctavish, E. J.; Sánchez-Reyes, L. L.; Holder, M. T. OpenTree: A Python Package for Accessing and Analyzing Data from the Open Tree of Life. *Systematic Biology* **2021**, *70* (6), 1295–1301. <https://doi.org/10.1093/sysbio/syab033>.
- (146) Eaton, D. A. R. Toytree: A Minimalist Tree Visualization and Manipulation Library for Python. *Methods Ecol Evol* **2020**, *11* (1), 187–191. <https://doi.org/10.1111/2041-210X.13313>.
- (147) Frith, M. C. How Sequence Alignment Scores Correspond to Probability Models. *Bioinformatics* **2019**, btz576. <https://doi.org/10.1093/bioinformatics/btz576>.
- (148) Flouri, T.; Izquierdo-Carrasco, F.; Darriba, D.; Aberer, A. J.; Nguyen, L.-T.; Minh, B. Q.; Von Haeseler, A.; Stamatakis, A. The Phylogenetic Likelihood Library. *Systematic Biology* **2015**, *64* (2), 356–362. <https://doi.org/10.1093/sysbio/syu084>.
- (149) Felsenstein, J. CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP. *Evolution* **1985**, *39* (4), 783–791. <https://doi.org/10.1111/j.1558-5646.1985.tb00420.x>.
- (150) Abascal, F.; Zardoya, R.; Posada, D. ProtTest: Selection of Best-Fit Models of Protein Evolution. *Bioinformatics* **2005**, *21* (9), 2104–2105. <https://doi.org/10.1093/bioinformatics/bti263>.
- (151) Maddison, D. R.; Maddison, W. P. MacClade 4, 2000. <http://ib.berkeley.edu/courses/ib200/readings/MacClade%204%20Manual.pdf>.
- (152) Larsson, A. AliView: A Fast and Lightweight Alignment Viewer and Editor for Large Datasets. *Bioinformatics* **2014**, *30* (22), 3276–3278. <https://doi.org/10.1093/bioinformatics/btu531>.
- (153) Waterhouse, A. M.; Procter, J. B.; Martin, D. M. A.; Clamp, M.; Barton, G. J. Jalview Version 2--a Multiple Sequence Alignment Editor and Analysis Workbench. *Bioinformatics* **2009**, *25* (9), 1189–1191. <https://doi.org/10.1093/bioinformatics/btp033>.
- (154) Tamura, K.; Stecher, G.; Kumar, S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution* **2021**, *38* (7), 3022–3027. <https://doi.org/10.1093/molbev/msab120>.
- (155) Zheng, Y.; Zhang, L. Effect of Incomplete Lineage Sorting On Tree-Reconciliation-Based Inference of Gene Duplication. *IEEE/ACM Trans. Comput. Biol. and Bioinf.* **2014**, *11* (3), 477–485. <https://doi.org/10.1109/TCBB.2013.2297913>.
- (156) Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data. *Bioinformatics* **2012**, *28* (23), 3150–3152. <https://doi.org/10.1093/bioinformatics/bts565>.
- (157) Colless, D. H.; Wiley, E. O. Phylogenetics: The Theory and Practice of Phylogenetic Systematics. *Systematic Zoology* **1982**, *31* (1), 100. <https://doi.org/10.2307/2413420>.

- (158) Sukumaran, J.; Holder, M. T. DendroPy: A Python Library for Phylogenetic Computing. *Bioinformatics* **2010**, *26* (12), 1569–1571. <https://doi.org/10.1093/bioinformatics/btq228>.
- (159) Pattengale, N. D.; Alipour, M.; Bininda-Emonds, O. R. P.; Moret, B. M. E.; Stamatakis, A. How Many Bootstrap Replicates Are Necessary? *Journal of Computational Biology* **2010**, *17* (3), 337–354. <https://doi.org/10.1089/cmb.2009.0179>.
- (160) Rudd, K. E.; Johnson, S. C.; Agesa, K. M.; Shackelford, K. A.; Tsoi, D.; Kievlan, D. R.; Colombara, D. V.; Ikuta, K. S.; Kissoon, N.; Finfer, S.; Fleischmann-Struzek, C.; Machado, F. R.; Reinhart, K. K.; Rowan, K.; Seymour, C. W.; Watson, R. S.; West, T. E.; Marinho, F.; Hay, S. I.; Lozano, R.; Lopez, A. D.; Angus, D. C.; Murray, C. J. L.; Naghavi, M. Global, Regional, and National Sepsis Incidence and Mortality, 1990–2017: Analysis for the Global Burden of Disease Study. *The Lancet* **2020**, *395* (10219), 200–211. [https://doi.org/10.1016/S0140-6736\(19\)32989-7](https://doi.org/10.1016/S0140-6736(19)32989-7).
- (161) Kim, H.-J.; Kim, H.; Lee, J.-H.; Hwangbo, C. Toll-like Receptor 4 (TLR4): New Insight Immune and Aging. *Immun Ageing* **2023**, *20* (1), 67. <https://doi.org/10.1186/s12979-023-00383-3>.
- (162) Yan, B.; Yu, X.; Cai, X.; Huang, X.; Xie, B.; Lian, D.; Chen, J.; Li, W.; Lin, Y.; Ye, J.; Li, J. A Review: The Significance of Toll-Like Receptors 2 and 4, and NF- κ B Signaling in Endothelial Cells during Atherosclerosis. *Front. Biosci. (Landmark Ed)* **2024**, *29* (4), 161. <https://doi.org/10.31083/j.fbl2904161>.
- (163) Nagai, Y.; Akashi, S.; Nagafuku, M.; Ogata, M.; Iwakura, Y.; Akira, S.; Kitamura, T.; Kosugi, A.; Kimoto, M.; Miyake, K. Essential Role of MD-2 in LPS Responsiveness and TLR4 Distribution. *Nat Immunol* **2002**, *3* (7), 667–672. <https://doi.org/10.1038/ni809>.
- (164) Viriyakosol, S.; Tobias, P. S.; Kitchens, R. L.; Kirkland, T. N. MD-2 Binds to Bacterial Lipopolysaccharide. *Journal of Biological Chemistry* **2001**, *276* (41), 38044–38051. <https://doi.org/10.1074/jbc.M105228200>.
- (165) Hailman, E.; Lichenstein, H. S.; Wurfel, M. M.; Miller, D. S.; Johnson, D. A.; Kelley, M.; Busse, L. A.; Zukowski, M. M.; Wright, S. D. Lipopolysaccharide (LPS)-Binding Protein Accelerates the Binding of LPS to CD14. *The Journal of experimental medicine* **1994**, *179* (1), 269–277. <https://doi.org/10.1084/jem.179.1.269>.
- (166) Gioannini, T. L.; Weiss, J. P. Regulation of Interactions of Gram-Negative Bacterial Endotoxins with Mammalian Cells. *Immunol Res* **2007**, *39* (1–3), 249–260. <https://doi.org/10.1007/s12026-007-0069-0>.
- (167) Prohinar, P.; Re, F.; Widstrom, R.; Zhang, D.; Teghanemt, A.; Weiss, J. P.; Gioannini, T. L. Specific High Affinity Interactions of Monomeric Endotoxin-Protein Complexes with Toll-like Receptor 4 Ectodomain. *Journal of Biological Chemistry* **2007**, *282* (2), 1010–1017. <https://doi.org/10.1074/jbc.M609400200>.
- (168) Ryu, J.-K.; Kim, S. J.; Rah, S.-H.; Kang, J. I.; Jung, H. E.; Lee, D.; Lee, H. K.; Lee, J.-O.; Park, B. S.; Yoon, T.-Y.; Kim, H. M. Reconstruction of LPS Transfer Cascade Reveals

Structural Determinants within LBP, CD14, and TLR4-MD2 for Efficient LPS Recognition and Transfer. *Immunity* **2017**, *46* (1), 38–50.
<https://doi.org/10.1016/j.immuni.2016.11.007>.

- (169) Haziot, A.; Ferrero, E.; Köntgen, F.; Hijiya, N.; Yamamoto, S.; Silver, J.; Stewart, C. L.; Goyert, S. M. Resistance to Endotoxin Shock and Reduced Dissemination of Gram-Negative Bacteria in CD14-Deficient Mice. *Immunity* **1996**, *4* (4), 407–414.
[https://doi.org/10.1016/S1074-7613\(00\)80254-X](https://doi.org/10.1016/S1074-7613(00)80254-X).
- (170) Tan, Y.; Kagan, J. C. A Cross-Disciplinary Perspective on the Innate Immune Responses to Bacterial Lipopolysaccharide. *Molecular Cell* **2014**, *54* (2), 212–223.
<https://doi.org/10.1016/j.molcel.2014.03.012>.
- (171) O’Neill, L. A. J.; Bowie, A. G. The Family of Five: TIR-Domain-Containing Adaptors in Toll-like Receptor Signalling. *Nat Rev Immunol* **2007**, *7* (5), 353–364.
<https://doi.org/10.1038/nri2079>.
- (172) Watters, T. M.; Kenny, E. F.; O’Neill, L. A. J. Structure, Function and Regulation of the Toll/IL-1 Receptor Adaptor Proteins. *Immunol Cell Biol* **2007**, *85* (6), 411–419.
<https://doi.org/10.1038/sj.icb.7100095>.
- (173) Ve, T.; J. Gay, N.; Mansell, A.; Kobe, B.; Kellie, S. Adaptors in Toll-Like Receptor Signaling and Their Potential as Therapeutic Targets. *CDT* **2012**, *13* (11), 1360–1374.
<https://doi.org/10.2174/138945012803530260>.
- (174) Mata-Haro, V.; Cekic, C.; Martin, M.; Chilton, P. M.; Casella, C. R.; Mitchell, T. C. The Vaccine Adjuvant Monophosphoryl Lipid A as a TRIF-Biased Agonist of TLR4. *Science* **2007**, *316* (5831), 1628–1632. <https://doi.org/10.1126/science.1138963>.
- (175) Li, Y.; Wang, Z.; Chen, J.; Ernst, R.; Wang, X. Influence of Lipid A Acylation Pattern on Membrane Permeability and Innate Immune Stimulation. *Marine Drugs* **2013**, *11* (9), 3197–3208. <https://doi.org/10.3390/md11093197>.
- (176) Needham, B. D.; Carroll, S. M.; Giles, D. K.; Georgiou, G.; Whiteley, M.; Trent, M. S. Modulating the Innate Immune Response by Combinatorial Engineering of Endotoxin. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110* (4), 1464–1469.
<https://doi.org/10.1073/pnas.1218080110>.
- (177) Rietschel, E. T.; Kirikae, T.; Schade, F. U.; Mamat, U.; Schmidt, G.; Loppnow, H.; Ulmer, A. J.; Zähringer, U.; Seydel, U.; Di Padova, F.; Schreier, M.; Brade, H. Bacterial Endotoxin: Molecular Relationships of Structure to Activity and Function. *FASEB j.* **1994**, *8* (2), 217–225. <https://doi.org/10.1096/fasebj.8.2.8119492>.
- (178) Scott, A. J.; Oyler, B. L.; Goodlett, D. R.; Ernst, R. K. Lipid A Structural Modifications in Extreme Conditions and Identification of Unique Modifying Enzymes to Define the Toll-like Receptor 4 Structure-Activity Relationship. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **2017**, *1862* (11), 1439–1450.
<https://doi.org/10.1016/j.bbalip.2017.01.004>.

- (179) Xie, Y.; Meijer, A. H.; Schaaf, M. J. M. Modeling Inflammation in Zebrafish for the Development of Anti-Inflammatory Drugs. *Front. Cell Dev. Biol.* **2021**, *8*, 620984. <https://doi.org/10.3389/fcell.2020.620984>.
- (180) Gauthier, A. E.; Chandler, C. E.; Poli, V.; Gardner, F. M.; Tekiau, A.; Smith, R.; Bonham, K. S.; Cordes, E. E.; Shank, T. M.; Zanoni, I.; Goodlett, D. R.; Biller, S. J.; Ernst, R. K.; Rotjan, R. D.; Kagan, J. C. Deep-Sea Microbes as Tools to Refine the Rules of Innate Immune Pattern Recognition. *Sci. Immunol.* **2021**, *6* (57), eabe0531. <https://doi.org/10.1126/sciimmunol.abe0531>.
- (181) Chilton, P. M.; Embry, C. A.; Mitchell, T. C. Effects of Differences in Lipid A Structure on TLR4 Pro-Inflammatory Signaling and Inflammasome Activation. *Front. Immun.* **2012**, *3*. <https://doi.org/10.3389/fimmu.2012.00154>.
- (182) Orlandi, K. N.; Phillips, S. R.; Sailer, Z. R.; Harman, J. L.; Harms, M. J. Topiary: Pruning the Manual Labor from Ancestral Sequence Reconstruction. *Protein Science* **2023**, *32* (2), e4551. <https://doi.org/10.1002/pro.4551>.
- (183) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Židek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>.
- (184) Evans, R.; O'Neill, M.; Pritzel, A.; Antropova, N.; Senior, A.; Green, T.; Židek, A.; Bates, R.; Blackwell, S.; Yim, J.; Ronneberger, O.; Bodenstein, S.; Zielinski, M.; Bridgland, A.; Potapenko, A.; Cowie, A.; Tunyasuvunakool, K.; Jain, R.; Clancy, E.; Kohli, P.; Jumper, J.; Hassabis, D. Protein Complex Prediction with AlphaFold-Multimer. October 4, 2021. <https://doi.org/10.1101/2021.10.04.463034>.
- (185) Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. ColabFold: Making Protein Folding Accessible to All. *Nat Methods* **2022**, *19* (6), 679–682. <https://doi.org/10.1038/s41592-022-01488-1>.
- (186) Bates, J. M.; Akerlund, J.; Mittge, E.; Guillemin, K. Intestinal Alkaline Phosphatase Detoxifies Lipopolysaccharide and Prevents Inflammation in Zebrafish in Response to the Gut Microbiota. *Cell Host & Microbe* **2007**, *2* (6), 371–382. <https://doi.org/10.1016/j.chom.2007.10.010>.
- (187) Rice, T. W.; Wheeler, A. P.; Bernard, G. R.; Vincent, J.-L.; Angus, D. C.; Aikawa, N.; Demeyer, I.; Sainati, S.; Amlot, N.; Cao, C.; Ii, M.; Matsuda, H.; Mouri, K.; Cohen, J. A Randomized, Double-Blind, Placebo-Controlled Trial of TAK-242 for the Treatment of Severe Sepsis*: *Critical Care Medicine* **2010**, *38* (8), 1685–1694. <https://doi.org/10.1097/CCM.0b013e3181e7c5c9>.

- (188) Ono, Y.; Maejima, Y.; Saito, M.; Sakamoto, K.; Horita, S.; Shimomura, K.; Inoue, S.; Kotani, J. TAK-242, a Specific Inhibitor of Toll-like Receptor 4 Signalling, Prevents Endotoxemia-Induced Skeletal Muscle Wasting in Mice. *Sci Rep* **2020**, *10* (1), 694. <https://doi.org/10.1038/s41598-020-57714-3>.
- (189) Yang, L.; Jiménez, J. A.; Earley, A. M.; Hamlin, V.; Kwon, V.; Dixon, C. T.; Shiao, C. E. Drainage of Inflammatory Macromolecules from the Brain to Periphery Targets the Liver for Macrophage Infiltration. *eLife* **2020**, *9*, e58191. <https://doi.org/10.7554/eLife.58191>.
- (190) Matsunaga, N.; Tsuchimori, N.; Matsumoto, T.; Ii, M. TAK-242 (Resatorvid), a Small-Molecule Inhibitor of Toll-Like Receptor (TLR) 4 Signaling, Binds Selectively to TLR4 and Interferes with Interactions between TLR4 and Its Adaptor Molecules. *Mol Pharmacol* **2011**, *79* (1), 34–41. <https://doi.org/10.1124/mol.110.068064>.
- (191) MacKenzie, S. A.; Roher, N.; Boltaña, S.; Goetz, F. W. Peptidoglycan, Not Endotoxin, Is the Key Mediator of Cytokine Gene Expression Induced in Rainbow Trout Macrophages by Crude LPS☆. *Molecular Immunology* **2010**, *47* (7–8), 1450–1457. <https://doi.org/10.1016/j.molimm.2010.02.009>.
- (192) Anderson, J. A.; Loes, A. N.; Waddell, G. L.; Harms, M. J. Tracing the Evolution of Novel Features of Human Toll-like Receptor 4. *Protein Science* **2019**, *28* (7), 1350–1358. <https://doi.org/10.1002/pro.3644>.
- (193) Ohto, U.; Miyake, K.; Shimizu, T. Crystal Structures of Mouse and Human RP105/MD-1 Complexes Reveal Unique Dimer Organization of the Toll-Like Receptor Family. *Journal of Molecular Biology* **2011**, *413* (4), 815–825. <https://doi.org/10.1016/j.jmb.2011.09.020>.
- (194) Ogata, H.; Su, I.; Miyake, K.; Nagai, Y.; Akashi, S.; Mecklenbräuker, I.; Rajewsky, K.; Kimoto, M.; Tarakhovskiy, A. The Toll-like Receptor Protein Rp105 Regulates Lipopolysaccharide Signaling in B Cells. *The Journal of Experimental Medicine* **2000**, *192* (1), 23–30. <https://doi.org/10.1084/jem.192.1.23>.
- (195) Divanovic, S.; Trompette, A.; Atabani, S. F.; Madan, R.; Golenbock, D. T.; Visintin, A.; Finberg, R. W.; Tarakhovskiy, A.; Vogel, S. N.; Belkaid, Y.; Kurt-Jones, E. A.; Karp, C. L. Inhibition of TLR-4/MD-2 Signaling by RP105/MD-1. *J. Endotoxin Res.* **2005**, *11* (6), 363–368. <https://doi.org/10.1179/096805105X67300>.
- (196) Oliveira-Nascimento, L.; Massari, P.; Wetzler, L. M. The Role of TLR2 in Infection and Immunity. *Front. Immun.* **2012**, *3*. <https://doi.org/10.3389/fimmu.2012.00079>.
- (197) Stafford, J. L.; Neumann, N. F.; Belosevic, M. Products of Proteolytic Cleavage of Transferrin Induce Nitric Oxide Response of Goldfish Macrophages. *Developmental & Comparative Immunology* **2001**, *25* (2), 101–115. [https://doi.org/10.1016/S0145-305X\(00\)00048-3](https://doi.org/10.1016/S0145-305X(00)00048-3).
- (198) Haddad, G.; Belosevic, M. Transferrin-Derived Synthetic Peptide Induces Highly Conserved pro-Inflammatory Responses of Macrophages. *Molecular Immunology* **2009**, *46* (4), 576–586. <https://doi.org/10.1016/j.molimm.2008.07.030>.

- (199) Trites, M. J.; Barreda, D. R. Contributions of Transferrin to Acute Inflammation in the Goldfish, *C. Auratus*. *Developmental & Comparative Immunology* **2017**, *67*, 300–309. <https://doi.org/10.1016/j.dci.2016.09.004>.
- (200) Husebye, H.; Halaas, Ø.; Stenmark, H.; Tunheim, G.; Sandanger, Ø.; Bogen, B.; Brech, A.; Latz, E.; Espevik, T. Endocytic Pathways Regulate Toll-like Receptor 4 Signaling and Link Innate and Adaptive Immunity. *EMBO J* **2006**, *25* (4), 683–692. <https://doi.org/10.1038/sj.emboj.7600991>.
- (201) Meyer, A.; Schartl, M. Gene and Genome Duplications in Vertebrates: The One-to-Four (-to-Eight in Fish) Rule and the Evolution of Novel Gene Functions. *Current Opinion in Cell Biology* **1999**, *11* (6), 699–704. [https://doi.org/10.1016/S0955-0674\(99\)00039-3](https://doi.org/10.1016/S0955-0674(99)00039-3).
- (202) Sullivan, C.; Charette, J.; Catchen, J.; Lage, C. R.; Giasson, G.; Postlethwait, J. H.; Millard, P. J.; Kim, C. H. The Gene History of Zebrafish *Tlr4a* and *Tlr4b* Is Predictive of Their Divergent Functions. *The Journal of Immunology* **2009**, *183* (9), 5896–5908. <https://doi.org/10.4049/jimmunol.0803285>.
- (203) Ohno, S. *Evolution by Gene Duplication*; Springer Berlin Heidelberg: Berlin, Heidelberg, 1970. <https://doi.org/10.1007/978-3-642-86659-3>.
- (204) Kayagaki, N.; Warming, S.; Lamkanfi, M.; Walle, L. V.; Louie, S.; Dong, J.; Newton, K.; Qu, Y.; Liu, J.; Heldens, S.; Zhang, J.; Lee, W. P.; Roose-Girma, M.; Dixit, V. M. Non-Canonical Inflammasome Activation Targets Caspase-11. *Nature* **2011**, *479* (7371), 117–121. <https://doi.org/10.1038/nature10558>.
- (205) Shi, J.; Zhao, Y.; Wang, Y.; Gao, W.; Ding, J.; Li, P.; Hu, L.; Shao, F. Inflammatory Caspases Are Innate Immune Receptors for Intracellular LPS. *Nature* **2014**, *514* (7521), 187–192. <https://doi.org/10.1038/nature13683>.
- (206) Hagar, J. A.; Powell, D. A.; Aachoui, Y.; Ernst, R. K.; Miao, E. A. Cytoplasmic LPS Activates Caspase-11: Implications in TLR4-Independent Endotoxic Shock. *Science* **2013**, *341* (6151), 1250–1253. <https://doi.org/10.1126/science.1240988>.
- (207) Yang, D.; Zheng, X.; Chen, S.; Wang, Z.; Xu, W.; Tan, J.; Hu, T.; Hou, M.; Wang, W.; Gu, Z.; Wang, Q.; Zhang, R.; Zhang, Y.; Liu, Q. Sensing of Cytosolic LPS through Casp2 Pyrin Domain Mediates Noncanonical Inflammasome Activation in Zebrafish. *Nat Commun* **2018**, *9* (1), 3052. <https://doi.org/10.1038/s41467-018-04984-1>.
- (208) Teufel, F.; Almagro Armenteros, J. J.; Johansen, A. R.; Gíslason, M. H.; Pihl, S. I.; Tsirigos, K. D.; Winther, O.; Brunak, S.; Von Heijne, G.; Nielsen, H. SignalP 6.0 Predicts All Five Types of Signal Peptides Using Protein Language Models. *Nat Biotechnol* **2022**, *40* (7), 1023–1025. <https://doi.org/10.1038/s41587-021-01156-3>.
- (209) Rojas, A.; Shiao, C. Brain-Localized and Intravenous Microinjections in the Larval Zebrafish to Assess Innate Immune Response. *BIO-PROTOCOL* **2021**, *11* (7). <https://doi.org/10.21769/BioProtoc.3978>.

- (210) Wang, S.; Song, R.; Wang, Z.; Jing, Z.; Wang, S.; Ma, J. S100A8/A9 in Inflammation. *Front. Immunol.* **2018**, *9*, 1298. <https://doi.org/10.3389/fimmu.2018.01298>.
- (211) Leukert, N.; Vogl, T.; Strupat, K.; Reichelt, R.; Sorg, C.; Roth, J. Calcium-Dependent Tetramer Formation of S100A8 and S100A9 Is Essential for Biological Activity. *Journal of Molecular Biology* **2006**, *359* (4), 961–972. <https://doi.org/10.1016/j.jmb.2006.04.009>.
- (212) Odink, K.; Cerletti, N.; Brügggen, J.; Clerc, R. G.; Tarcsay, L.; Zwadlo, G.; Gerhards, G.; Schlegel, R.; Sorg, C. Two Calcium-Binding Proteins in Infiltrate Macrophages of Rheumatoid Arthritis. *Nature* **1987**, *330* (6143), 80–82. <https://doi.org/10.1038/330080a0>.
- (213) Urban, C. F.; Ermert, D.; Schmid, M.; Abu-Abed, U.; Goosmann, C.; Nacken, W.; Brinkmann, V.; Jungblut, P. R.; Zychlinsky, A. Neutrophil Extracellular Traps Contain Calprotectin, a Cytosolic Protein Complex Involved in Host Defense against *Candida Albicans*. *PLoS Pathog* **2009**, *5* (10), e1000639. <https://doi.org/10.1371/journal.ppat.1000639>.
- (214) Hayden, J. A.; Brophy, M. B.; Cunden, L. S.; Nolan, E. M. High-Affinity Manganese Coordination by Human Calprotectin Is Calcium-Dependent and Requires the Histidine-Rich Site Formed at the Dimer Interface. *J. Am. Chem. Soc.* **2013**, *135* (2), 775–787. <https://doi.org/10.1021/ja3096416>.
- (215) Gagnon, D. M.; Brophy, M. B.; Bowman, S. E. J.; Stich, T. A.; Drennan, C. L.; Britt, R. D.; Nolan, E. M. Manganese Binding Properties of Human Calprotectin under Conditions of High and Low Calcium: X-Ray Crystallographic and Advanced Electron Paramagnetic Resonance Spectroscopic Analysis. *J. Am. Chem. Soc.* **2015**, *137* (8), 3004–3016. <https://doi.org/10.1021/ja512204s>.
- (216) Nakashige, T. G.; Stephan, J. R.; Cunden, L. S.; Brophy, M. B.; Wommack, A. J.; Keegan, B. C.; Shearer, J. M.; Nolan, E. M. The Hexahistidine Motif of Host-Defense Protein Human Calprotectin Contributes to Zinc Withholding and Its Functional Versatility. *J. Am. Chem. Soc.* **2016**, *138* (37), 12243–12251. <https://doi.org/10.1021/jacs.6b06845>.
- (217) Clark, H. L.; Jhingran, A.; Sun, Y.; Vareechon, C.; De Jesus Carrion, S.; Skaar, E. P.; Chazin, W. J.; Calera, J. A.; Hohl, T. M.; Pearlman, E. Zinc and Manganese Chelation by Neutrophil S100A8/A9 (Calprotectin) Limits Extracellular *Aspergillus Fumigatus* Hyphal Growth and Corneal Infection. *The Journal of Immunology* **2016**, *196* (1), 336–344. <https://doi.org/10.4049/jimmunol.1502037>.
- (218) Baker, T. M.; Nakashige, T. G.; Nolan, E. M.; Neidig, M. L. Magnetic Circular Dichroism Studies of Iron(II) Binding to Human Calprotectin. *Chem. Sci.* **2017**, *8* (2), 1369–1377. <https://doi.org/10.1039/C6SC03487J>.
- (219) Hadley, R. C.; Gagnon, D. M.; Brophy, M. B.; Gu, Y.; Nakashige, T. G.; Britt, R. D.; Nolan, E. M. Biochemical and Spectroscopic Observation of Mn(II) Sequestration from Bacterial Mn(II) Transport Machinery by Calprotectin. *J. Am. Chem. Soc.* **2018**, *140* (1), 110–113. <https://doi.org/10.1021/jacs.7b11207>.

- (220) Besold, A. N.; Gilston, B. A.; Radin, J. N.; Ramsoomair, C.; Culbertson, E. M.; Li, C. X.; Cormack, B. P.; Chazin, W. J.; Kehl-Fie, T. E.; Culotta, V. C. Role of Calprotectin in Withholding Zinc and Copper from *Candida Albicans*. *Infect Immun* **2018**, *86* (2), e00779-17. <https://doi.org/10.1128/IAI.00779-17>.
- (221) Manitz, M.-P.; Horst, B.; Seeliger, S.; Strey, A.; Skryabin, B. V.; Gunzer, M.; Frings, W.; Schünlau, F.; Roth, J.; Sorg, C.; Nacken, W. Loss of S100A9 (MRP14) Results in Reduced Interleukin-8-Induced CD11b Surface Expression, a Polarized Microfilament System, and Diminished Responsiveness to Chemoattractants In Vitro. *Molecular and Cellular Biology* **2003**, *23* (3), 1034–1043. <https://doi.org/10.1128/MCB.23.3.1034-1043.2003>.
- (222) Wang, C.; Iashchishyn, I. A.; Pansieri, J.; Nyström, S.; Klementieva, O.; Kara, J.; Horvath, I.; Moskalenko, R.; Rofougaran, R.; Gouras, G.; Kovacs, G. G.; Shankar, S. K.; Morozova-Roche, L. A. S100A9-Driven Amyloid-Neuroinflammatory Cascade in Traumatic Brain Injury as a Precursor State for Alzheimer’s Disease. *Sci Rep* **2018**, *8* (1), 12836. <https://doi.org/10.1038/s41598-018-31141-x>.
- (223) Grunwald, D. J.; Eisen, J. S. Headwaters of the Zebrafish — Emergence of a New Model Vertebrate. *Nat Rev Genet* **2002**, *3* (9), 717–724. <https://doi.org/10.1038/nrg892>.
- (224) Willett, C. E.; Cortes, A.; Zuasti, A.; Zapata, A. G. Early Hematopoiesis and Developing Lymphoid Organs in the Zebrafish. *Dev Dyn* **1999**, *214* (4), 323–336. [https://doi.org/10.1002/\(SICI\)1097-0177\(199904\)214:4<323::AID-AJA5>3.0.CO;2-3](https://doi.org/10.1002/(SICI)1097-0177(199904)214:4<323::AID-AJA5>3.0.CO;2-3).
- (225) Davidson, A. J.; Zon, L. I. The ‘Definitive’ (and ‘Primitive’) Guide to Zebrafish Hematopoiesis. *Oncogene* **2004**, *23* (43), 7233–7246. <https://doi.org/10.1038/sj.onc.1207943>.
- (226) Trede, N. S.; Langenau, D. M.; Traver, D.; Look, A. T.; Zon, L. I. The Use of Zebrafish to Understand Immunity. *Immunity* **2004**, *20* (4), 367–379. [https://doi.org/10.1016/S1074-7613\(04\)00084-6](https://doi.org/10.1016/S1074-7613(04)00084-6).
- (227) Lieschke, G. J.; Currie, P. D. Animal Models of Human Disease: Zebrafish Swim into View. *Nat Rev Genet* **2007**, *8* (5), 353–367. <https://doi.org/10.1038/nrg2091>.
- (228) Chen, H.; Xu, C.; Jin, Q.; Liu, Z. S100 Protein Family in Human Cancer. *Am J Cancer Res* **2014**, *4* (2), 89–115.
- (229) Marenholz, I.; Heizmann, C. W.; Fritz, G. S100 Proteins in Mouse and Man: From Evolution to Function and Pathology (Including an Update of the Nomenclature). *Biochemical and Biophysical Research Communications* **2004**, *322* (4), 1111–1122. <https://doi.org/10.1016/j.bbrc.2004.07.096>.
- (230) Kraemer, A. M.; Saraiva, L. R.; Korsching, S. I. Structural and Functional Diversification in the Teleost S100 Family of Calcium-Binding Proteins. *BMC Evol Biol* **2008**, *8* (1), 48. <https://doi.org/10.1186/1471-2148-8-48>.

- (231) Zhang, C.; Zhang, Q.; Wang, J.; Tian, J.; Song, Y.; Xie, H.; Chang, M.; Nie, P.; Gao, Q.; Zou, J. Transcriptomic Responses of S100 Family to Bacterial and Viral Infection in Zebrafish. *Fish & Shellfish Immunology* **2019**, *94*, 685–696. <https://doi.org/10.1016/j.fsi.2019.09.051>.
- (232) Wheeler, L. C.; Donor, M. T.; Prell, J. S.; Harms, M. J. Multiple Evolutionary Origins of Ubiquitous Cu²⁺ and Zn²⁺ Binding in the S100 Protein Family. *PLoS ONE* **2016**, *11* (10), e0164740. <https://doi.org/10.1371/journal.pone.0164740>.
- (233) Bozzi, A. T.; Nolan, E. M. Avian MRP126 Restricts Microbial Growth through Ca(II)-Dependent Zn(II) Sequestration. *Biochemistry* **2020**, *59* (6), 802–817. <https://doi.org/10.1021/acs.biochem.9b01012>.
- (234) Farnsworth, D. R.; Saunders, L. M.; Miller, A. C. A Single-Cell Transcriptome Atlas for Zebrafish Development. *Developmental Biology* **2020**, *459* (2), 100–108. <https://doi.org/10.1016/j.ydbio.2019.11.008>.
- (235) Hou, Y.; Lee, H. J.; Chen, Y.; Ge, J.; Osman, F. O. I.; McAdow, A. R.; Mokalled, M. H.; Johnson, S. L.; Zhao, G.; Wang, T. Cellular Diversity of the Regenerating Caudal Fin. *Sci. Adv.* **2020**, *6* (33), eaba2084. <https://doi.org/10.1126/sciadv.aba2084>.
- (236) Bhattacharya, S.; Chazin, W. J. Calcium-Driven Changes in S100A11 Structure Revealed. *Structure* **2003**, *11* (7), 738–740. [https://doi.org/10.1016/S0969-2126\(03\)00132-1](https://doi.org/10.1016/S0969-2126(03)00132-1).
- (237) Santamaria-Kisiel, L.; Rintala-Dempsey, A. C.; Shaw, G. S. Calcium-Dependent and -Independent Interactions of the S100 Protein Family. *Biochemical Journal* **2006**, *396* (2), 201–214. <https://doi.org/10.1042/BJ20060195>.
- (238) Harman, J. L.; Loes, A. N.; Warren, G. D.; Heaphy, M. C.; Lampi, K. J.; Harms, M. J. Evolution of Multifunctionality through a Pleiotropic Substitution in the Innate Immune Protein S100A9. *eLife* **2020**, *9*, e54100. <https://doi.org/10.7554/eLife.54100>.
- (239) Hadley, R. C.; Gu, Y.; Nolan, E. M. Initial Biochemical and Functional Evaluation of Murine Calprotectin Reveals Ca(II)-Dependence and Its Ability to Chelate Multiple Nutrient Transition Metal Ions. *Biochemistry* **2018**, *57* (19), 2846–2856. <https://doi.org/10.1021/acs.biochem.8b00309>.
- (240) Spratt, D. E.; Barber, K. R.; Marlatt, N. M.; Ngo, V.; Macklin, J. A.; Xiao, Y.; Konermann, L.; Duennwald, M. L.; Shaw, G. S. A Subset of Calcium-binding S100 Proteins Show Preferential Heterodimerization. *The FEBS Journal* **2019**, *286* (10), 1859–1876. <https://doi.org/10.1111/febs.14775>.