An Evolutionary and Biochemical Approach to Understanding the Mechanism of Activation of Innate Immune Receptor TLR4/MD-2 with Exogenous Lipopolysaccharides and Endogenous S100A9

by

Sophia Phillips

A dissertation accepted and approved in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in Chemistry

Dissertation Committee:

Dr. Bradley Nolen, Chair

Dr. Michael Harms, Advisor

Dr. Diana Libuda, Core Member

Dr. Calin Plesa, Core Member

Dr. Matthew Barber, Institutional Representative

University of Oregon

Spring 2025

DISSERTATION ABSTRACT

Sophia Rebecca Phillips

Doctor of Philosophy in Chemistry

Title: An Evolutionary and Biochemical Approach to Understanding the Mechanism of Activation of Innate Immune Receptor TLR4/MD-2 with Exogenous Lipopolysaccharides and Endogenous S100A9

The innate immune system is one of our first lines of defense against infection, and also recognizes and responds to internal tissue damage. TLR4 is an important innate immune receptor shared amongst vertebrates that performs both of these roles, recognizing lipopolysaccharides from Gram-negative bacteria as well the endogenous damage-associated protein S100A9. While TLR4 is widely studied in chronic inflammation, disease, and infection, many questions remain about the history and function of this receptor. In this work, I address two fundamental questions regarding TLR4: (1) How did TLR4 evolve species-specific recognition of different lipopolysaccharide structures, and (2) What is the binding mechanism of S100A9 and TLR4? I answer these questions combining evolutionary techniques such as ancestral sequence reconstruction and dN/dS, functional cell-based assays, mutagenesis, and *in silico* structure prediction and molecular dynamics simulations. The results of this work aim to uncover the evolutionary history of TLR4 specificity for microbial pathogens, as well as define a direct interaction between TLR4 and an endogenous protein which is implicated in many chronic inflammatory diseases that afflict human health.

This dissertation contains previously published and unpublished co-authored material.

CURRICULUM VITAE

NAME OF AUTHOR: Sophia Rebecca Phillips

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
California Polytechnic State University – San Luis Obispo

DEGREES AWARDED:

Doctor of Philosophy, Chemistry, 2025, University of Oregon
Bachelor of Science, Biochemistry, 2019, California Polytechnic State University - SLO

AREAS OF SPECIAL INTEREST:

Biochemistry
Biophysics
Molecular Biology
Protein Evolution

PROFESSIONAL EXPERIENCE:

Graduate Research and Teaching Assistant, University of Oregon, 2019-2024

GRANTS, AWARDS, AND HONORS:

NIH T32 Molecular Biology and Biophysics Training Program, University of Oregon, 2020-2022

PUBLICATIONS:

Svendsen J, Ford M, Asnes C, Oh S, Dorogin J, Fear K, O'Hara-Smith J, Chisholm L, Phillips S, Harms M, Hosseinzadeh P, Hettiaratchi M. "Applying computational protein design to engineer affibodies for affinity-controlled delivery of vascular endothelial growth factor and platelet-derived growth factor." Biomacromolecules. (2025). doi: 10.1021/acs.biomac.5c00097

Chisholm LO, Orlandi KN, Phillips SR, Shavlik MJ, Harms MJ (2024) Ancestral Reconstruction and the Evolution of Protein Energy Landscapes. Annual Review of Biophysics

53:127–146. https://www.annualreviews.org/content/journals/10.1146/annurevbiophys-030722-125440

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.

Harman, J.L.; Reardon, P.N.; Costello, S.M.; Warren, G.D.; Phillips, S.R.; Connor, P.J.; Marqusee, S.; Harms, M.J. "Evolution avoids a pathological stabilizing interaction in the immune protein S100A9.” (2022) PNAS 10.1073/pnas.2208029119

* Denotes authors contributed equally

ACKNOWLEDGEMENTS

First, I would like to thank my advisor, Dr. Mike Harms, for his mentorship over the past 5.5 years. Mike truly embodies leadership by example – whether it is helping graduate students or new professors he barely knows navigate transition in their academic careers, always having a thoughtful question to ask after seminars, polling our lab for input on his latest grant proposal and implementing feedback, spending dozens of hours reading primary literature and making original figures to improve his undergraduate course curriculum, or carving out a precious 45 minutes in his busy schedule to talk about the latest piece of data, Mike approaches his work with kindness, curiosity, creativity, and determination in spite of challenges and setbacks. He has inspired and encouraged me to face my own challenges with bravery and resilience, and I credit a large part of my persistence in graduate school and as a scientist to him.

Next, I would like to thank my academic community, my lab members and committee, for their help and feedback throughout my Ph.D. Previous lab members Dr. Anneliese Morrison, Dr. Michael Shavlik, Dr. Lauren Chisholm, Dr. Kona Shaw, and JJ Yin, as well as current lab members Brennan Fitzgerald, Natalie Jaeger, José Sanchez-Borbón, Evan LeGrand, and Amelia Kotamarti were instrumental – for passing the time during long lab days, for bouncing ideas off of, for teaching me newer or better techniques, for taking my LB plates out of the warm room, and for helping me complete experiments. I am grateful for the careful attention and suggestions of Dr. Brad Nolen, Dr. Diana Libuda, Dr. Matt Barber, and Dr. Calin Plesa to push forward and refine my thesis work.

My career as a scientist would not be possible without the support of my parents, Eloisa and Todd Phillips, and my brother Scott Phillips. They taught me to always be curious, and that it

is never wrong to follow my own path. Even though my parents don't quite understand what I do, they have the DOIs of all my papers saved on their home computer.

My time in graduate school was infinitely improved by the friends I made along the way. Dr. Acadia DiNardo and Molly Shallow were my family away from home, and were there for countless home-cooked dinners, early-morning drives to the mountains to ski or hike, several run-throughs of the Twilight movie series, commiseration about the woes of science, encouragement to send that late email, and everything in between. I would also like to thank Sam Horst, Dr. Jared Freedman, William Crow, Emily Dennis, Max Horrocks, Hannah Wilson, Brenden Campbell, Rachael Giersch, Sofia Carlson, Sophia Doerr, and Avery Bush for caring about me, and for so many antics including volleyball, tailgating, skiing, pickleball, dance, hiking, running, baking, trivia, and board games that took my mind off grad school and reminded me to find joy in every day.

This work was supported in part by the NIH T32 Molecular Biology and Biophysics Training Grant (5T32GM007759) awarded to me, and the R01-GM146114 grant awarded to Dr. Mike Harms at the University of Oregon.

TABLE OF CONTENTS

Chapter Page

Chapter Page

Chapter Page

CSV: MD2 PAML AND HYPHY RESULTS SUMMARY

CSV: LIST OF CONTACTS BETWEEN TLR4, MD-2, AND A9 IN BINDING MODELS

LIST OF FIGURES

Figure Page

Figure Page

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

This dissertation covers three bodies of work I have contributed to pertaining to protein evolution broadly, as well as a specific receptor of the innate immune system.

Chapter II is a published co-author manuscript of a bioinformatics pipeline for ancestral protein sequence reconstruction. Kona Shaw (nèe Orlandi) and myself were co-lead authors of this manuscript, Zachary Sailer and Joseph Harman were co-authors, and Michael Harms was the project and software development lead as well as a major writing contributor.

Chapter III is an unpublished manuscript investigating the evolutionary history of ligand specificity for the innate immune receptor TLR4. As lead author, I designed and performed a majority of the experiments, as well as being a major contributor to writing and editing. Lauren Chisholm contributed to methodology and data curation. Maggy Barry and Laurel Moneysmith contributed equally to data curation. Michael Harms is the corresponding author, and conducted molecular dynamics simulations and analysis, in addition to providing experimental guidance and overseeing writing and editing.

Chapter IV is unpublished work describing biochemical insight into how an endogenous protein, S100A9, activates TLR4. I provide functional and computational studies that propose future work to discern the mechanism by which this interaction occurs.

The innate immune system responds to foreign pathogens

The innate immune system is present in all vertebrates and is the first line of defense against infection. It includes physical barriers such as the skin, mucosal membranes in the gastrointestinal tract, and tears in the eyes. If microbes successful bypass these physical barriers,

they encounter a slew of different pattern recognition receptors (PRRs) found on the surface of immune cells or excreted in the extracellular matrix. PRRs have evolved over millennia to recognize microbe-associated molecule patterns (MAMPs) that may be dangerous to a host.[1–3] When PRRs encounter MAMPs they recognize, they are able to initiate downstream signaling cascades that mediate the inflammatory response, primarily by releasing proteins called cytokines and other small molecules which can further recruit other immune cells for wound healing, detection and killing of foreign microbes, removal of cellular debris, or activation of the adaptive immune system. Physiologically, a host may experience the inflammatory response as redness, swelling, pain, and heat at a site of infection or damage. In severe cases, other broad symptoms may emerge such as fever, organ dysfunction, and an increase in mucus production experienced in the nose, throat, and eyes.

Microbes and their hosts are often engaged in an evolutionary "arm's race", whereby microbes evolve MAMPs to evade host immune recognition and host immune systems evolve to recognize these new MAMPs. The consequence of hosts losing this arms' race can be fatal, and poses a threat to global health. Several of the leading causes of death globally are a result of innate immune dysfunction or chronic inflammation, including cardiovascular disease. It is estimated that 1 in 3 deaths globally occur due to immune dysfunction.[4]

Timely and coordinated recognition of MAMPs by the innate immune system are essential for proper clearance of foreign pathogens that maintain host health. On the other hand, dysregulation of the innate immune system can result in internal damage that is fatal for the host. Sepsis is a condition whereby an initial infection or damage can trigger the increasing production of cytokines, termed a "cytokine storm", which ratchets forward the inflammatory response until irreparable damage is done to the host. If not caught quickly, sepsis can rapidly onset and lead to

death within a few days of initial infection.[5,6] This is of critical concern to human health – one study on U.S. hospitals estimated that between 30-50% of hospital deaths occurred due to sepsis[7].

Toll-like receptors are a class of PRRs that mediate inflammation

Toll-like receptors, initially named for their similarity to the *Drosophila melanogaster* Toll receptor, are a vertebral class of PRRs that recognize MAMPs. TLRs have diversified across vertebrates, with 27 being identified in total.[8,9] However, different lineages have evolved their own sets of TLRs, with birds (for example) having 10 and mammals having 13. While different TLRs have evolved to bind different MAMPs, each TLR shares characteristic structural motifs. All TLRs have repeating loops of leucine-rich repeats (LRRs) which give them a candy cane-like structure. TLRs require binding of a ligand to drive dimerization with another TLR to form either a homo- or heterodimer, and dimerization is what activates the receptor to signal downstream pathways for an inflammatory response.

TLRs are found in different cell types and cellular compartments, with some being constitutively expressed while others are upregulated or present for activation only when recruited.[10] Some TLRs are transmembrane proteins with intra- and extra-cellular regions, while others are only accessible on one side of a membrane. There is some redundancy in TLR function and ligand recognition, with the duplication, gain, and loss of TLRs in various lineages.

Toll-like receptor 4 recognizes both foreign pathogens and internal damage

Toll-like receptor 4 (TLR4) is a TLR found across vertebrates, and is present on the surface of immune and other cell types. It is a large transmembrane protein with an extracellular

portion largely consisting of LRRs, a transmembrane portion, and an intracellular TIR domain.[11] Interestingly, TLR4 is the only TLR that requires a co-receptor, MD-2, to activate. TLR4 and MD-2 natively for a heterodimer, and some studies suggest that co-expression of MD-2 with TLR4 is necessary for proper trafficking of TLR4 to the cell surface.[12] TLR4 was originally identified in 1998 as the receptor that recognizes lipopolysaccharides, which are found on the surface of Gram-negative bacteria.[13] The mechanism by which LPS interacts with TLR4/MD-2 has since been well-characterized. The hydrophobic acyl chains of the lipid A moiety tuck into a hydrophobic pocket in MD-2, creating a complementary surface that drives dimerization with a second TLR4/MD-2:LPS complex. Dimerization of the TLR4 complex initiates the intracellular NF-kB signaling cascade, which triggers the production of proinflammatory cytokines and thus the inflammatory response. A wide collection of studies have illuminated how these different endotoxin features fine-tune the TLR4/MD-2 binding interaction as well as the inflammatory response for pathogen detection or evasion.[14–16]

More recently, it was found that TLR4 can recognize and initiate inflammation in response to another class of molecules – damage-associated molecular patterns, or DAMPs. DAMPs are host molecules or proteins that are released under cellular damage or death, dispersing them into the extracellular matrix where they may be recognized by PRRs such as TLR4.[17,18] Many DAMPs, including S100s, HMGB1, tenascin-C, HSP70, and others have been identified as activators of inflammation in a TLR4-dependent manner. While interactions with DAMPs may be beneficial for the host to remove or repair damaged tissue, over-activation of TLR4 by these molecules can result in disease states that are dangerous to a host.[19] For example, increased levels of DAMPs have been observed in cardiovascular disease, arthritis, some cancers, inflammatory bowel diseases, and others chronic inflammatory diseases.[20–22]

Importantly, these interactions can be dangerous because the host innate immune system is being activated by "self" molecules which are presumably necessary for other important cellular functions. It can be challenging to treat these diseases as the initiators of inflammation, DAMPs, cannot be easily removed. Thus, an important treatment strategy for reducing DAMP-mediated inflammation is to disrupt their interaction with TLR4.

The mechanisms by which DAMPs interact with and activate TLR4 are poorly understood. An obvious reason for this is that many DAMPs have no structural similarity to LPS, the canonical ligand of TLR4. Importantly, the only known binding site on the TLR4 complex is primarily encapsulated by the hydrophobic pocket of MD-2. Many DAMPs that are known to activate TLR4 do not physically fit in the MD-2 binding pocket. Thus, several models have been proposed to describe DAMP activation of TLR4, including DAMP proteolysis which would allow for peptide fragments to bind MD-2, directly or indirectly inducing crowding of the membrane to spatially confine and drive TLR4 dimerization, or direct binding on a novel, unknown site on TLR4.[23–26] Currently, none of these models have been widely accepted due to a lack of biophysical and high-quality structural evidence. Practically, several drug candidates which putatively disrupt DAMP-TLR4 binding interactions have failed in clinical trials, revealing that our knowledge of these interactions are lacking.[27]

S100A9 acts as a DAMP for TLR4

S100A9 (A9) is a small homodimeric calcium-binding protein that is abundant in neutrophils. Under non-stress conditions, A9 has several function including metabolism of arachidonic acid and microtubule organization for phagocyte migration.[28,29] It can also form a heterodimer with S100A8 and has antimicrobial properties in this state. It is part of the family of

S100 proteins in tetrapods, of which 21 homologs have been identified in humans. S100s appear to have emerged about 500 million years ago in the ancestor we last shared with jawless fish, although a recent work in our lab found that some early diverging fish such as zebrafish do not have an S100 ortholog.[30–32] TLR4 also appears to have emerged about 500 million years ago in the ancestor of all vertebrates, although it appears that the interaction between these two proteins occurred more recently, in the ancestor of amniotes ~320 million years ago.[8,33,34]

TLR4 evolution is constrained by maintaining recognition for a variety of ligands

Proteins that are involved in pathogen recognition are key targets for natural selection, especially that which results in a change in protein sequence and biochemical features.[1] However, TLR4 and MD-2 have many constraints they must balance to maintain their function. Notably, TLR4/MD-2 must (1) stably association with each other but not activate the receptor (i.e. form a dimer of dimers) in the absence of ligand, (2) specifically bind certain endotoxin variants in agonistic or antagonistic orientations, (3) dimerize when agonistic ligand is bound, and (4) associate with accessory or other downstream signaling proteins (CD14, MyD88, TRAM, etc.). It is assumed that many of these features are conserved to maintain fitness of the host, given that many studies demonstrate how dysregulation or mutational studies leading to constitutive activity, non-native or weak endotoxin recognition, or inability to bind to other signaling proteins lead to adverse and often fatal health effects on the host.[35–37]

Several groups have made observations between species regarding the differences in their sequence features, structure, and activation with different LPS variants (ref, ref, ref). Many groups have performed mutagenesis on TLR4 and MD-2, either heavily informed by published structural information or broadly and un-biasedly scanning the full sequence.[38–41] In fact, the

Meng group has proposed a full set of seven residues in TLR4/MD-2 which, when species-swapped between human and mouse, completely switch antagonism for agonism (respectively), although it has previously been unclear if these positions are descriptive of all species-specific TLR4/MD-2 ligand specificity.[42] Notably, all studies to date have lacked rigorous evolutionary data to help inform models for currently observed ligand specificity. There has been a lack of data from earlier diverging extant creatures including sauropsids and monotremes which could help reveal ligand specificity in early TLR4/MD-2 history, as well as a potential evolutionary trend in ligand specificity. Structure-based studies also generally yield candidate mutations that obviously perturb the binding interaction, but we were curious if we could identify non-obvious second or third shell positions that affect activity. Further, there is an abundance of high-quality sequence data that has not been fully utilized to show the real evolutionary trajectories and sites that are important for protein function and specificity. Evolutionary studies in tandem with structural information can reveal non-obvious but necessary substitutions that occurred that enable functional substitutions in a binding pocket to perform their function, which were non-obvious with structural studies alone.[43,44] Conserved versus divergent sequence and structural features between species can help reveal the underpinnings of function.

Modern day proteins evolve from ancestral states

Modern day proteins arrive at their current sequence, structure, and functional states from previous template(s) – their ancestral predecessors. Thus, all proteins encode artefacts of historical pressures and constraints they faced during their evolution. Understanding the evolutionary history of a protein can uncover many important pieces of information. For example, what were the environmental pressures a creature faced when it existed? What

sequences and structures were viable in their native cellular environment, accounting for other existing proteins and potentially competing molecules? What function did it perform? Tracking over the evolutionary trajectory of a protein, there is also an implicit but important question to answer: what where the biophysical, evolutionary, and genetic constraints that defined this protein's evolution? While incredibly valuable for understanding protein evolution, we cannot reasonably track the answers to many of these questions for many proteins in real time.

Some ancestors have fossils which can be collected and sequenced, but it is likely that we will not be able to know with perfect certainty for many ancestors what the sequences and ancestral states of proteins were.[45] Because it is difficult or impossible to go back in time and measure protein evolution, and because we likely cannot accurately model the specific pressures that shaped protein evolution, we can instead simplify this question by using modern-day, high quality sequence information to estimate what previous ancestral states were.[46] Considering that closely-related species shared a common ancestor more recently, and their respective protein orthologs have had less time to accumulate sequence changes, we can use the species tree as well as modern day sequences, we can track how recently or anciently amino acid states appeared, and implement a substitution-based model to estimate the sequence of ancestral proteins.

Ancestral sequence reconstruction is a bioinformatics methods that has improved studies on protein structure, function, and evolution since it was first proposed in 1963.[47] It is broadly applicable to any protein family study, as long as a sufficient quantity and quality of modern day sequences exist. A significant barrier to incorporating this method has previously been the computational expertise and familiarity with existing programs necessary to intelligently string together a functional multi-step pipeline, which often requires extensive reformatting and organization of files to "communicate" between programs. A great benefit to this field would

reduce the technological skill needed to do ASR so that scientists with expert knowledge of their protein of interest can perform more in-depth studies which may complement other methods.

Bridge to Chapter II

In Chapter I, I introduce the role the innate immune system plays in recognizing and responding to both foreign microbes and internal damage. I describe what is known mechanistically about how the innate immune receptor TLR4 recognizes specific microbe- and damage-associated molecular patterns LPS and S100A9, respectively. Further, I provide background for what is currently known about how recognition of these ligands evolved. I finally end the chapter with a brief explanation of what ancestral sequence reconstruction is and what questions it can help us answer in protein evolution. In Chapter II, I present a bioinformatics tool, topiary, I helped to develop which automates most steps of ancestral sequence reconstruction that were previously burdensome and time-consuming. This chapter also provides example usage of the pipeline, as well as validation of the results and tips on how to interpret results of an ASR 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[47,48]. 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[46]. For example, ASR has been used for crystallographic and kinetic studies on ancestral proteins when their modern-day descendants were not amenable to crystallization[49], for bioengineering enzymes that are both thermally stable and catalytically active using ancestral enzymes as templates[50], and in the discovery of an ancestral coagulation factor VIII protein that is now used as a therapeutic for people with hemophilia[51]. These, and many other studies[48,51–56] 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[48]. The problem is made worse because some often-used software is no longer maintained: for example, PAML4 was last updated in 2007[57]. 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 choice[58]: 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 2.1**). 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").

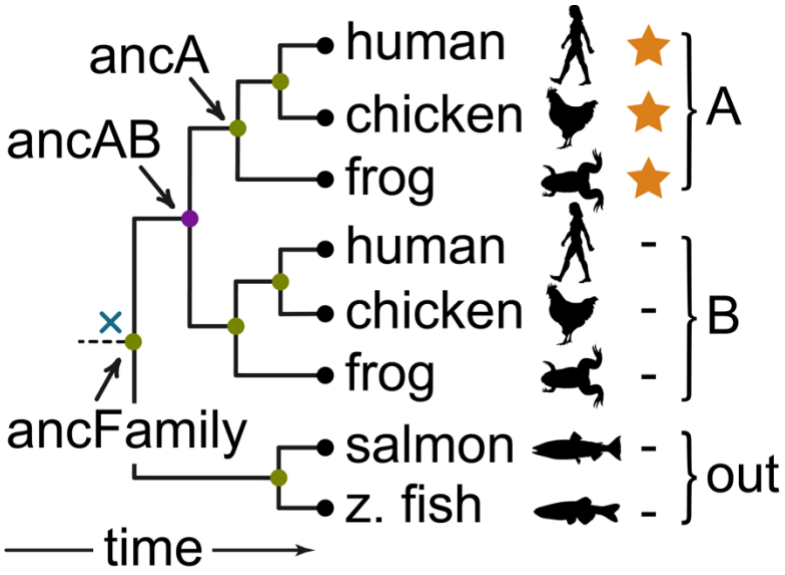

Figure 2.1. Define the ancestral reconstruction problem. The panel shows the evolutionary history of a hypothetical protein family with two paralogs A and B. The tree is rooted: ancestors are arranged from ancient to recent, left to right. Black circles at the tips of the tree denote modern protein sequences 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[59]. 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[59]. 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[60,61] and/or by manual evaluation and editing[62,63]. 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[64].

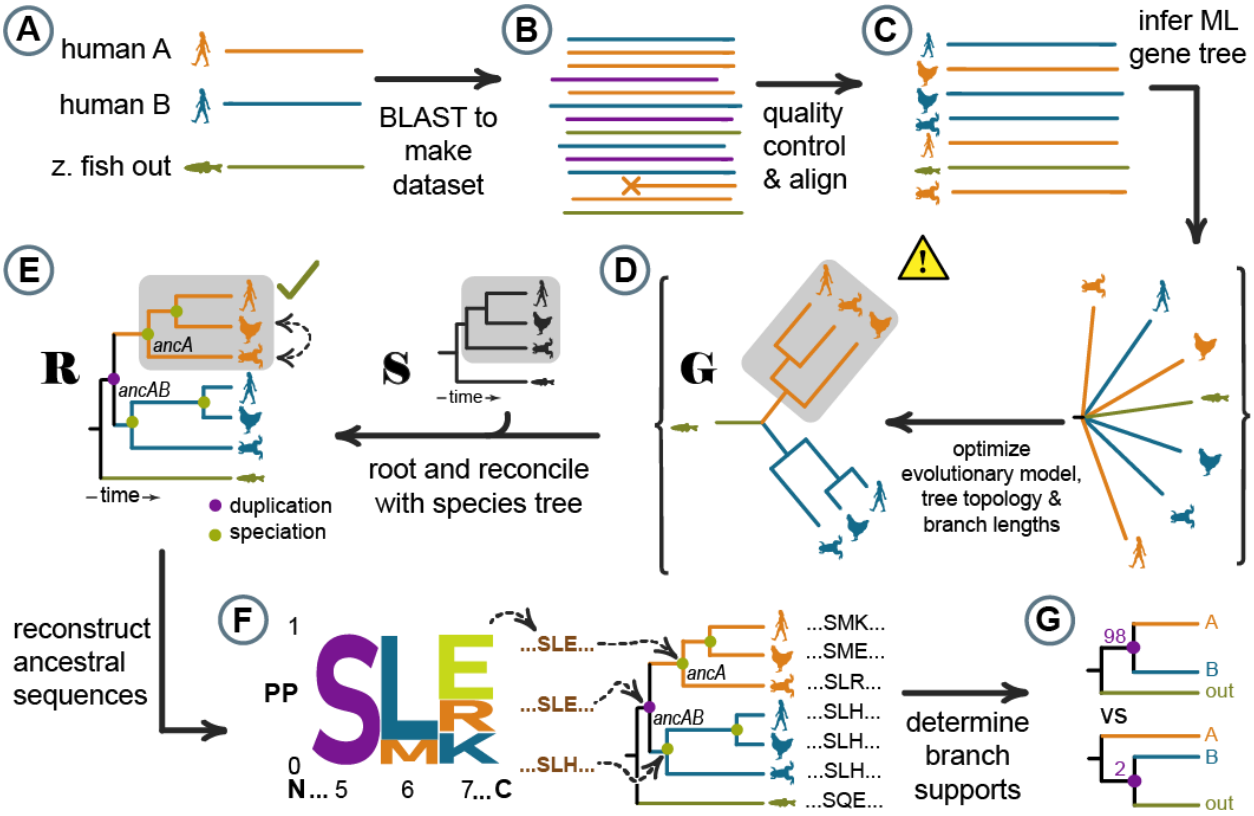

Figure 2.2. Ancestral sequence reconstruction has six main steps. (**A**) Start with a handful of homologous protein sequences spanning the paralogs of interest and their taxonomic distribution. Throughout the figure, color indicates the identity of the protein (orange: paralog A, blue: paralog B, and 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 (that is, 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 [48,56,65].

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** and **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[66].

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[67].

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 species.[68]) As a result, ASR studies of microbial proteins have

generally relied on unreconciled gene trees[69]. 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[70]. 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[71]. 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, $PP_i = L_i/sum(L_{aa})$, where L_i is the likelihood of amino acid i and $sum(L_{aa})$ 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 *The Topiary Pipeline* section 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[72]. 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[52,54,73–76].

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 *The Topiary Pipeline* section; 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 *Protocol* section 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 alignment[77]; RAxML-NG for maximum likelihood gene tree and ancestral sequence inference[78]; GeneRax for gene-species tree reconciliation[66]; and PastML for gap reconstruction[79]. Under the hood, it uses the ETE 3 library for tree manipulations[80]; Biopython to access NCBI BLAST and the NCBI database[81,82]; python-opentree to interact with the Open Tree of Life taxonomic database[58,83]; and toytree for drawing trees[84]. 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/harmslab/topiary. A collection of example datasets and Jupyter notebooks are available at https://github.com/harmslab/topiary-examples.

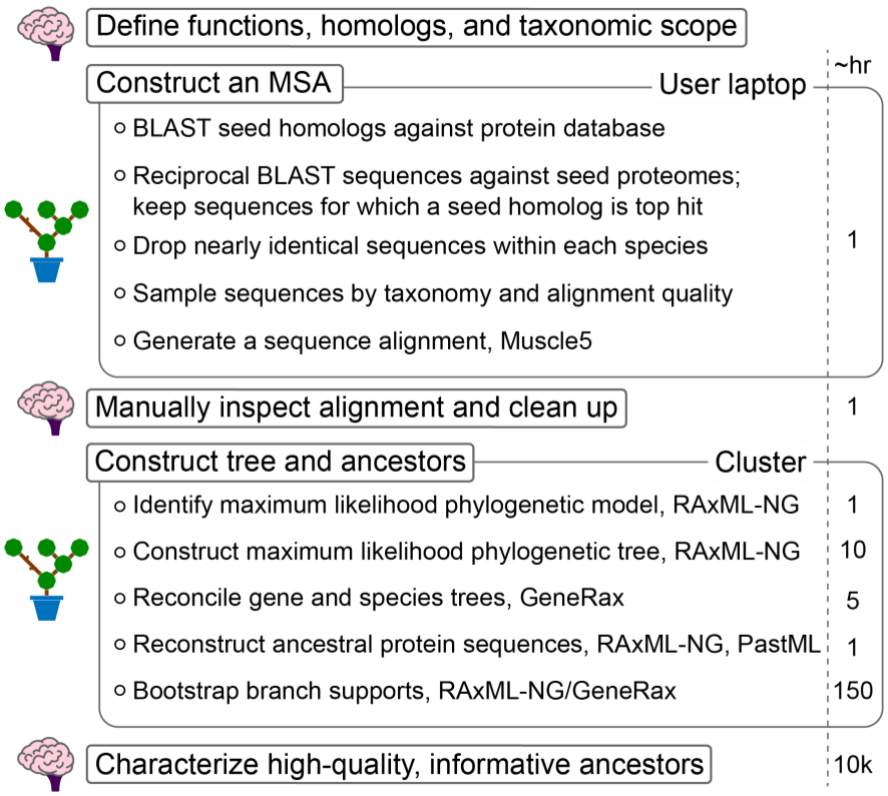

Figure 2.3. Summarized topiary ASR pipeline. The pipeline is a series of human and automatic steps (indicated on the left with brain and topiary icons, respectively). The approximate time, in hours, required for each step is indicated on the right.

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 the *Protocol* section below. 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 5,000 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 *Protocol* for details about defining aliases.) It weights each hit by $2^{s/t}$ 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[85]. 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** to **2.2C**). 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 4**).

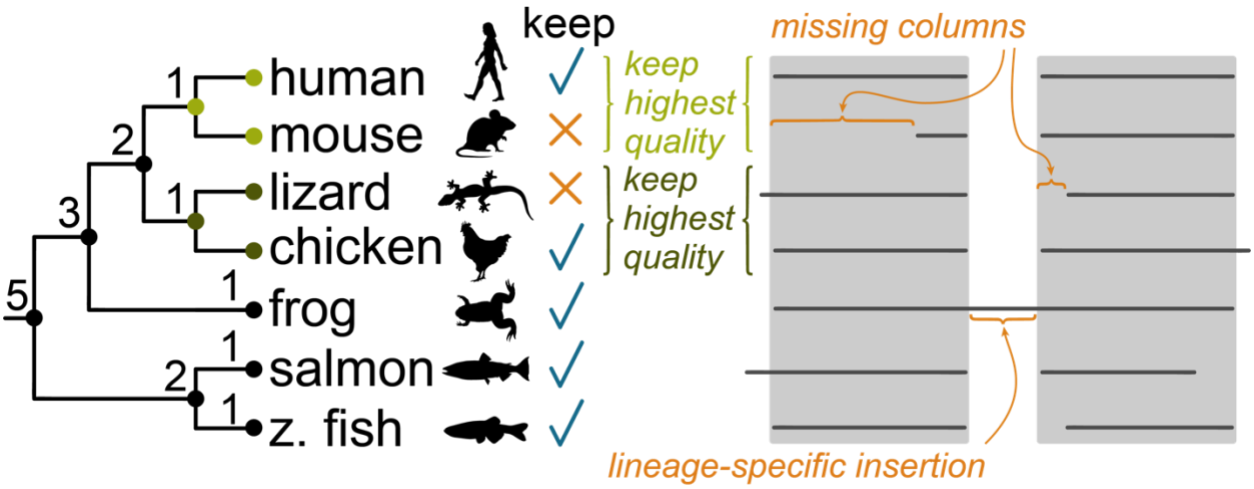

Figure 2.4. Topiary redundancy reduction and quality control. This analysis starts with seven sequences (taken from seven organisms) with the goal of retaining five for the downstream analysis. The numbers next to the ancestral nodes on the tree are the budget allocated for all descendants: 5 for all organisms, 2 for the fishes, 3 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**)[86]. 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[62,63]. 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[60]. In practice, most of our manual work consists

of deleting a handful of problematic sequences, followed by global realignment in Muscle5. (See the *Protocol* section for details.)

Stage 2: Alignment to Ancestors

In stage 2, we go from our alignment to ancestral sequences (**Figure 2.2C-G; Figure 2.3**). We selected RAxML-NG[78] 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 RAxML-NG use the same underlying computational phylogenetics library—libpll[87]—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[88]. 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[89]. It uses RAxML-NG 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[67]. 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)[66]. 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[58,83]. (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 parameters[66]: 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[71]. (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 sites[72]; 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 DOWNPASS[90] algorithm as implemented by the PastML package[79]. 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 *Protocol* section).

Branch supports

To determine branch supports (**Figure 2.2G**), topiary uses non-parametric bootstrapping[88]. 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 1,000 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 2.2A**). 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 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 AliView[91], JALView[92], and MEGA[93]. We generally use AliView because of its balance of utility and simplicity.

There are differing views on whether to manually edit an alignment[62,63]; 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 (~1,000 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—i.e., for microbial genes—only the steps shown in Figures 2.5A and 2.5D 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[94]. 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[66]. 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.

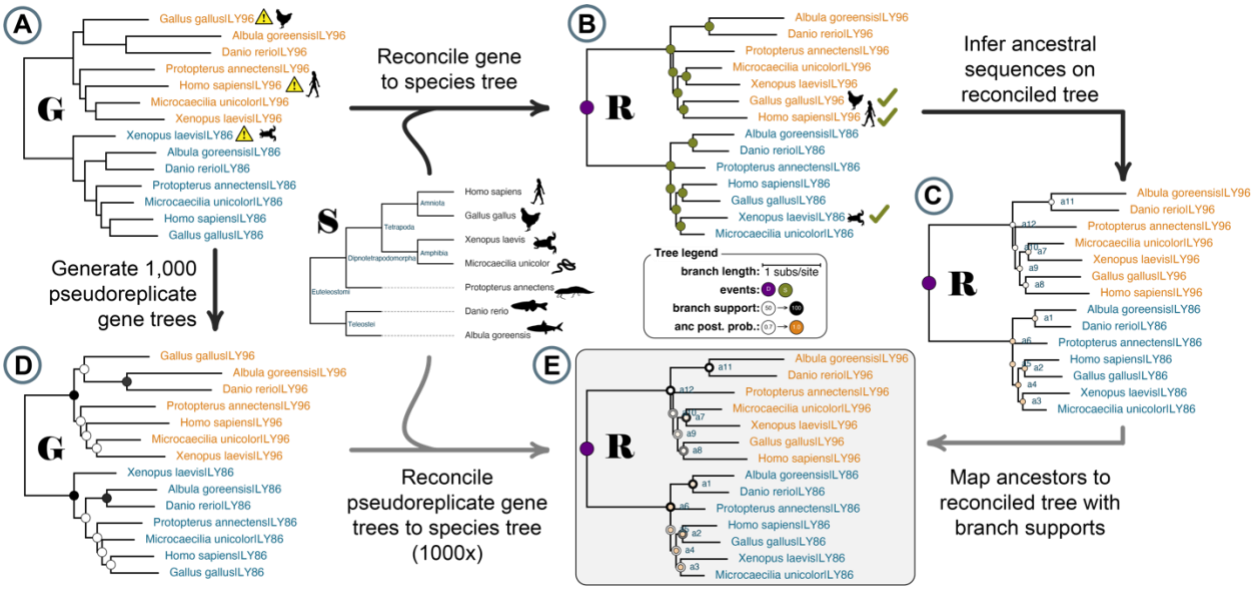

Figure 2.5: Example trees at each step in the ASR calculation. Summary trees from an ASR inference using a toy alignment with seven LY96 sequences (orange) and seven LY86 sequences (blue). Black arrows indicate steps done by the first script (*alignment-to-ancestors*); gray arrows indicate steps done 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 1,000 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's 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.

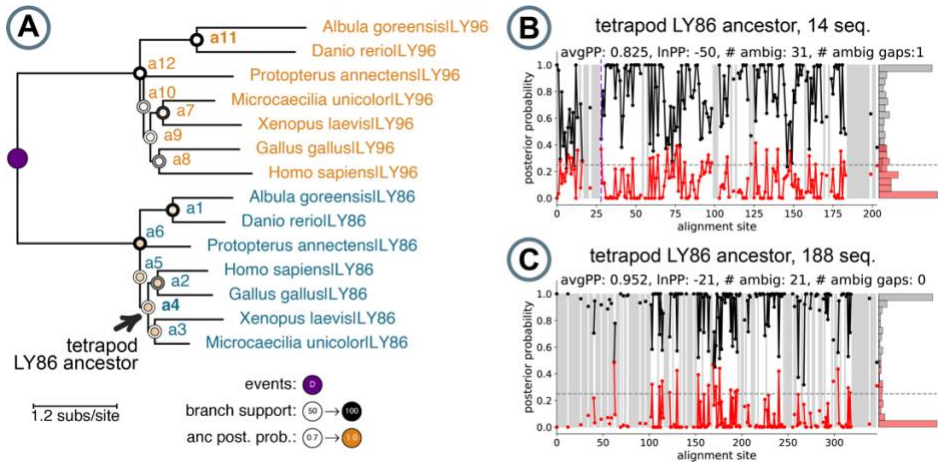

Figure 2.6: Graphs for evaluating ancestor quality. (**A**) The final bootstrap supported gene-species reconciled tree built from an example set of 14 sequences. Reconstructed ancestral sequences at each node are labeled with a unique name. Duplication events are marked in purple. Each node is labeled 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** and **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.

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 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** and **2.6C** 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[72]. 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[52,54,73–76].

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/topiary-examples). 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 RAxML-NG 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.

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

Much of what topiary does is 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-HIT[95], 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 Index[96] as calculated by DendroPy[97] (**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.

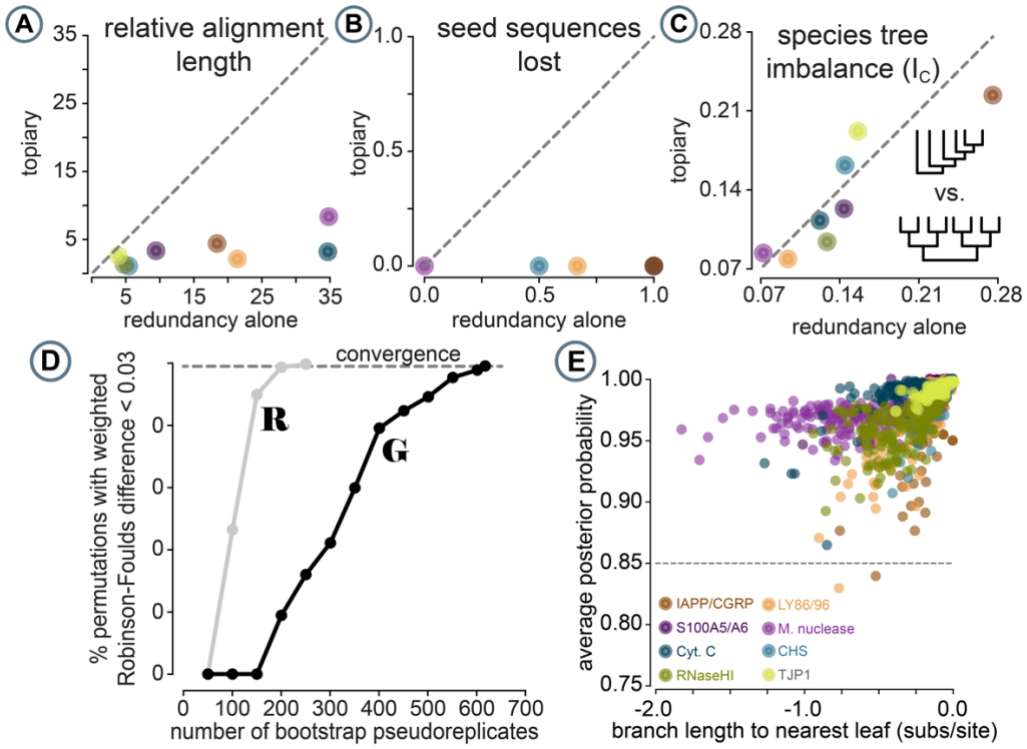

Figure 2.7: Validation of the topiary pipeline. Panels show topiary results generated for the eight protein families from Table 2. Colors indicate the family in question (see panel E for color legend). Panels A-C show topiary alignment quality as measured by three metrics: (**A**) Relative alignment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

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[98]. 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 less than 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 (1,027 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 1,027 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, 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

In Chapter II I presented topiary, an ancestral sequence reconstruction pipeline which streamlines the computational task of reconstructing high-quality ancestral proteins. Topiary integrates best-practice software and methods into a user-friendly interface to make ASR more accessible for a variety of scientists who may lack computational expertise but whose work may greatly benefit from ancestral information. In Chapter III, I pivot to discussing my work on the evolution of species-specific LPS specificity for an innate immune receptor, TLR4. I utilize both bioinformatic evolutionary methods such as ASR and dN/dS in conjunction with *in silico* modeling and functional characterization of TLR4 from various species or mutant states. These methods help me discern both historical patterns in ligand specificity as well as providing further insight into the relationship between sequence, structure, and function that shape TLR4 ligand specificity.

CHAPTER III

FRIEND OR FOE: HOW THE INNATE IMMUNE RECEPTOR TLR4 EVOLVED SPECIFICITY FOR LPS

*This chapter contains unpublished co-authored material.

Author contributions: S. R. P.: Conceptualization (lead); data curation (lead); investigation (lead); methodology (lead); software (lead); validation (lead); visualization (lead); writing – original draft (lead); writing – review and editing (lead). L. O. C.: methodology (supporting); software (supporting). L. M.: data curation (supporting); validation (supporting). M. B.: data curation (supporting); validation (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

Toll-like receptor 4 (TLR4) of the vertebrate innate immune system differentiates between members of its host's microbial community via recognition of diverse lipopolysaccharides (LPS) from Gram-negative bacteria. TLR4 either agonistically binds LPS variants which initiates the inflammatory response, or antagonistically binds for immunosilencing. Despite the importance and prevalence of this interaction, the evolutionary history and mechanism by which TLR4 toggles specificity for these ligands has previously been understudied. The work presented here characterized TLR4 specificity in modern day and reconstructed ancestral sauropsids and found that while agonism for hexa- and hepta-acylated LPS is highly conserved, antagonism for tetra- and penta-acylated LPS has been evolutionarily labile. A previous model proposing the necessary and sufficient set of species-swapping mutations between human and mouse to switch tetra-acylated LPS (L4) activity proved to be inconsistent with creatures we characterized. This inspired us to perform dN/dS calculations, molecular dynamics simulations, and functional characterization of single-point mutants to understand if sites that are under diversifying selection and directly interact with acyl chains are responsible for maintaining agonism with L4. We found that agonism of L4 is sensitive to and easily damaged by mutation of sites under selection, regardless of the species origin of the mutation. Molecular dynamics simulations with mouse TLR4 crystal structures revealed that even agonistic binding of L4 induced fewer contacts than L6 in the dimerization interface of the TLR4 receptor, and that the E122K mutation to co-receptor MD-2 primarily affecting L4 binding interactions without affect to L6. Taken together, this work provides evidence for how the innate immune system balances competing selection and biophysical constraints throughout evolution to maintain specificity for an ever-changing microbial environment.

Introduction

The animal innate immune system continuously evolves to recognize and respond to microbes it encounters in its environment.[1,99] Some of these microbes are pathogenic, others commensal, and still others conditionally pathogenic. While it is unknown exactly how many different types of bacteria inhabit or interact with individual creatures, one study on human microbiome diversity found that at least 395 bacterial phylotypes are represented, and other regions such as the mouth and skin have their own lively bacterial communities that contribute significantly to total diversity.[100,101] This diversity places innate immune system genes under possibly conflicting evolutionary pressures to both recognize and differentiate between pathogenic and commensal molecules, often with high structural similarity.

Three of the key players involved in this response are Toll-like receptor 4 (TLR4) and its co-receptors MD-2 and CD14 (Fig. 3.1A). These proteins, found on the surface of animal immune cells, are early responders to Gram-negative bacterial infection. They trigger an inflammatory response, inducing physiological symptoms within hours of infection across mammals.[102–105] This response is critical in humans for response to a wide variety of disease including pneumonia, typhoid fever, whooping cough, meningitis, urinary tract infections, and bloodstream infection.[106,107] When dysregulated, however, this response can be deadly. Sepsis—an overactive and self-destructive TLR4 response to LPS—is a leading cause of death worldwide.[7,108]

One potent activator of the TLR4 complex is lipopolysaccharide (LPS), an endotoxin found on the cell membrane of Gram-negative bacteria. Lipopolysaccharides are generally composed of a "lipid A" moiety of acyl chains attached to a decorated diglucosamine backbone, a "core" consisting of short sugar chains, and an "O-antigen" made of longer carbohydrate

polymers (Fig. 3.1B).[109] The canonical LPS structure which was first characterized, and has been broadly shown to activate the TLR4 receptor across animals, is hexa-acylated LPS from *E. coli*.[13,110–112] Since then, it has been shown that Gram-negative bacteria produce a variety of LPS structures that vary in the number, length, and saturation of their lipid A acyl chains, the identity and side groups of the lipid A backbone, as well as the carbohydrate composition of their core and O-antigens.[113] Variation in LPS structure occurs both between species and within individual organisms.

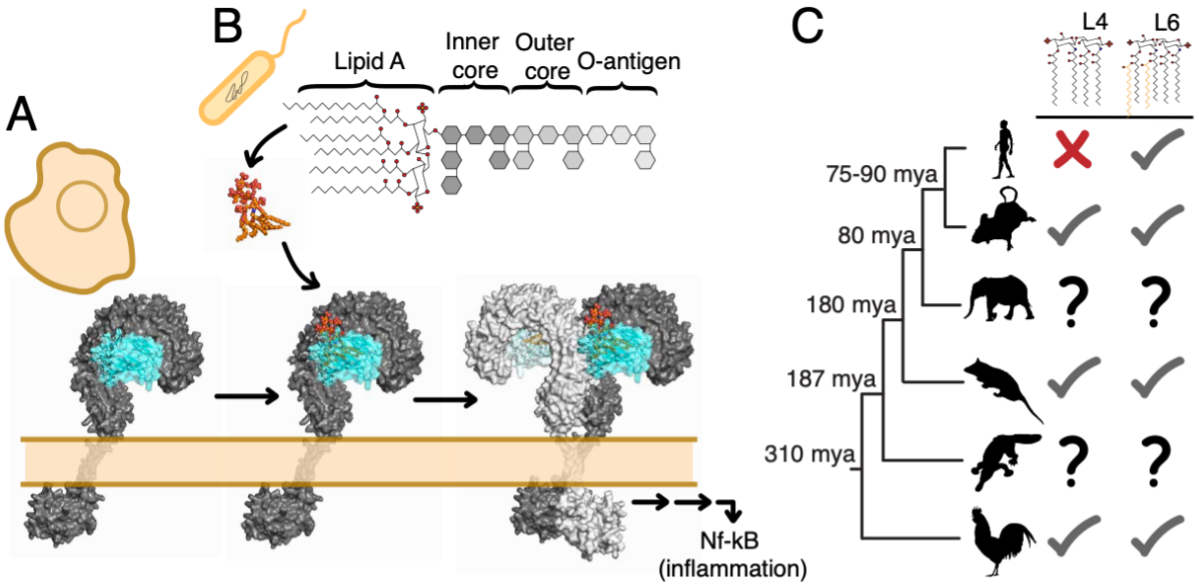

Figure 3.1. TLR4/MD-2 recognize lipopolysaccharides (LPS) to activate inflammation. (A) Mechanism for TLR4/MD-2 binding to hexa-acylated LPS on the surface of a cell. Cartoon depicts human TLR4/MD-2 crystal structure (PDB: 3FXI).[11] (B) Cartoon structure of LPS found on the surface of Gram-negative bacteria. Common structural motifs are labeled. (C) Evolutionary species tree of sauropsids with TLR4/MD-2 activity (mya = million years ago). Human, mouse, opossum, and chicken have known activation with L4 and L6; elephant and platypus were previously unknown.

While it is obvious that the TLR4 complex must recognize and initiate an immune response for certain LPS variants, it is equally important for TLR4 to not respond to variants made by commensal microbes. For example, a recent study on total human gut LPS found that

the most abundant Gram-negative clade in the commensal gut microbiome is *Bacteroidetes*, which predominantly produces variants of LPS with four and five acyl chains that are unable to activate human TLR4/MD-2.[114] Indeed, these "hypo-acylated" LPS variants antagonize human TLR4 by binding non-productively to the binding pocket. Healthy human individuals appear to have a 6:1 – 18:1 ratio of antagonizing versus agonizing LPS in their gut, pointing to the prevalence and importance of immunosilencing in the gut for human health.[115]

No studies to date have systematically investigated the evolutionary history of specificity for hyper- and hypo-acylated LPS - that is, if agonism has been an evolved trait in certain lineages or stochastically appearing in different species. Additionally, it has been unclear if there are trade-offs in recognizing hypo-acylated LPS at the expense of hexa- or hepta-acylated LPS. Thus, we hypothesized two possible modes of evolution for TLR4/MD-2 endotoxin recognition: one, that agonism (or antagonism) was an ancestral trait that slowly evolved specificity over time for different LPS acylation states, or two, that there has not been an evolutionary pattern and specificity is highly species-specific.

Here, we address the evolutionary history of ligand specificity through functional characterization of modern-day and reconstructed ancestral TLR4s with several LPS variants, dN/dS, mutagenesis studies, and molecular dynamics (MD) simulations. Functional and structural studies provide greater context to the biophysical laws that constrain protein evolution.

Results

The ability to recognize hypo-acylated LPS fluctuated across mammalian lineages.

Based on previous work, we hypothesized that broad LPS recognition was an ancestral feature of sauropsid TLR4 complexes, followed by the loss of the ability to respond to hypo-

acylated LPS in the human lineage (Fig 3.1C). To test this hypothesis, we characterized the specificity of the African elephant (*Loxodonta africana*) and platypus (*Ornithorhynchus anatinus*) TLR4 complexes. We also measured the activity of the human, mouse, opossum, and chicken complexes. We measured activity using a well-established NF-κB reporter assay. Briefly, we transfect plasmids containing a luciferase reporter and species-matched TLR4, MD-2, and CD14 into HEK293T cells. We then treat the cells with increasing concentrations of endotoxin and measure the dose-dependent luciferase output. Finally, we fit a mathematical model to each dataset, extracting both an EC_{50} and maximum activation for each LPS/complex pair (see Materials and Methods).

We measured the activity of endotoxin variants with four, five, six, and seven acyl chains. *E. coli* LPS—the canonical endotoxin used across myriad studies —has six acyl chains (L6). This molecule activates every previously characterized sauropsid TLR4 complex.[110–112] As a representative molecule with seven acyl chains (L7), we used LPS from *Salmonella enterica* (serotype typhimurium), which is responsible for typhoid fever.[14] For our molecule with four acyl chains (L4), we used lipid IVa, a biosynthetic precursor to *E. coli* LPS. Finally, use used the penta-acylated LPS from *Rhodobacter sphaeroides* as our representative L5. Both L4 and L5 antagonize the human complex but initiate an immune response in several other species (Fig. 3.1C).[116–118]

As has been observed previously, chicken and mouse TLR4 activated in response to all four LPS variants (Fig. 3.2A).[116] In contrast, human TLR4 activated in response to L6 and L7, but not L4 and L5 (Fig. 3.2A). We also reproduced the previous observation that L4 binds, but cannot activate, human TLR4. We did so by measuring the ability of a mixed LPS treatment

consisting of L6 and 10-fold excess L4 to activate the human complex. As expected, L4 competed for the binding site and lowered the L6 activation (Fig. 3.2B).

To our surprise, the platypus and elephant complexes behaved more similarly to the human complex than mouse, opossum, or chicken. These complexes responded robustly to L6 and L7, but weakly—or not at all—to L4 and L5 (Fig. 3.2A). Further, like the human complex, L4 binds but does not activate these complexes, as shown by the competitive activation assay (Fig. 3.2B).

To better understand the phylogenetic distribution of complex specificity, we plotted the maximum activation for the TLR4 complex of each species in response to each LPS variant (Fig. 3.2C). We found no statistically significant difference between maximum activation with L6 versus L7 in any species examined. In contrast, human, elephant, and platypus had significantly lower responses to L4 and L5 compared to L6 (p-value < 0.001). In the complexes we studied, we found that L4 antagonism was highly correlated with L5 antagonism. This suggests that structural features that allow for L4 activation also mediate L5 activation. We did not find any complex that activated with L5 but not L4.

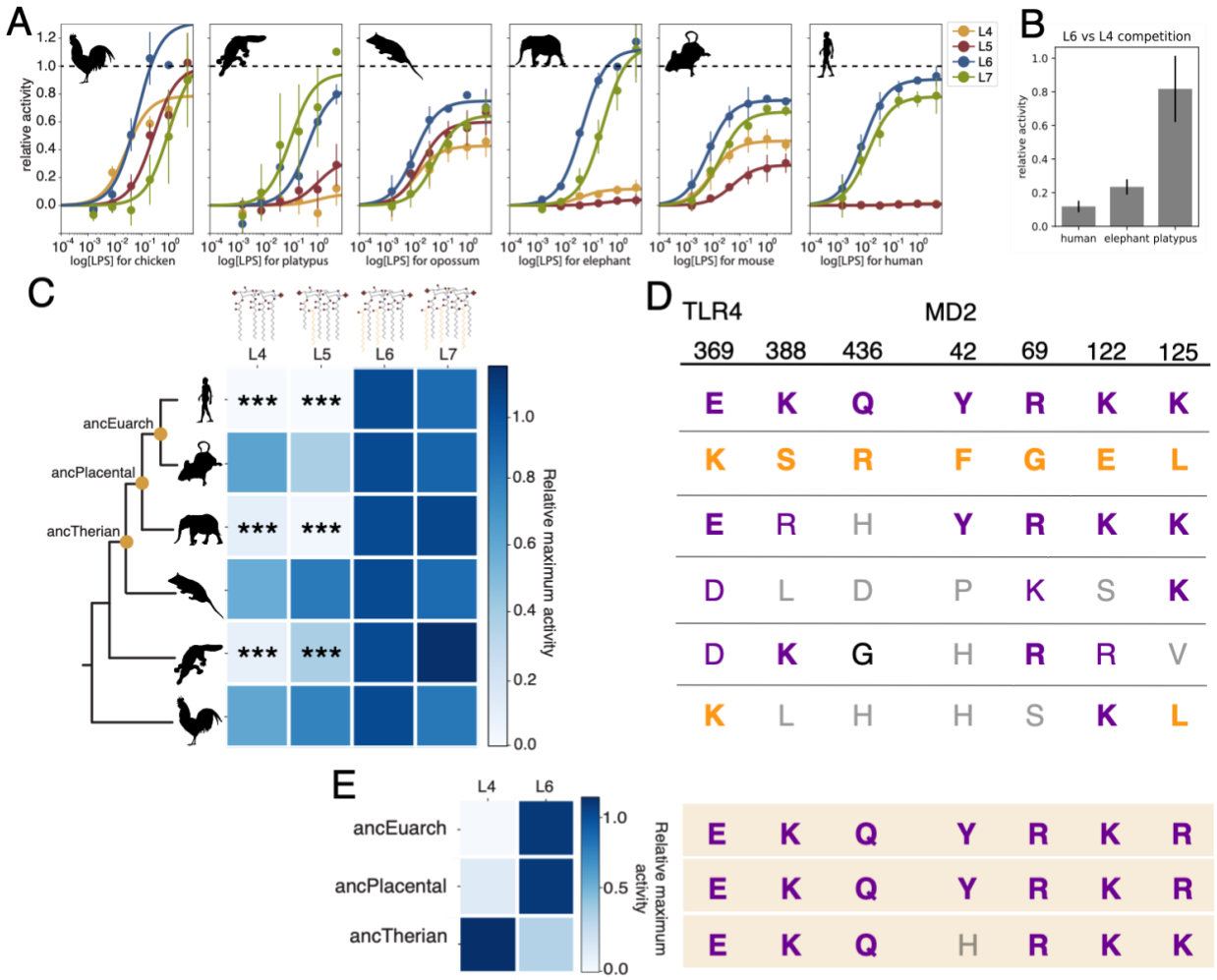

Figure 3.2. LPS specificity has fluctuated over evolutionary time. (A) Dose response curves for TLR4/MD-2 from sauropsid species treated with LPS variants ranging from 4-7 acyl chains. LPS variants are *E. coli* biosynthetic precursor lipid IVa (L4; yellow), *R. sphaeroides* LPS (L5, red); *E. coli* LPS (L6, blue), and *S. typhimurium* (L7, green). Activity was measured using an NF-κB reporter in HEK293T cells transiently transfected with complex components, normalized to XX. (B) Competition between L4 and L6 for human, elephant, and platypus complexes. Relative activity is background-subtracted and normalized such that L6 = 1.0. (C) Heatmaps summarizing maximum activation from (A) for each species and LPS variant. *** denote p-value < 0.001 difference from L6 activation using Tukey's HSD post-hoc test applied to a two-level ANOVA. Characterized ancestral nodes are labeled in orange circles. (D) Alignment of notable homologous positions (human TLR4/MD2 numbering) that are known to have structural and functional importance for binding LPS and inducing inflammation.[42] Exact conserved residues for human (purple) and mouse (orange) are bolded, similar residues are in regular font. Non-conserved residues with either species background are gray. (E) Alignment of reconstructed ancestors and maximum activation for L4 and L6.

Previously identified sites important for specificity cannot explain this pattern

We next sought to understand whether this pattern represented lineage-specific gains or losses of L4 activation. Many studies have been done previously to disentangle the structural features of TLR4/MD-2 that are responsible for endotoxin specificity.[119,120] This includes a careful study that identified the necessary and sufficient set of mutations to switch the human and mouse proteins between L4 agonism and antagonism.[42] This set includes seven sites, three from TLR4 and four from MD-2.

We looked at the amino acid states of our six species at these seven sites (Fig. 3.2D). We found that most species were more similar to the human than mouse sequence. The elephant had five identical amino acids, the opossum three, the platypus four, and the chicken one. Despite having a similar ability to respond to L4 and L5, the opossum shared no sequence overlap with the mouse at these seven sites. This is consistent with the independent acquisition of L4 agonism in these two lineages. Likewise, despite a similar phenotype to the mouse, the chicken only shares two sites.

Based on this analysis, we predicted that L4 antagonism was the ancestral state, with L4 agonism independently acquired several times. To test this hypothesis, we turned to ancestral sequence reconstruction. We generated ancestors of TLR4 complexes in this tree. This is a relatively challenging reconstruction, as it requires accurate reconstruction of three rapidly-evolving genes (TLR4, MD-2, and CD14). We reconstructed high-quality ancestors for ancEuarch (the last common ancestor of Euarchontoglires, which include humans and mice) and ancPlacental (the last common ancestor of placental mammals). The reconstruction was marginal for ancTherian and relatively poor for ancMammal and ancSauropsid. Within each of the ancestral complexes, the gene with the lowest posterior probability was 0.993 (ancEuarch), 0.985

(ancPlacental), 0.876 (ancTherian), and below 0.8 for ancMammal and ancSauropsid. These summary statistics, however, conceal some uncertainty. Another way to describe uncertainty is with the number of ambiguous sites, defined as those where the posterior probability of the next-best reconstructed state was >0.25. For example, while the ancEuarch reconstruction had only 37 ambiguous sites, the ancTherian had 205.

We first characterized the sensitivity of the well-reconstructed ancEuarch complex to L4 and L6. These proteins were identical to the human protein at the seven sites of interest (all sites in this region had posterior probability > 0.98). Consistent with the sequence analysis, this protein was activated by L6 but not L4. We next characterized ancPlacental, finding it behaved identically to ancEuarch: it shared the human sites and activated with L6 but not L4. Consistent with the modern species data, this demonstrates that L4 activation was newly acquired in mice from an ancestor that exhibited L4 antagonism.

We then characterized ancTherian. This protein shared six of the seven sites with human, but its specificity changed radically. This complex now activates with L4 but not L6. Because none of the known sauropsid TLR4 complexes exhibited this pattern of specificity, we believe this is an artifact of the reconstruction rather than the reflection of a real ancestral state. This does, however, reveal that it is biochemically possible to achieve a TLR4 complex with inverted specificity. Given the low quality reconstruction scores for the deeper ancestors—and the strange behavior of the ancTherian ancestor—we elected not to characterize the deeper ancestors.

TLR4 has sites that are rapidly diversifying

We next sought to understand the evolutionary pattern in which L4 sensitivity was continuously gained from a "default" position of L4 antagonism. We hypothesized that rapidly

evolving TLR4 and MD-2 sites were responsible for the observed acquisition of L4 activation. To test this hypothesis, we first generated alignments of 20 mammalian species broadly spanning the mammalian clade for both TLR4 and MD-2 (See Supplement Files) and then tested for regions and sites with elevated rates of evolution. Given the known binding location of endotoxin in the complex, we focused our evolutionary analyses on the extracellular domain of TLR4 and full-length MD-2.[11,121]

We first used GARD to probe for distinct regions of each protein evolving with different evolutionary rates. The TLR4 ectodomain was best described by three regions evolving at different rates (Fig. 3.3A). One region (sites 291-375, $\omega = 0.35$) was evolving faster than the other regions (sites 1-290, $\omega = 0.22$ and sites 376-608, $\omega = 0.27$). In contrast, we found that MD-2 evolution could be described with a single evolutionary rate.

We next sought to identify regions and sites under diversifying selection using the programs FEL, FUBAR, MEME and PAML. We applied the methods individually to each of the three TLR4 regions identified by GARD, as well as to full-length MD-2 (see Materials and Methods). Overall, we found that 68 of 608 (11.2%) of sites on TLR4 were under positive selection in at least one test, while 145/608 (23.8%) were under negative selection (see Supplemental Files). For MD-2, 24/160 (15.0%) were under positive selection in at least one test, while 35/160 (21.9%) were under negative selection.

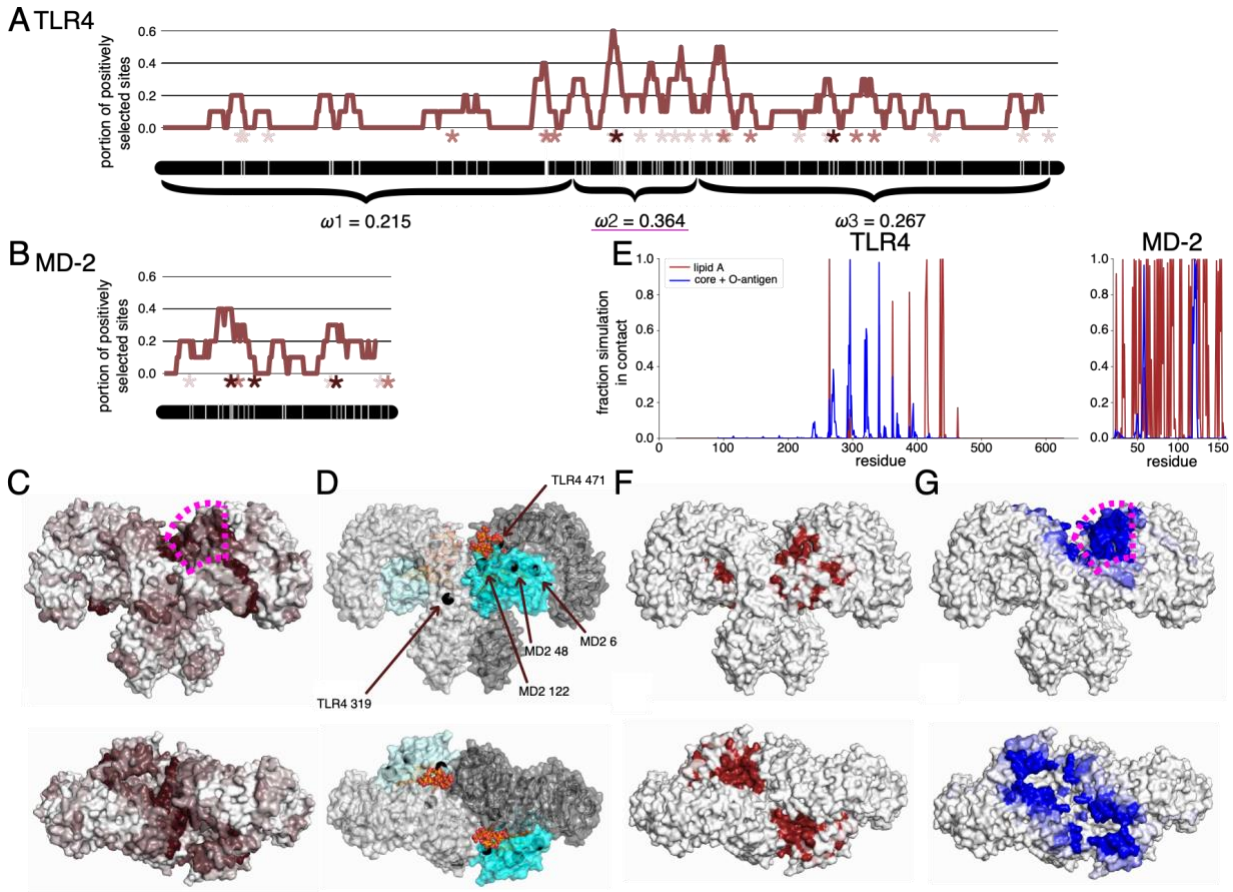

Figure 3.3. Sites under selection on TLR4 and MD-2 recognize different features of LPS. Sliding window analysis of sites under selection identified in up to four dN/dS calculations for (A) TLR4 and (B) MD-2. Window slides along ten-residue increments. Sites identified under positive selection in one test (white tick), two tests (light pink star), three tests (medium pink star), or all four tests (dark pink star) are denoted. (C) Crystal structure of human TLR4/MD-2 colored by sliding window portion of positively selected sites from (A) and (B). Magenta dashed line surround TLR4 GARD breakpoint with highest rate of evolution. (D) Crystal structure of human TLR4/MD-2 with chains colored. Large black spheres denote residues under highest selection (i.e. identified in all four evolutionary tests) on TLR4 and MD2. (E) Fraction of simulation time TLR4 (left panel) or MD-2 (right panel), by residue, are in contact within 5 angstroms of lipid A (red) or core + O-antigen (blue) portions of hexa-acylated LPS (see Methods and Materials). Lipid A (F) and core + O-antigen (G) contacts are mapped to structure. Maximum color saturation is set to sites that form contacts >20% of the simulation time.

The sites under positive selection on TLR4 and MD-2 form a continuous surface when mapped onto the crystal structure of the complex (Fig. 3.3C). To understand the origins of this

surface, we performed molecular dynamics simulations with docked LPS molecules with five O-antigens. We then asked which regions of the LPS interacted with which regions of TLR4/MD-2. We found that the fastest evolving region of TLR4 interacted with the O-antigen and core of the bound endotoxin.[11] Both endotoxin features are highly variable between bacterial species, suggesting that this interface is a mechanism through which TLR4 attempts to "keep up" with and recognize new ligand variants it encounters. The slower-evolving regions, by contrast, corresponded to conserved protein-protein interaction interfaces within the complex. Sites 1-290 span the interface between TLR4 and MD-2, while sites 376-608 correspond to the TLR4 dimerization interface. Some of residues in the latter region are involved with recognition of acyl chains from bound endotoxin; however, most form direct contacts with the opposite TLR4 chain.

MD-2 has sites that are under diversifying selection

Because we are interested in the evolution of specificity for acyl chain number, we focused the remainder of our investigation on MD-2. This is because the vast majority of contacts between the acyl chains and the complex are mediated by MD-2 (Fig. 3.1A; 3.4A). Further, previous studies have provided evidence that MD-2 primarily defines the binding mode of endotoxin.[42] For example, co-transfection in a human cell line of mTLR4/hMD-2 treated with tetra-acylated lipid IVa results in a near complete loss of activation, which is consistent with native human activation but not mouse activation; meanwhile, activity with hexa-acylated LPS appears to be unaffected, demonstrating that the chimeric receptor complex is still functional.

We further validated that MD-2 forms the primary interaction with the lipid A portion of LPS via molecular dynamics simulations sampling the surfaces of TLR4 and MD-2 that interact with the lipid A versus the core and O-antigen of hexa-acylated LPS from *E. coli* (Fig. 3.3E-G). We found that while TLR4 does form several contacts with lipid A, almost every residue in MD-

2 interacts directly with the lipid A portion. On the other hand, TLR4 forms the majority of the core and O-antigen contacts, which are also known to diversify as a potential bacterial host-evasion strategy.[122]

Three sites on MD-2 were identified as under positive selection by all four methods (FEL, FUBAR, MEME and PAML): sites 48, 64, and 122 (Fig. 3.3B). Position 122 has been well-documented through structural, functional, and mutagenesis studies to be essential for forming productive contacts with both endotoxin and the dimerizing TLR4.[42,123,124] Other positions which were identified have an unknown function in MD-2. Side chains at position 48 face inward and interact with the acyl chains of endotoxin tucked into MD-2. Position 64 has no known functional or structural importance – it is located away from the mouth of MD-2 (i.e. away from the binding pocket) but also too far away from any region of TLR4 to participate in complex formation. The side chain for this residue also faces outward. This was an intriguing position to explore how positions away from the binding interface might contribute to LPS specificity.

In addition to sites 48, 64, and 122, we also selected positions 42, 86, and 125 for further study. MD-2 position 42 is attractive as a site under positive selection, given that this position is known to form important contacts with the N-terminal end of TLR4 for initial association, and previous mutagenesis studies suggest that introducing a mutation at this position can (de)stabilize the TLR4/MD-2 interaction in a way that affects dimerization and thus activation of the receptor.[11,42] Position 86 has yet to be investigated, but appears to be directly in the dimerization interface and interact with acyl chains on bound endotoxin. Finally, position 125 is directly in the mouth of the binding pocket, and has been shown previously in mutagenesis and structural

studies to contribute to association of MD-2 with TLR4, endotoxin orientation stabilization, and dimerization of the ligand-bound receptor.

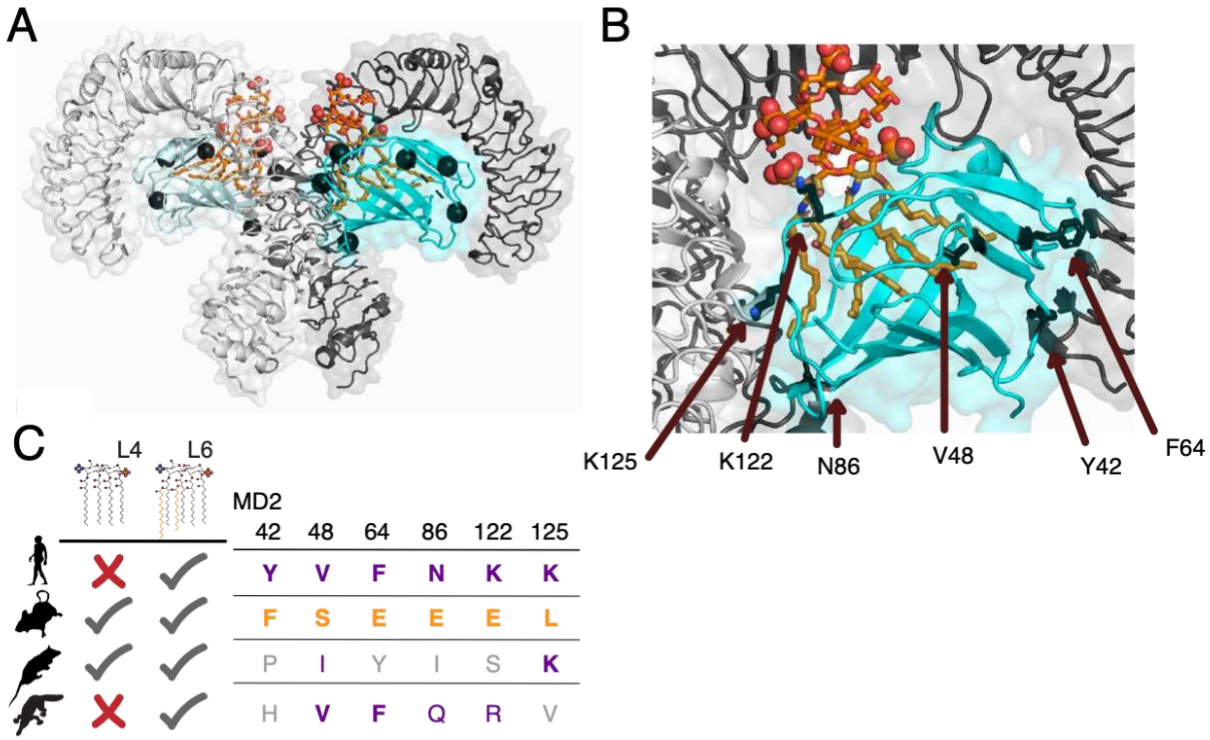

Figure 3.4. Candidate MD-2 sites identified under positive selection for mutational analysis. (A) Crystal structure of human TLR4/MD-2 bound to hexa-acylated LPS from *E. coli*, including the lipid A and inner core moieties (PDB 3FXI). Candidate MD-2 sites labeled with the alpha carbon as black spheres. (B) Close-up of candidate MD-2 sites with amino acid state and numbering in human MD-2. (C) Amino acid state at each candidate MD-2 position, for each species of interest. Ability to activate with L4 and L6 denoted with a check (agonism) or an "x" (antagonism).

Key sequence features interact epistatically to allow species to toggle between specificity for smaller endotoxin.

Our functional characterization suggested that the mouse and opossum convergently acquired the ability to recognize L4 as an agonist. We hypothesized that this was due to selection to recognize this class of ligands. To test this hypothesis, we took the six sites identified in our dN/dS analysis (Fig. 3.4C) and introduced the observed states into three genetic backgrounds:

human, mouse, and opossum. (We also attempted to measure the effects of mutations in the platypus complex; however, the overall activation of the complex was too low to confidently resolve the effects of mutations. We therefore did not proceed with mutations in this species background).

We made two predictions based on our hypothesis. First, we predicted that mutations at sites under selection would primarily effect L4, not L6, activation. Second, we predicted that mutations originating from species that are antagonized by L4 (human and platypus) would reduce L4 activity, whereas mutations originating from species that are agonized by L4 (mouse and opossum) would increase L4 activity.

We introduced mutations at each of six MD-2 candidate sites into three different species backgrounds and measured the activity of these mutants challenged with L6 and L4 (Fig. 3.4C, 3.5A). For each species' wildtype TLR4/MD-2, we considered the relative activity with L4 versus L6 to be the "wildtype ratio". Any mutation that globally increased or decreased both L4 and L6 activity but retained this ratio affected the overall ability of the receptor complex to activate, but not affect the specificity for either ligand. For example, this L6:L4 ratio in human is very low (0.01) because human minimally activates with L4 compared to its activity with L6. Alternatively, mouse and opossum have a L4:L6 ratio closer to 1 (0.7 and 0.9, respectively) because they activate with either ligand similarly. We then measured the difference (i.e. distance from the wildtype ratio line) between a mutant L4:L6 ratio versus the wildtype L4:L6 ratio for each species background; if the difference is negative, it indicates that the mutant disrupts L4 activity relative to L6 activity (Fig. 3.5C). If the difference is positive, it indicates that the mutation increases L4 activity relative to L6 activity.

For our first prediction, we found that it was only true in some species that MD-2 mutations as a whole affected L4 activation more than L6 activation. For example, the human background did not start with L4 agonism, and no single mutation created L4 agonism (Fig. 3.5B). Some mutations to the human background did thus cause slight variations from wildtype. In the mouse background, mutations introduced small perturbations to both L4 and L6 activation (Fig. 3.5C). Finally, in the opossum background mutations as a whole did significantly affect L4 activation more than L6 activation (Fig. 3.5D).

As for our second prediction, we found a different but robust trend: while specificity for larger endotoxin is generally conserved and difficult to break through single mutations, sensitivity to smaller endotoxin is easy to damage with a single mutation (Fig. 3.5B). Notably in all species backgrounds studied, no single mutation at these sites imparted measurable sensitivity to L4. We found no mutations with a statistically significant positive effect on L4 activity, but 13 mutations that statistically decreased L4 activity with respect to wildtype.

To better understand what features impact L4 activation, we performed a three-way ANOVA on our dataset with position, species background, and species origin of mutation as factors and distance from the wildtype L4:L6 ratio line, d, as our dependent variable. While we found that position and species background were significant factors ($p = 0.008$ and 1.28E-5, respectively), species of origin for mutation was not ($p = 0.698$). There was also a significant interaction effect between position and species ($p = 0.00128$), suggesting that different species backgrounds were more sensitive to mutation at some sites versus others.

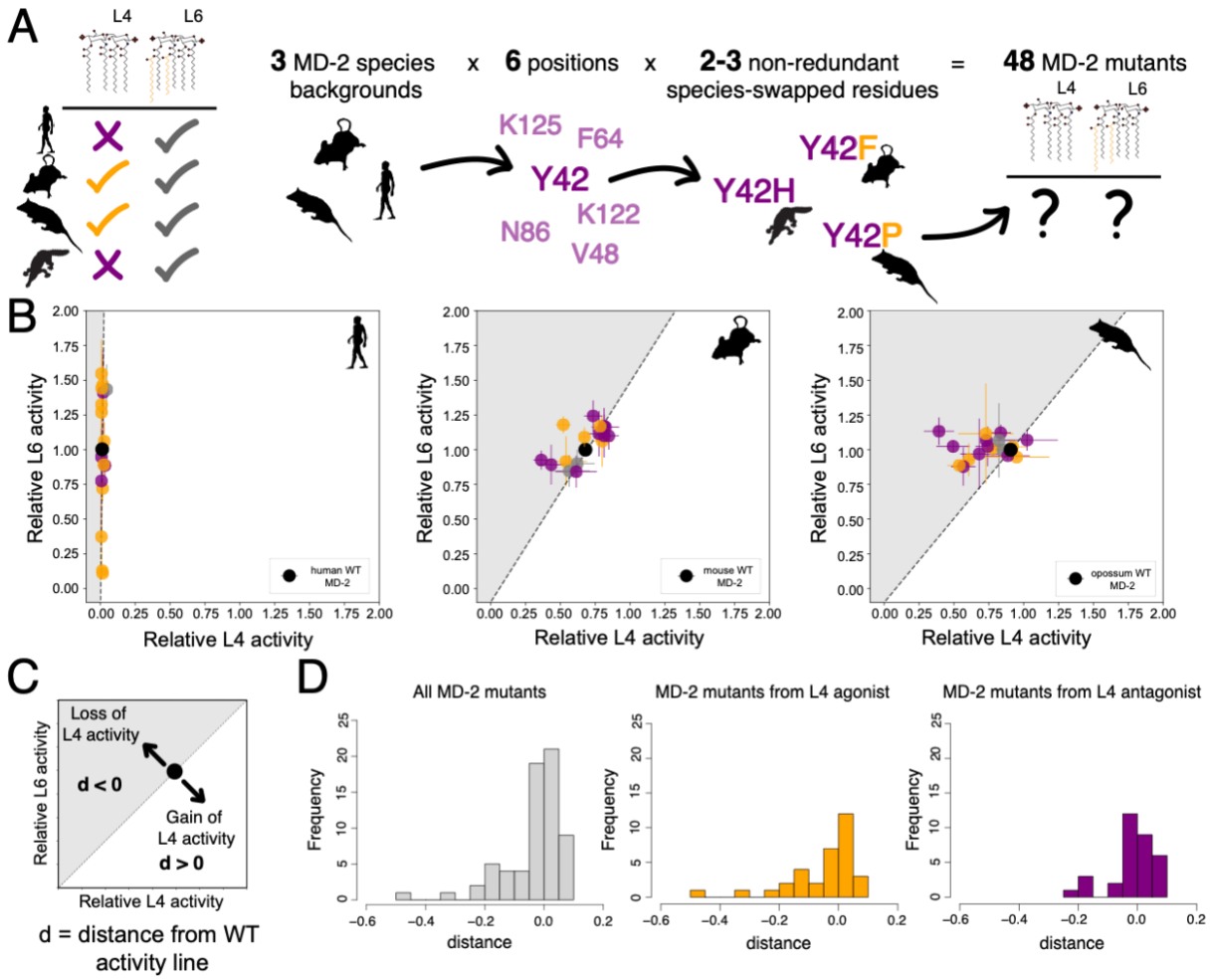

Figure 3.5. Mutations to MD-2 reveal robustness of L6 activation, fragility of L4 activation. (A) Schematic of species-swapping MD-2 mutations introduced into human, mouse, and opossum MD-2 backgrounds. Residues originating from species that are agonized by L4 (mouse and opossum) shown in orange; residues originating from species that are antagonized by L4 (human and platypus) shown in purple. (B) Relative ratio of L6 activation versus L4 activation. Wildtype L6:L4 ratio for each species shown as a dashed line, with ratio of L4 activation to normalized relative activity of L6 (1) shown as a black dot. The grey region to the left of the wildtype line denotes that a mutation introduced a relative loss in L4 activation with respect to wildtype. (C) Schematic describing the metric for measuring the difference in L4 activation an MD-2 mutation yielded. Movement along wildtype L6:L4 ratio line denotes global increase or decrease in overall ability of TLR4/MD-2 to activate. Distance to the left of the wildtype line ($d < 0$) denotes loss in L4 activity with respect to L6 activity, whereas distance to the right if the wildtype line ($d > 0$) denotes increase in L4 activity with respect to L6 activity. (D) Histograms of distances from (left panel) all data, (middle panel) MD-2 mutations originating from a species agonized by L4, and (right panel) MD-2 mutations originating from a species antagonized by L4. Two-sample Kolmogorov-Smirnov (KS) tests show no significant difference between distributions of middle and right panel (p-value = 0.228), or either subplot with total data ($p = 0.853$ for total versus L4 agonists, $p = 0.529$ for total versus L4 antagonists).

We also observed the distribution of distances from the wildtype ratio for mutations originating from L4 agonists versus L4 antagonists (Fig. 3.5D). Using the Kolmogorov-Smirnov Test, we found that there was no significant difference in these distributions ($p = 0.452$). This further suggests that there is no unique underlying mechanism by which species' origin of MD-2 mutation specifically affects L4 specificity. The result that origin of mutation was not a significant factor suggests that different species along their evolutionary lineages optimize L4 specificity for their own bacterial landscape through a variety of sequence-level and structural features, which are not interchangeable and often deleterious between species.

It is possible that selection does drive L4 specificity. However, we propose that L4 antagonism was likely the default ancestral state due to L4 agonism being a more difficult biophysical challenge and thus constraining for evolution. In this case, selection pressure contributed to certain lineages to evolve L4 agonism.

Is the strength of the binding interaction responsible for measured activation of the TLR4 complex?

In the mouse MD-2 background, mutation of E122K (to the human residue) resulted in no significant change to L6 activation but the most significant loss of L4 activation, with a drop from 0.669 to 0.458, or loss of ~1/3 of wildtype activity. Given that there is crystallographic information for how L6 and L4 bind to the mouse TLR4/MD-2 receptor and forms an activated tetramer, we were interested in further studying the structural role this mutation plays in forming contacts for dimerization.

MD-2 mutation E122K has been identified previously as a critical position for maintaining or breaking activity with hypo-acylated endotoxin in human and mouse, and appears

to have the largest effect size of sites under diversifying selection.[42] Other papers have suggested the biophysical role this residue plays in both species' backgrounds: in human MD-2, positively-charged K122 interacts with the negatively charged 4' phosphate on both penta- and tetra-acylated LPS (~180° opposite orientation of hexa-acylated LPS) to draw these structures closer into the MD-2 binding pocket, minimizing the contacts endotoxin can make with dimerizing TLR4/MD-2 and preventing effective dimerization.[125] Mouse TLR4/MD-2 natively express E122. Introduction of E122K into the mouse background has the opposite effect, where L4 activity is greatly diminished.

We hypothesized that different agonists for the TLR4 receptor display greater or weaker activity due to the amount of atomic contacts they induce in the dimerization interface. That is to say, there would be more contacts at the wildtype dimerization interface (i.e. between dimerizing TLR4 and MD-2:ligand) when bound to L6 versus L4, given that mouse activates with L4 at ~70% of what it activates with L6 (Fig. 3.2A, 3.5B). Further, we hypothesized that the E122K mutation would appreciably reduce the number of contacts formed between dimerizing TLR4 and MD-2ΔE122K:L4 while not inducing a large change between dimerizing TLR4 and MD-2ΔE122K:L6.

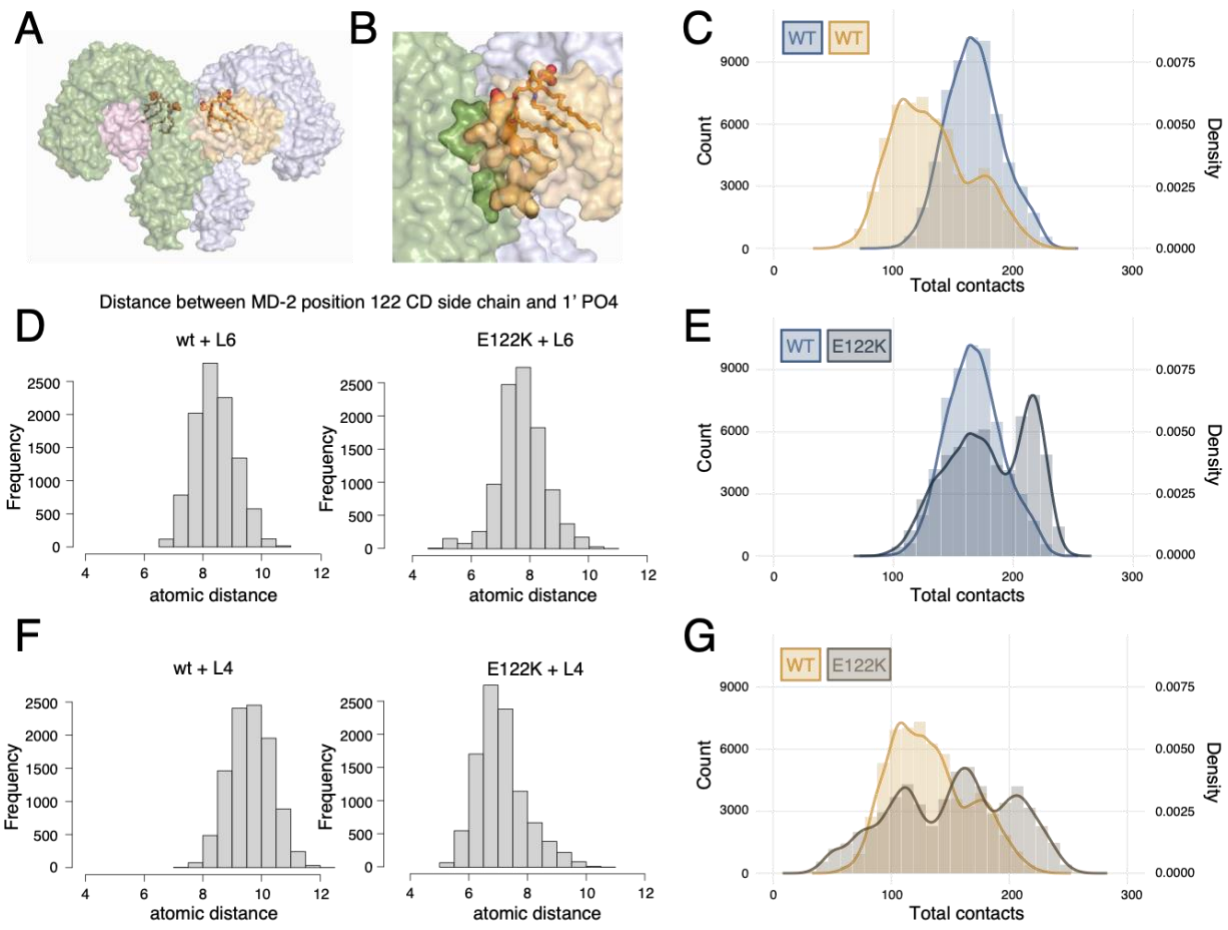

Figure 3.6. Hypo-acylated LPS induces fewer contacts in wildtype and E122K mouse TLR4/MD-2 dimerization interface *in silico*. (A) Representative crystal structure of wildtype mouse TLR4/MD-2 bound to lipid IVa (PDB: 3VQ1) that was used as an initial pose for MD simulations.[119] (B) Close-up of dimerization interface shown as opaque surfaces. MD-2 (light orange) and ligand (orange) atomic contacts within 5 angstroms of the dimerizing TLR4 (green) were counted for each frame of each trajectory. (C) Total MD-2 and ligand atomic contacts maintained over 100 ns MD simulations for wildtype with L6 and L4. (D) Distance (in angstroms) between the CD atom of glutamate (in wildtype E122) or lysine (in mutant E122K) and the 1' phosphate of L6. (E) Total MD-2 and ligand atomic contacts maintained by wildtype or E122K with L6. (F) Distance (in angstroms) between the CD atom of glutamate (in wildtype E122) or lysine (in mutant E122K) and the 1' phosphate of L4. (G) Total MD-2 and ligand atomic contacts maintained by wildtype or E122K with L4.

L4 exhibits greater mobility in the MD-2 binding pocket than L6.

To investigate this, we performed molecular dynamics (MD) simulations of the tetrameric receptor, including the entire extracellular portion of mouse TLR4 and MD-2 as well

as MD-2ΔE122K:ligand (i.e. lipid IVa or hexa-acylated LPS from *E. coli*) to observe structural changes that occur in these two bound structures (Fig. 3.6A). Simulations were started from crystal structures as the starting poses.[119] These simulations were allowed to proceed for 100ns to observe initial relaxation with ligand bound (see Methods and Materials). We then observed the number of atomic contacts that were maintained at a 5-angstrom radius during the trajectory between both MD-2 and bound ligand to the dimerizing TLR4 (Fig. 3.6B). We collected these values for each dimerizing partner in the tetramer, as the dimerization of the complex involves the dimerization of two heterodimers.

We found that overall, our prediction was correct - the L6-bound complex maintained more contacts with the dimerizing TLR4 than the L4-bound complex (Fig. 3.6C). The total number of contacts between MD-2:L6 and dimerizing TLR4 appear to fluctuate around an average value (~160 atomic contacts), suggesting that there is a primary, most stable binding conformation that is induced by L6. Contacts formed by the dimerizing TLR4 and L6 alone appeared to have a highly likely conformation, as well as a less common but tighter binding interaction. These are not offset by MD-2 binding. It is possible that these additional contacts by L6 are the mobile R2' acyl chain adjusting slightly to be more or less exposed outside of the MD-2 binding pocket.

Meanwhile, L4 induced very different contacts in the tetrameric mouse complex. Overall, L4 on average induces fewer contacts, both contributing from MD-2 and ligand, than L6 does (Fig. 3.6C). There appears to be bimodal binding interactions between both MD-2 and L4 with the dimerizing TLR4. Looking at individual replicates, it does not appear that MD-2 contacts necessarily compensate for L4 contacts when L4 is more hidden in the MD-2 binding pocket.

The observation that multiple states exist suggest that L4 is much more mobile and thus unstable in its binding interaction with TLR4/MD-2.

MD-2 mutant E122K destabilizes the TLR4/MD-2 complex in a ligand-dependent manner.

We next wanted to investigate the potential mechanistic and structural change induced by the MD-2 mutant E122K, which was shown to have a deleterious effect on L4 activity but had no significant effect on L6 activity (Fig. 3.5B). We found that, compared to the wildtype L4-bound structure, E122K:L4 demonstrated up to four different binding states, with a wider range of contacts formed than wildtype+L4 (Fig. 3.6G). Interestingly, the contacts formed between wildtype+L4 and E122K+L4 were not significantly different (peaks for wt:L4 at 110 and 175; peaks for E122K:L4 at 80, 110, 162, and 207), but the mutation appeared to shift the complex towards favoring a binding state with more contacts between the dimerizing TLR4 and L4. Similarly to total contacts, there appeared to be a broader range of contacts formed between E122K:L4 MD-2 versus the wiltype+L4 MD2.

Conversely, the E122K mutant with L6 appeared to have also introduced additional contacts between MD-2:L6 and the dimerizing TLR4, as seen by the shift towards more contacts with the dimerizing TLR4 and a bimodal distribution (peak for wt:L6 at 164; peak at E122K:L6 at 176 and 217). This was primarily observed as a great increase in contacts between L6 and the dimerizing TLR4 (Fig. 3.6E).

Taken together, the MD-2 E122K mutation induces instability that reduces contacts with the L4-bound structure but creates contacts with the L6 structure. In conjunction with functional data of this mutation (Fig. 3.5B), these results suggest that the ability of the lipid A portion of

endotoxin to activate the TLR4/MD-2 is highly dependent on its ability to stably maintain sufficient contacts between MD-2:ligand and the dimerizing TLR4.

MD-2 position 122 interacts with the 1' PO4- of lipid IVa and hexa-acylated LPS

Previous papers have suggested that E122 repulses the negative charge on the 1' PO4 of endotoxin that sticks out of the MD-2 binding pocket, and that this repulsion helps orient endotoxin from sticking too deeply in the MD-2 pocket and thus is accessible to the dimerization interface.[42,119] On the other hand, K122 in the human background draws in the 4' PO4 of lipid IVa which shields lipid IVa from interacting with dimerizing TLR4. Given that the primary contact position 122 makes is with this phosphate, we were curious to know how this interaction was altered by mutation. We measured the distance of the side chain delta carbon (CD) of the native E122 and mutant E122K, respectively, with the phosphate on 1' PO4 (Fig. 3.6D,F).

In the wildtype background, we found that the negatively charged glutamate does indeed repulse the 1' PO4- from either L4 and L6. Given that a carbonyl oxygen would extend about 1.2-1.4 angstroms further, and 1.4-1.6 angstroms for the phosphate – oxygen interaction, a distance of 7 angstroms would allow for a hydrogen bond of less than 5 angstroms in distance to occur. Given that distances of no less than 7 angstroms were observed in the wildtype E122 with the L4 1' PO4', and <10% with L6, the interaction between this residue and the 1' PO4- is considered to be insignificant.

In the mutant E122K background, a much closer interaction between E122K and L4 or L6 is observed. A significant change is observed between the wildtype versus the mutant E122K background with L4, with an average distance of ~9.5A observed in wt+L4 versus ~6.2A in E122K+L4. Meanwhile, the difference in the binding interaction between the wildtype E122 and

mutant E122K background with L6 is much less pronounced – the average goes from ~8.2 to ~7.2A. This is likely due to the fact that L6 doesn't have much room to move in the hydrophobic pocket on MD-2, while L4 has much more room to move in the pocket. We interpret this decrease in distance when E122K is introduced with L4 to mean that E122K encourages L4 to bind deeply in the pocket by forming a highly favorable hydrogen bond with the 1' PO4-. The change induced by this mutation is much more pronounced when L4 is bound due to L4 having much more room to interact in the binding pocket. This is further evidence that L4 agonism is more biochemically challenging to maintain, especially if L6 agonism is what is primarily selected for.

Discussion

In this work, we investigated the evolutionary history and structural basis of antagonism for hypo-acylated LPS in sauropsids. Overall, we found that human TLR4/MD-2 did not lose agonism for smaller endotoxin, but rather antagonism for hypo-acylated LPS is a feature that has appeared repeatedly in evolution. Work from reconstructed ancestors of Euarchontoglires (including human and mouse) and placental mammals suggest that this antagonism was the ancestral state in recent (~80 million years) history, and may have appeared even earlier given that platypus also displays hypo-acylated LPS antagonism. Agonism for hexa- and hepta-acylated LPS has been conserved in all modern-day creatures studied, and agonism for larger endotoxin does not correlate with agonism for smaller endotoxin. It is unclear what the significance of the cutoff between hexa- and hepta-acylated versus hypo-acylated endotoxin is; Gram-negative bacteria containing all numbers of acyl chains are represented throughout all of the bacterial species tree, although 4-7 acyl chains are most common.[113] Still, there was a non-

significant difference in L4 and L5 activation in any creature tested, whether agonistic or antagonistic.

We were then curious if we could identify sequence-level features that explained why some species are agonized by hypo-acylated LPS while others are antagonized. A previous mutagenesis study identified seven sites in TLR4 and MD-2 that are sufficient to switch between species-specific L4 agonism versus antagonism in the human and mouse background.[42] Using this model, we compared the sequences of all modern-day species we characterized, in addition to three reconstructed ancestors. We found that the sites and amino acids utilized to achieve this specificity vary in different lineages, and that there was not an obvious pattern of necessary amino acids at certain sites to confer hypo-acylated LPS agonism or antagonism. For example, we found that opossum shares three of the human amino acids at these sites and none of the mouse amino acids, even though both of these creatures are agonized by L4 (Fig. 2D). Further, a reconstructed ancestor of all mammals shares all seven amino acids with human at these key sites, yet displays L4 agonism and L6 antagonism, which is opposite to human. We conclude that this model does not broadly describe specificity for hypo-acylated LPS, and that other sequence features were necessary to evolve L4 specificity. This drove us to investigate a different hypothesis – that perhaps sites that are rapidly diversifying are responsible for the changes we observed in L4 specificity.

Using a variety of selection models including PAML, FEL, FUBAR, and MEME, we identified a collection of sites on TLR4 and MD-2 in mammals that are under positive (diversifying) selection. Interestingly, we identified a region on TLR4 (residues 291-376) that had the highest rate of evolution. We have previously performed MD simulations with hexa-acylated LPS, including a long O-antigen, which sampled the contacts on human TLR4 and MD-

2 that interact with the O-antigen. We found great overlap on TLR4 between this fastest-evolving region and the surface where the O-antigen interacts. We thus concluded that this region of TLR4 diversifies rapidly in order to recognize the O-antigen of endotoxin, which correspondingly evolves rapidly to evade TLR4 detection.[126] Because TLR4 minimally interacts with the lipid A moiety of LPS including the acyl chains, and lipid IVa does not have core or O-antigen moieties, we decided to focus on MD-2 in our study on selection as a driving force for L4 specificity. We identified six MD-2 sites under positive selection to perturb in two natively L4-agonized (mouse and opossum) and one natively L4 antagonized (human) species to investigate the role of selection on ligand specificity.

We emphasize here that diversifying selection in co-receptor MD-2 is of critical interest for the evolution and function of the TLR4/MD-2 receptor. This is especially because, out of the 13 TLRs identified in mammals, all other TLRs directly bind their ligands.[9] It is interesting that TLR4 has previously been found in primates to have a higher rate of diversification compared to 10 other TLRs, given that TLR4 itself does not constitute the primary binding pocket for endotoxin.[121]

From our MD-2 mutagenesis studies, we made several striking findings. First, while it was easy to damage L4 agonism with a single mutation, it was impossible to create or improve L4 activation alone (i.e. without proportionally affecting L6 activity) with a single mutation. Second, the origin of the mutation (from an L4-agonized or L4-antagonized specie) didn't matter in terms of the effect a mutation would have on L4 activity. This suggests that species that are agonized by L4 (mouse and opossum) evolve different sequence and structural features along their lineages to arrive at this specificity, and that these features are not interchangeable between

species. We identified mouse MD-2 E122K as an attractive mutation to investigate further, as this mutation did not affect L6 activation but did appreciably reduce L4 activation.

We next performed MD simulations to better understand the molecular basis by which L4 and L6 interact in the TLR4/MD-2 binding pocket, as well as identify any possible changes in the binding interaction induced by the MD-2 mutation E122K with L4 but to a lesser or no degree, L6. We indeed validated through this *in silico* method that in wildtype mouse TLR4/MD-2, binding of L4 maintains fewer contacts in the dimerization interface than L6. We then found that E122K broadens the amount of possible contacts that are formed when L4 is bound, including a shift towards fewer atomic contacts being populated. This is suggestive of a less specific interaction between dimerizing TLR4 and MD-2ΔE122K:L4, and may provide a molecular explanation for the functional loss of L4 activation we observed from this mutation.

Our mutagenesis functional studies and MD simulations suggest that L4 agonism is biophysically difficult to maintain. This is unsurprising in the context that L4 does not fully occupy the MD-2 binding pocket. ~5 acyl chains from L6 can fill the hydrophobic binding pocket of MD-2 and still expose ~1 acyl chain to form the dimerization interface. Additional structural features are necessary to orient L4 such that the hydrophobic acyl chains are not fully and preferentially concealed in the hydrophobic pocket of MD-2, but rather out and available for interacting with the dimerizing TLR4. These structural features must also not conflict with structural features that are necessary for L6 agonism.

Our evolutionary studies further suggest that L4 antagonism was the default, ancestral state of the TLR4 receptor, but that selection pressure along certain lineages can switch specificity towards L4 agonism. While we did not investigate combinations of MD-2 mutations, the observation that no single mutation increased L4 activity suggests that evolutionary

trajectories for L4 agonism are somewhat constrained. Conversely, the ability for single mutations to decrease L4 activation suggests that L4 agonism is easily lost during evolution if selection pressure is not maintained.

In the broader scope of structural-functional evolution, we were curious to know if these ligand-specificity tradeoffs are observed in other protein phylogenies. Ultimately, it appears to be a feature of TLR4/MD-2 to have highly conserved and robust specificity for some ligands while occasionally allowing specificity for smaller ligands. Given the life-preserving function of this receptor complex to prevent deadly infection and/or sepsis, it seems reasonable that single mutations are not sufficient to create off-target specificity for ligands derived from non-fatal bacteria. Other receptors with multiple ligands have been observed to easily toggle between specificity for different ligands, although the function of these receptors is less detrimental to individual fitness and can more safely explore their specificity landscape.

Materials and Methods

Plasmid Construction and Mutagenesis

All TLR4, MD-2, and CD14 genes were purchased or gifted in the pcDNA3.1(+) backbone, following a constitutive CMV promoter (see Table 3.1). We used the pGL3-elam-luc plasmid which encodes a firefly luciferase gene downstream of an NF-κB promoter to measure TLR4/MD-2 mediated NF-κB activation. The pRL-TK plasmid encoding constitutively active renilla was used to measure total cell count.

Site-directed mutagenesis was performed on human, mouse, opossum, and platypus MD-2 constructs using recommended product specifications (New England Biolabs). Primers were designed using the NEBaseChanger online server and ordered from Eurofins. Plasmids were re-

circularized using KLD enzyme mix (New England Biolabs). Mutants were validated using Sanger sequencing (Azenta).

Functional Activity Assay

We measured activation of TLR4/MD-2 via a well-established NF-κB assay in HEK293T cells as previously described[33,116,127] Cells were passaged up to 30 times and maintained in DMEM with the addition of 10% FBS and Antibiotic-Antimycotic, at 37°C with 5% CO_2. For each experiment, 135 µL of cells at ~25% confluency were seeded in a 96-well plate. Each well of cells were then transiently co-transfected with 65µL aliquots of 100ng plasmids consisting of 10ng TLR4, 1ng MD-2, 1 ng CD14, 20ng pGL3-elam-luc, 1ng Renilla pRL-TK, and 67ng empty pcDNA3.1(+) vector diluted in Opti-MEM (ThermoFisher Scientific). PLUS and lipofectamine were used as recommended to enhance transfection efficiency (ThermoFisher Scientific). Endotoxin samples were diluted in endotoxin-free water and sonicated in a jewelry ultrasonicator for 15 minutes at room temperature at 180 W immediately before use in order to uniformly disrupt micelles. Transfected cells were incubated for 20-24 hours. After this, transfection mix was removed and replaced with 100µL per well treatments including biologically-relevant concentrations of 200 ng of L5, L6, or L7, or 2 µg of L4 diluted in 25 µL endotoxin-free PBS and 75 µL serum-free DMEM. Cells were incubated with treatments for 3 hours, then activity for each treated well was assessed using the Dual-Glo® Luciferase Assay System (Promega) with the SpectraMax i3 plate reader. All treatment conditions were tested in technical triplicate wells. Raw luciferase values per well were corrected for cell count by dividing luciferase by renilla values to yield a "relative activity" value per well. The relative activity for technical triplicate wells was averaged and considered as one biological replicate. All data presented were done in

(at least) biological triplicate. Biological replicates for each species' TLR4/MD-2/CD14 always contained the following controls: triplicate wells treated with mock treatment (no endotoxin added) for background subtraction, and 200 ng per well of L6, which was normalized to 1 and used as a normalization factor for all other treatments.

Dose Response Curves

Dose response curves for ligand were done using concentrations that were 1/625X, 1/125X, 1/25X, 1/5X, 1X, and 5X the standard dosage (described above). The Hill equation was then used to fit a curve to functional data in order to calculate EC_{50} and maximum activation values for each ligand, for each species' TLR4/MD-2 receptor complex.[128] Parameters for EC_{50} and maximum activity were optimized using the Python scipy.optimize.least_squares package fitted to Equation 1.[129]

$$relative\ activity\ = \frac{[LPS]}{([LPS] + EC_{50})}\, x\ maximum\ activity + background$$

MD-2 Mutant Analysis

Transfection of MD-2 mutants was done as described above but with 1 ng of MD-2 mutant plasmid instead of wildtype. All MD-2 mutants were then treated (per well) with: 200 ng of L6, 2 μg of L4, 200 ng L6 + 2 μg L4, and mock treatment. Biological replicates of MD-2 mutants were done in tandem (i.e. on the same 96-well plate) with their respective wildtype species' background.

Data from MD-2 mutants was normalized to the relative activity of their respective species' wildtype TLR4/MD-2/CD14 after background subtraction and treatment with L6. We

then compared the specificity of MD-2 mutants for L4 versus L6. To do this, we first calculated a ratio of L4 versus L6 for the wildtype TLR4 receptor for each species.

$$WT\ L4{:}\,L6\ ratio\ = \frac{relative\ WT\ activity\ with\ L4}{relative\ WT\ activity\ with\ L6}$$

We considered this ratio to be the specificity of each species' TLR4 complex to activate with L4 versus L6. We then measured the relative activity of each mutant MD-2 with L4 and L6, with respect to wildtype L6 activation.

$$relative\ mut\ activity\ with\ [L4\ or\ L6]\ = \frac{relative\ mut\ activity\ with\ [L4\ or\ L6]}{relative\ WT\ activity\ with\ L6}$$

The ratio of activation with L6 versus L4 was calculated for each wildtype species' TLR4/MD-2/CD14. The change in specificity, d, for L4 was then calculated as the difference of a mutant L6/L4 ratio from the wildtype ratio line (Fig. 5C). Mathematically, this is calculated as the perpendicular distance that a mutant MD-2 L6/L4 has from the wildtype ratio line.

$$change\ in\ specificity, d\ = \frac{\frac{mut\ L4}{WT\ L4} - \frac{mut\ L6}{WT\ L6}}{\sqrt{1 + \left(\frac{wt\ L6}{wt\ L4}\right)^2}}$$

Statistical analyses (three-way ANOVA and two-sample Kolmogorov-Smirnov tests) were done using R code.

Ancestral Sequence Reconstruction

Ancestral sequence reconstruction was done previously to produce full-length reconstructions of TLR4, MD-2, and CD14 from the Euarchontoglires, eutherian, and therian ancestors.[116,130] These ancestors were human codon-optimized and purchased subcloned into the pcDNA3.1(+) vector backbone (GenScript).

Evolutionary studies (HyPhy and PAML)

Multiple sequence alignment

Full-length TLR4 and MD-2 DNA sequences were obtained from the NCBI Orthologs database, filtering for mammalian sequences only. Of these, 20 sequences were chosen that span major clades of the mammalian species tree including primates, placentals, marsupials, and monotremes, with slight emphasis on primate sequences. The same species were selected for both TLR4 and MD-2 alignments. DNA sequences for each gene were then visualized in AliView and aligned using Muscle5.[77] The alignments were further processed to remove extensions at termini, insertions, and stop codons. In-frame triplicate columns (i.e. codons) that had less than 75% sequence coverage (i.e. more than 5 sequences out of 20 were missing a codon) in the alignment were removed. The final trimmed sequences were then aligned again to remove all gaps. These alignments were used for downstream evolutionary analyses.

PAML

The PAML NS sites program was used for TLR4 and MD-2 coding sequences to test for evidence of positive selection using the codon model F3x4, as described previously.[57,127,131] A species tree was generated using the PhyML command line tool with the GTR substitution model and 1,000 bootstraps. The M1+M2 and M7+M8 models of selection were used in likelihood ratio tests to determine the likelihood of genes evolving under positive selection.[132] Naïve Empirical Bayes (NEB) and Bayes Empirical Bayes (BEB) were tested to identify sites on TLR4 and MD-2 under positive selection, as described previously.[1] We proceeded with the BEB for selection analysis using the NS sites package for evolutionary Model 2 for both TLR4 and MD-2. Sites

with a posterior probability > 0.95 were considered to be strongly supported as under positive selection.[57]

HyPHY analyses

We constructed our own phylogenetic trees (including selection of an evolutionary model and branch lengths) using RaxML-ng for command line, which is a species-tree informed ML phylogenetic tree algorithm.[78,133] We used the "--opt-model" and "--opt-branches" options to optimize model parameters and branch lengths, respectively. RaxML found that the JTT+G8m and JTT evolutionary models were best via AIC for TLR4 and MD-2, respectively. We asserted these phylogenetic trees in all HyPhy analyses. Default trees estimated by HyPhy for TLR4 and MD-2 yielded trees that are unlikely given the known species tree, such as including several unrealistic duplications (data unshown). We ran HyPhy analyses on the command line.

GARD

We used GARD to detect breakpoints in TLR4 and MD-2.[134] While GARD breakpoints suggest genetic regions of recombination, they can also identify regions where the rate of evolution differs in a gene sequence. Following this analysis, only the TLR4 ectodomain (composing three different partitions; amino acids 0-290, 291-375, and 375-608) were used for downstream HyPhy analyses. As MD-2 was not found to have any breakpoints, we proceeded with downstream HyPhy analyses (FEL, FUBAR, and MEME) with the processed alignment.

FEL/FUBAR/MEME

Each of these analyses identify sites under selection implementing different assumptions. FEL and FUBAR are capable of identifying sites under positive or negative selection, while MEME can only identify sites under positive selection. Sites with a posterior probability > 0.9 were accepted for FUBAR as under positive selection. Sites with a p-value < 0.1 were accepted as under diversifying selection with MEME and FEL. R code was used to visually summarize results.[131,135]

Molecular dynamics simulations parameters

Molecular dynamics simulations were done as described previously.[127] For all simulations, we used GROMACS 2023 with the CHARMM36 2021 forcefield and TIP3P waters.[1,2] We generated LPS and lipid IVa coordinates and forcefield parameters using LPS Modeler as implemented in CHARMM-GUI, and aligned these structures to the mouse TLR4/MD-2 bound structures with LPS and lipid IVa (3VQ2 and 3VQ1, respectively). These bound structures were used as the initial poses to perform simulations.[119] PyMOL was used to visualize and prepare initial structures. The MD-2 E122K mutation was introduced *in silico* using the PyMOL mutagenesis function, and initial side chain rotamer pose was selected referring to the native human MD-2 K122 (PDB: 3FXI) rotamer, and ensuring the rotamer did not introduce unreasonable steric hindrance.[11] The structures for lipid IVa and hexa-acylated LPS with their respective GROMACS parameter files were obtained from CHARMM GUI.[138] Simulations were run as described previously in triplicate for both wildtype mTLR4/MD-2:lipidIVa and mutant mTLR4/MD-2ΔE122K:lipidIVa for 100ns. The MDAnalysis Python package was used to analyze simulations.[139,140] The University of Oregon Talapas High Performance Computing Cluster was used to run these simulations.

Table 3.1. Reagents used in Chapter III.

Reagent	Source/reference	Identifier/Additional Information
Human TLR4 plasmid	gift from Ruslan Medzhitov	RRID: Addgene 13086
Human MD-2 plasmid	Genscript	LY96_OHu26610C_pcDNA3.1(+)
Human CD14 plasmid	gift from Doug Golenbock	RRID: Addgene 13645
pGL3-ELAM-luc	gift from Doug Golenbock	RRID: Addgene 13029
Renilla pRL-TK	Promega	E2241
pcDNA3.1(+)	Addgene	V790-20
Mouse TLR4 plasmid	gift from Doug Golenbock	RRID: Addgene #13086
Mouse MD-2 plasmid	Genscript	UniProt #Q9JHF9 in pcDNA3.1(+) backbone
Mouse CD14 plasmid	Genscript	UniProt #P10810 in pcDNA3.1(+) backbone
Opossum TLR4	Genscript	UniProt #F6Y6W8 in pcDNA3.1(+) backbone
Opossum MD-2	Genscript	UniProt #F6QBE6 in pcDNA3.1(+) backbone
Opossum CD14	Genscript	NCB Accession #XP_007473804.1 in pcDNA3.1(+) backbone
Platypus TLR4	Genscript	GenEZ ORF Clone: TLR4_OOc01166C_pcDNA3.1(+)
Platypus MD-2	Genscript	NCBI reference: XP_028925171.1 (subcloned into pcDNA3.1(+))
Platypus CD14	Genscript	GenEZ ORF Clone: CD14_OOc137521C_pcDNA3.1(+)
LPS-ST (hepta-acylated LPS; L7)	Sigma Aldrich	L6143-1MG; from *Salmonella enterica* serotype typhimurium
LPS (hexa-acylated; L6)	Invivogen	tlrl-peklps; rough (no O-antigen) *E. coli*

LPS-RS (penta-acylated LPS; L5)	Invivogen	tlrl-rslps; *Rhodobacter Sphaeroides*
Lipid IVa (tetra-acylated LPS; L4)	Biosynth	CLP-24006-S
Q5® Hot Start High-Fidelity 2X Master Mix	New England Biolabs	M0494L
KLD Enzyme Mix	New England Biolabs	M0554S
DMEM	ThermoFisher	11995-073
FBS	ThermoFisher	A5669701
Antibiotic antimycotic	Gibco	1524006
PLUS	ThermoFisher Scientific	11514-015
Lipofectamine	ThermoFisher Scientific	18324-012
Opti-MEM	ThermoFisher Scientific	11058-021
Endotoxin-free PBS	Sigma	TMS-012-A
Dual-Glo® Luciferase Assay System	Promega	E2940

Bridge to Chapter IV

In Chapter III, I discussed how TLR4/MD-2 in sauropsids evolved specificity for hypo-acylated LPS – MD-2 primarily evolves in a highly species-specific manner to maintain a network of contacts to orient L4 into a productive orientation to drive dimerization, and that L4 agonism is easily lost and but difficult to create with single mutations. LPS is a MAMP, and is the canonical ligand that TLR4/MD-2 likely first evolved to recognize. However, TLR4/MD-2 can bind to and induce inflammation with other ligands. In Chapter IV, I introduce work I've done to describe the binding mechanism and evolution of TLR4/MD-2 with an endogenous ligand, S100A9. I also discuss complications with identifying a unique binding interface between

S100A9 and TLR4/MD-2, given that there is likely significant overlap between LPS and S100A9 binding on TLR4/MD-2.

CHAPTER IV

INSIGHTS ON A DIRECT BINDING MECHANISM BETWEEN S100A9 AND TLR4

INTRODUCTION

The immune response recruits cellular machinery to sites of tissue damage, which can be life-saving in both infection from foreign pathogens as well as internal stress or damage in a host. Damage-associated molecular patterns (DAMPs) are endogenous "self" molecules that are released or created (e.g. from sheared or necrotic cells, from a perturbation to homeostatic biological pathways, etc.), and can be recognized by the immune system to initiate an inflammatory response (reviewed here[19]). Successful engagement of the immune system in response to DAMPs is timely and highly regulated to clear the host from damage. However, a prolonged or overactive response can lead to chronic inflammatory diseases that afflict human health. Arthritis, atherosclerosis, inflammatory bowel disease, cardiovascular disease, and others are examples of diseases that are correlated with a dysregulated and persistent immune response to these DAMPs.[141] The World Health Organization reported that heart disease was the global leading cause of death in 2021, accounting for 13% of the world's total deaths.[142] Other diseases in the top 10 list of causes for death which are implicated with immune dysregulation were lower respiratory infections, lung cancer, kidney diseases, and Alzheimer's. Understanding the mechanism by which the immune system interacts with DAMPs is essential for generating therapeutics to help manage and treat these diseases.[143]

TLR4/MD-2 is an innate immune receptor that recognizes and initiates inflammation in response to DAMPs. One such DAMP is the calcium-binding protein, S100A9 (A9). This protein is expressed in many cell types but is especially abundant in macrophages. Although A9 normally performs other housekeeping functions in healthy cells, it can spill out of damaged or

dying cells into the extracellular matrix, making it accessible to interact with surface TLR4/MD-2 on immune cells. It has been well-documented that an increase in A9 is found around sites of tissue damage in several chronic inflammatory diseases such as arthritis, atherosclerosis, cardiovascular disease, and several studies have validated that this inflammation is TLR4/MD-2 dependent. Thus, creating therapeutics that block the interaction between A9 and TLR4 could improve the quality and span of human lives globally.

Despite the great potential A9 therapeutics could serve for human health, the mechanism by which A9 interacts with and activates TLR4/MD-2 has not been fully described (Fig. 4.1A).[144] Notably, TLR4/MD-2 only have one known ligand binding site – the hydrophobic pocket of MD-2. Binding of the canonical ligand, LPS, into the MD-2 binding pocket creates complementary hydrophobic and charged regions in the dimerization interface that ultimately bring the receptor complex together.[11] Confoundingly, A9 in its native, homodimeric form cannot fit into this pocket. Several groups have brought forth evidence that there is likely a direct interaction between A9 and TLR4/MD-2. For example, previous work has shown that A9 colocalizes with TLR4 and the accessory protein CD14 in live monocytes cells using gold-bead labeling.[145] Other studies have shown using surface plasmon resonance that A9 binds to TLR4/MD-2 with a Kd of 2.1 nM.[146] Yet another study found that peptide fragments of A9 are sufficient to bind to TLR4/MD-2 using mass spectrometry.[26] Despite an abundance of evidence that of A9 forms a direct binding interaction with TLR4/MD-2, a binding site on TLR4/MD-2 for A9 has yet to be found. Further, it is unclear if this interaction is necessary for A9 to drive dimerization and activation of TLR4/MD-2.

The only TLR4/MD-2 ligands for which there is a known binding site supported by crystallographic evidence is hexa-acylated LPS, penta-acylated Eritoran, and tetra-acylated lipid

IVa.[11,119,125] As described in the previous chapter, the canonical binding site on TLR4/MD-2 is the hydrophobic binding pocket of MD-2, which contains a small, highly hydrophobic volume. There are also many other DAMPs that have been reported as mediators of inflammation in a TLR4/MD-2 dependent manner, but it is unclear what their binding interaction with TLR4/MD-2 is. Notably, many DAMPs vary in size and charge and cannot physically fit into the MD-2 binding pocket. A pressing question for endogenous activation of TLR4 activation is whether these DAMPs utilize different mechanisms to activate TLR4, or if some share mechanisms for activation. Thus, elucidation of an interaction between A9 and TLR4 could open the door for defining other important DAMP interactions with TLR4.

Here, we provide evidence that A9 does not bind in the canonical MD-2 binding site via competition assays with lipid IVa, an antagonist of human TLR4/MD-2. We used MALDI-TOF mass spectrometry to investigate the binding stoichiometry between A9 and TLR4/MD-2, although these results were inconclusive. We then used both Rosetta ligand docking and AlphaFold3 to generate several possible binding models, allowing me to create lists of candidate sites that can be later explored to identify a binding interaction between A9 and TLR4. Finally, we characterized mutations probing a most-likely model, a top-binding model for A9 binding with TLR4/MD-2. Although we have not confidently identified a binding mechanism yet, the work provided here paves the way for future study.

Results

A9 interacts with TLR4/MD-2 via different binding site than LPS

We first tested the simple hypothesis that A9 shared a binding site with endotoxin. To do so, we performed a competition experiment, measuring the ability of A9 to activate in the

presence of increasing amounts of lipid IVa, a tetra-acylated LPS antagonist that binds in the MD-2 binding pocket. If the A9 binding site overlaps with the lipid IVa binding site, we predicted competitive inhibition, with the apparent EC_{50} for A9 activation shifting to the right upon addition of lipid IVa. If they use different binding sites, we predicted no inhibition. This is because lipid IVa inhibits TLR4/MD-2 activation by non-productively binding in the MD-2 binding pocket—not by blocking dimerization from occurring. If A9 promotes dimerization via a different site or different mechanism, this could occur independently of the occupancy of the MD-2 binding pocket.

We measured the activity of increasing amounts of A9 in the presence of increasing amounts of lipid IVa (Fig 4.1B). As we have seen previously, A9 has an EC_{50} of about 1 μM. We then added increasing amounts of lipid IVa. There was a small drop in activity between 0 and 0.02 μg/mL lipid IVa, then no further drop in A9 activity with up to 2 μg/mL lipid IVa. Further, we saw no apparent shift in TLR4/MD-2's sensitivity to A9 with increasing lipid IVa, ruling out a simple competitive binding mode.

One complication in our experiment was that we expressed and purified A9 from *E. coli*, whose outer membranes contain the TLR4/MD-2 agonist hexa-acylated LPS. Although we do several steps to specifically remove LPS, trace amounts of LPS could contaminate the purified A9, resulting in a spurious inflammatory signal. To control for this, we did two experiments.

First, we measured the activity of 1.7 μM A9 without lipid IVa in the presence of polymyxin B (PB). PB is a strong chelator of LPS, thus allowing us to determine what fraction of the signal we measured was due to A9 versus contaminating LPS. (It also chelates lipid IVa, meaning we could not include it in our main competition experiment). The addition of PB to a pure A9 sample did indeed drop the signal; however, it dropped it by the same amount as

saturating concentrations of lipid IVa. This explains the initial drop in signal with the addition of lipid IVa: it is displacing an LPS contaminant, revealing whatever activity is due to A9 itself. The A9 signal is then robust to increasing lipid IVa concentrations.

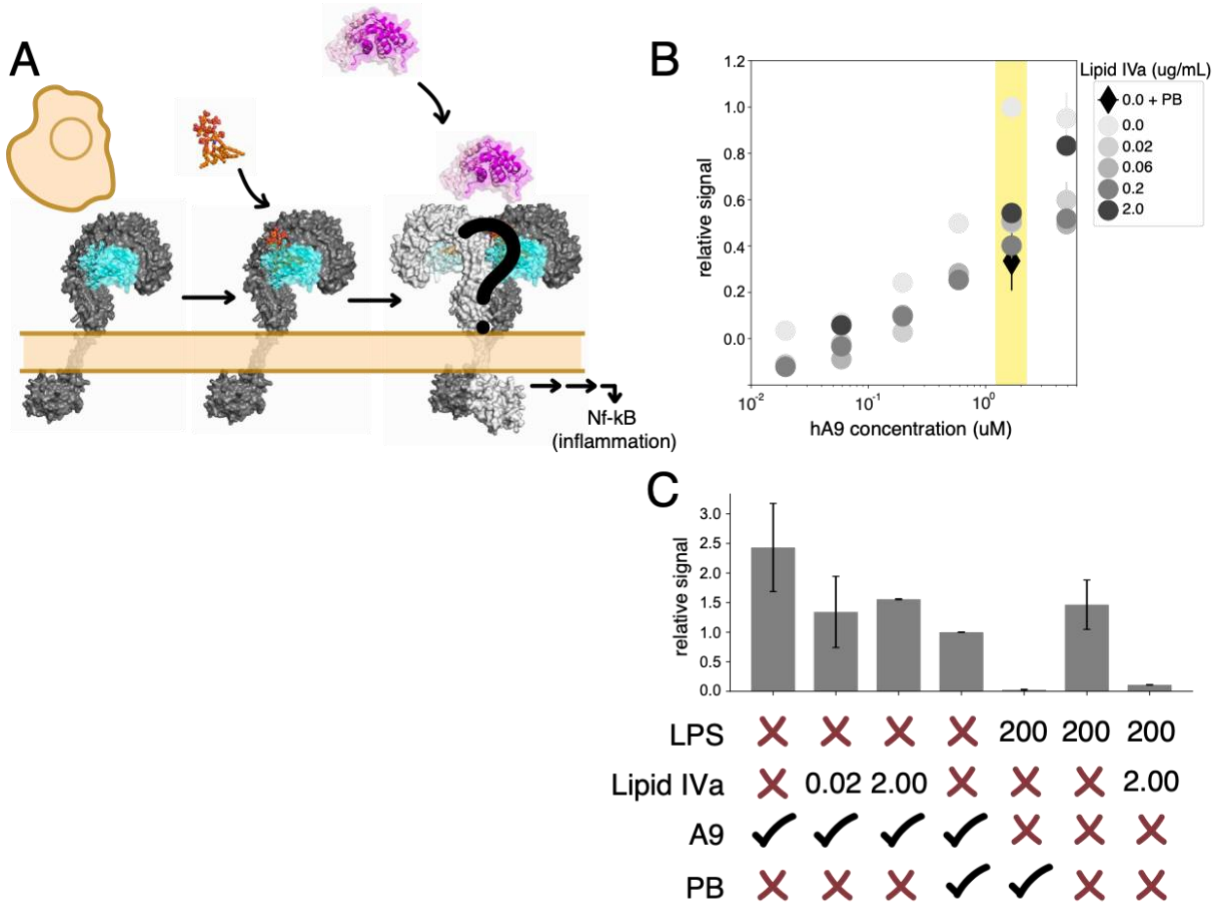

Figure 4.1. S100A9 activates inflammation via TLR4/MD-2 through a non-canonical interaction. (A) The direct binding interaction between A9 (homodimer in magenta and pink) and TLR4/MD-2 is unknown. A9 cannot bind in the hydrophobic MD-2 binding pocket, but does initiate NF-κB inflammation similar to hexa-acylated LPS. (B) Relative activity of HEK293T cells co-transfected with human TLR4/MD-2/CD14 and treated with mixtures of increasing amounts of hA9 and lipid IVa (hTLR4/MD-2 antagonist). Black diamond depicts 1.7uM A9 + polymyxin B (PB). Highlighted yellow region at 1.7uM A9 has further experiments in (C). Relative activity of selected combinations of hexa-acylated LPS (ng/well), lipid IVa (μg/well), 1.7μM A9, and PB.

As a second control, we tested the ability of lipid IVa to compete with LPS to activate TLR4/MD-2. We asked what would happen if the entire signal we observed from our A9 sample was due to contaminating LPS. We thus found a concentration of LPS that was sufficient to activate TLR4/MD-2 to the same extent as our 1.7 μM A9 treatment (200 ng/well LPS, as it turned out; Fig. 4.1C). We then added 2 μg/well lipid IVa to this sample. Unlike in the A9 experiment, lipid IVa completely ameliorated this signal. As with the previous experiments, this is consistent A9 activating TLR4/MD-2 by a different site—or at least different mechanism—than LPS.

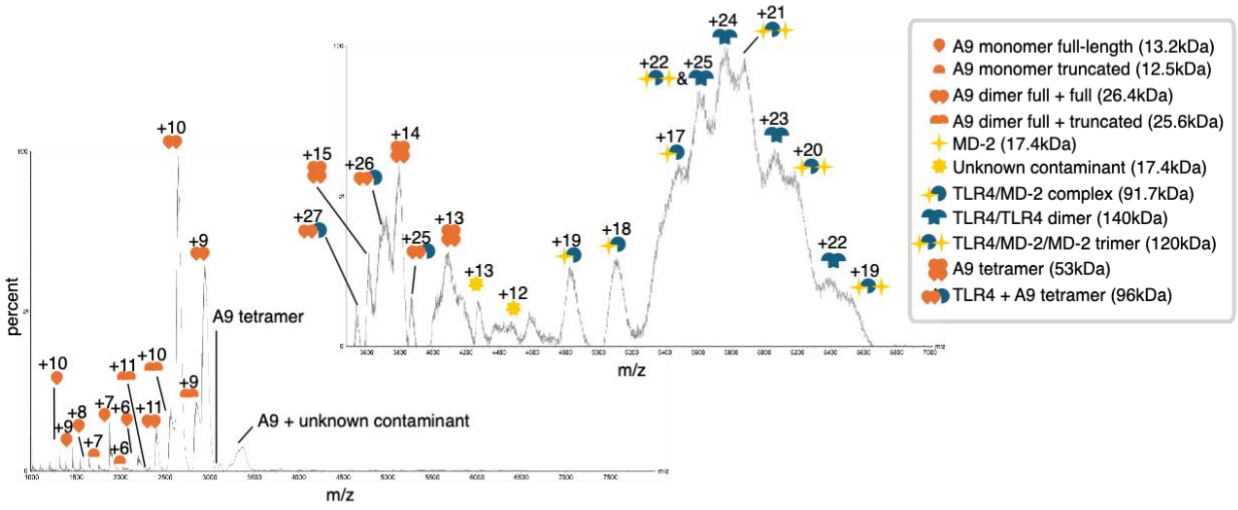

Figure 4.2. Mass spectrometry of A9 with TLR4/MD-2 does not obviously capture binding stoichiometries. Native mass spectrum showing stoichiometries of A9, TLR4, MD-2, and contaminating species.

The stoichiometry of the complex remains unclear

Several groups have proposed binding models for A9 and TLR4/MD-2. These relatively unsophisticated docking calculations, however, and have been minimally tested. Even the stoichiometry of the model interaction remains unknown. The simplest model would have a dimer of A9 activating a heterodimer of TLR4/MD-2; however, other combinations are possible.

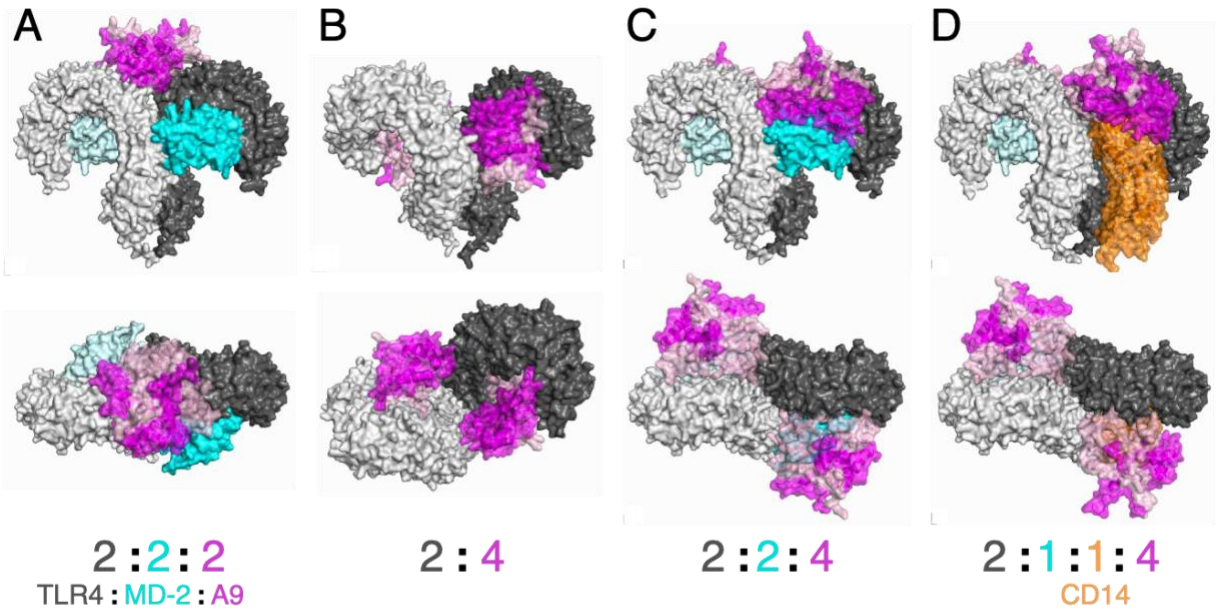

Figure 4.3. Putative binding models and stoichiometries for A9 with the TLR4 complex. Models were generated using plausible stoichiometries of human TLR4, MD-2, A9, and CD14 in AlphaFold3. (A) Top-binding model, whereby A9 drives dimerization by holding both TLR4/MD-2 heterodimers in place. (B) MD-2 displacement model, where A9 replaces MD-2 but provides additional contacts that mediate dimerization of the complex. (C) A9 on top of MD-2 model, where A9 binds on top of MD-2 and provides additional contacts on a similar interface where the diglucosamine backbone, core, and O-antigen of LPS interacts. (D) CD14 (orange) replacement model, where CD14 (an accessory protein which is known to bind to and improve activation with A9) replaces one (or both) of the MD-2(s) and helps mediate a binding interaction between A9 and TLR4 that activates the receptor.

To attempt to constrain this aspect of the problem, we set out to measure the stoichiometry of the interaction *in vitro* using nano electrospray ionization MS, as well as MALDI-TOF MS (Fig. 4.2). For these experiments, we used our purified A9 protein and commercially available TLR4/MD-2 complexes. We found that a modest binding interaction was identifiable, but the stoichiometry of this interaction was difficult to confidently identify, as all concentrations were at the lower limits of this method. Further complicating the picture, native MD-2 is glycosylated at two sites in its native, active cellular environment; however, we had to

deglycosylase the proteins to have any hope of detecting a signal.[147] This experiment thus had contamination of deglycosylating proteins, making the results from this experiment difficult to interpret and decidedly inconclusive. We have some evidence for binding *in vitro*, but are not able to confidently assign stoichiometries.

Possible models for direct binding of A9 to TLR4/MD-2

We next used AlphaFold3 to generate a family of models with a collection of different stoichiometries for A9, TLR4, MD-2, and the important co-receptor CD14.[148] This identified four possible binding modes, each with different contacts (Fig 4.3A-D). We selected these structures based on their rank score.

The first model we identified was the "top binding" model (Fig. 4.3A). In this model, an A9 dimer bridges the horseshoe-like structures of two TLR4 chains. MD-2 is present, but not involved in a direct contact with A9. CD14 is not part of the complex, consistent with a transient participation with the TLR4/MD-2 complex. There is perhaps the most evidence for this model, which is consistent with the binding mode amongst other TLRs. TLR1/TLR2 and TLR3 both activate with their respective ligands, Pam3CSK4 lipopeptide and dsRNA, via top-binding which mediates dimerization of the receptor complex.[149,150]

The second model had A9 displacing MD-2 and driving activation through a direct interaction (Fig. 4.3B). This has been suggested previously as a potential binding model for how DAMPs which cannot occupy the canonical LPS-binding site could still drive TLR4 dimerization.[25] Previous studies have shown that MD-2 is required for activation of TLR4 by A9; however, co-expression of MD-2 alongside TLR4 is known to be necessary for proper folding and translocation of TLR4 to the outer cellular membrane.[12,147] It is possible that the

TLR4/MD-2 complex is trafficked to the cell surface, followed by displacement of MD-2 with A9 upon activation. MD-2 can remain soluble un-bound to TLR4.

The third model is a variant of the top-down model with a different stoichiometry. In this model, two dimers of A9, together, bridge the complex (Fig. 4.3C). Rather than primarily interacting with TLR4, A9 shares interaction interfaces with TLR4 and MD-2. This model is supported by what is known about canonical LPS binding to TLR4/MD-2 – A9 contacts surfaces that have high overlap with regions on TLR4/MD-2 that are known to interact with the diglucosamine backbone, core, and O-antigen of LPS which contribute to dimerization (Fig. 4.3E,G).

The fourth model has CD14 and A9 together displace MD-2. This is consistent with other results from our lab, which have shown that CD14 increases the A9-induced TLR4/MD-2 signal by a factor of five. This model has CD14 locking A9 into place with TLR4/MD-2, possibly forming larger protein complex that is stable for long enough to be internalized. Previous studies have shown that A9 can activate TLR4/MD-2 without the presence of CD14, however this signaling likely only occurs at the membrane and activates the MyD88/TIRAP pathway, which results in a weaker inflammatory response.[127]

Testing the top binding model

Of the four candidates, the top-binding model seemed the most plausible given what is known about other members of the Toll-like receptor family. Knowing that A9 does not bind in the canonical binding site and that top-binding is the typical mechanism by which ligand-binding drives dimerization of TLRs, we hypothesized that the most plausible binding model is that A9 directly interacts with TLR4/MD-2 via a top-binding model. To further probe this model, we

identified 24 sites on TLR4 and 6 sites on A9 that we predicted would perturb this putative binding interaction. We created single point mutants for each of these sites, introducing mutations that were most likely deleterious (e.g. charge reversals, replacing large hydrophobic amino acids with alanine, etc.).

We collected dose response curves of TLR4 mutants with 0-5 μM A9, as well as a single concentration of LPS at around ~75% of the maximum activation. We hypothesized that sites that uniquely affected an A9 binding interaction would decrease A9 activation without having a significant impact on LPS activation. Mutations that affected activity with both ligands were considered to have globally affected the ability of TLR4/MD-2 to dimerize and activate.

From this study, we found that mutants E278K, E321F, and D325K caused the most significant drop in A9 activation without significantly affecting LPS activation (Fig. 4.2B). Mutants N305F, R322E, and K324F had a global effect on both A9 and LPS to decrease activation. Mapped to the top-binding model, it is unclear why these positions would have the largest effect. It is possible that LPS activation was negatively affected because of disruption to a surface that interacts with the O-antigen, although presumably the O-antigen can interact with other postions studied (which were not negatively affected by mutation).

Further, we studied 6 mutations on A9 that were expected to perturb a potential top-binding interaction. We found that all modestly decreased A9 activation without significantly impacting A9 secondary structure, which suggests that these mutations did not notably affect A9 folding (Fig. 4.4C; CD data not shown). H28A and D30A decreased A9 activation most significantly. However, none of these mutations fully abrogated A9 binding.

We introduced a few double mutants (D273K_D299K and R322G_K324F) that globally decreased activation for both LPS and A9, or had little effect from wildtype, respectively (Fig.

4.2B). There was modest support for this model but we were unable to entirely break this binding interaction uniquely with just A9.

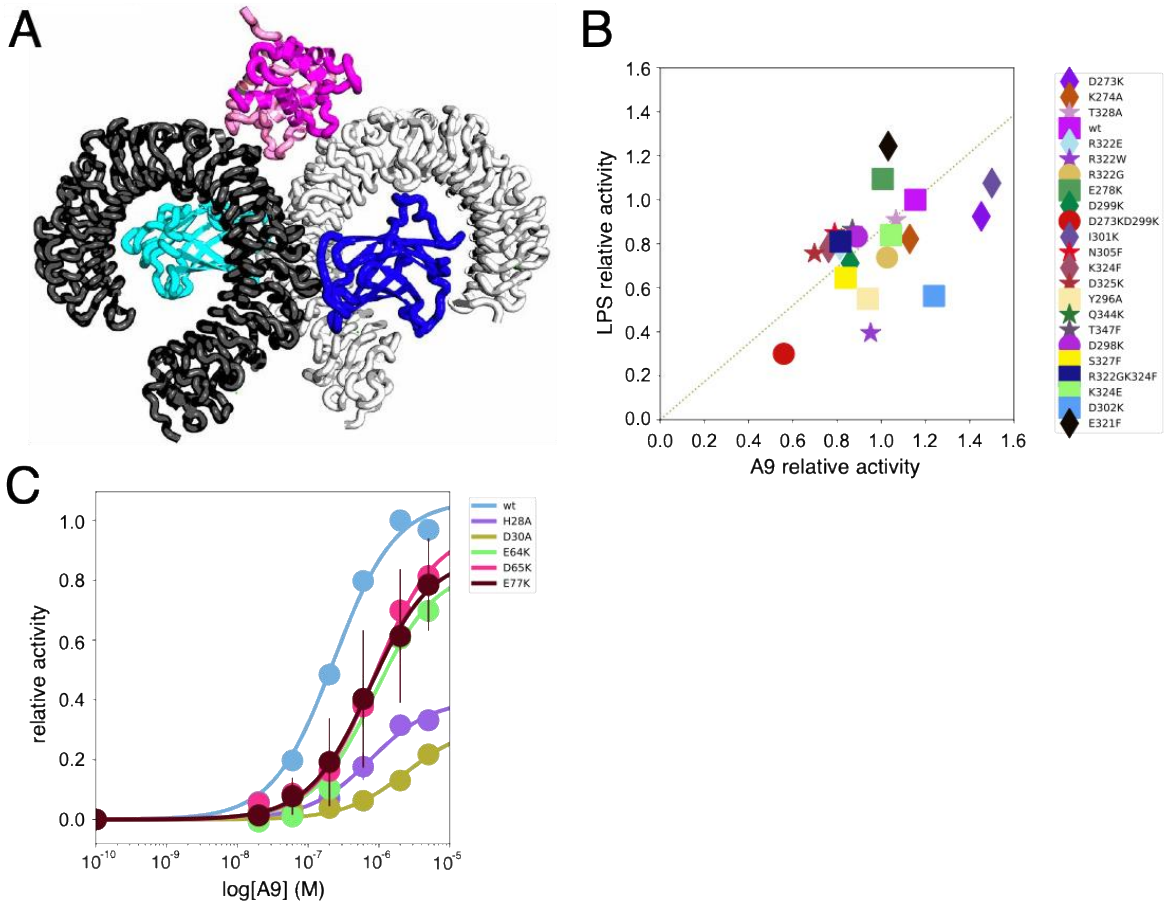

Figure 4.4. A9 top-binding model is not strongly supported by single TLR4 point mutations. (A) Proposed top-binding model between A9 and TLR4/MD-2 which docks on top of each TLR4/MD-2 heterodimer to drive tetramerization of the receptor complex for activation. Model was stable via Rosetta ligand docking.[151] (B) Relative activation of 24 candidate TLR4 mutants with respect to hA9 maximum activation (calculated from dose response curves, see Materials and Methods) and LPS (200ng/well) in HEK293T cells. (C) Dose response curves showing relative activation of six candidate hA9 mutants.

Discussion

In this work, I show that A9 interacts with TLR4/MD-2 in a novel mechanism that is unique from the canonical LPS binding site. Identifying the binding stoichiometry of this interaction *in vitro* would narrow down the possible binding modes this complex could take.

However, attempts at discerning this binding stoichiometry using ion electrospray MALDI-TOF native mass spectrometry with A9 and TLR4/MD-2 were inconclusive due to the presence of contaminating deglycosylation proteins for MD-2, as well as low signal from using very small amounts of TLR4/MD-2 that were difficult to resolve. Despite this, we proceeded to generate putative binding models exploring different plausible stoichiometries using AlphaFold3. From these candidate models, we functionally probed a most-likely model, the top-binding model whereby A9 drives dimerization of the complex by bridging the tops of each TLR4. We introduced 24 mutations to TLR4 and 6 mutations to A9 that would perturb this model, but found that using single and selected double point mutations, we were unable to confidently disrupt activity with A9. A9 mutations H28A and D30A had noteworthy effects on reducing activation by potentially removing important charged interactions with TLR4. The E278K mutant perturbed a site that was only found in the top-binding model.

This binding mechanism is in some part difficult to define because TLR4/MD-2, as a large transmembrane heterodimer, is hard to co-express and purify *in vitro* in large enough quantities for biochemical experimentation. In leu of this, thoughtful and well-defined structural models can help us tease apart important regions for binding through characterization of mutants. We find it important to note that the binding interaction between TLR4 and A9 may be transient or defined by a broad surface. In this case, it is likely that single mutations in the true binding interface may only introduce modest decreases in activation, and several mutations will be required to fully break the binding interaction.

The models proposed here compile knowledge from previous evolutionary, structural, and functional work, and describe our best understanding of possible interactions between A9 and TLR4. Future work defining the binding mechanism between A9 and the TLR4 complex

would benefit greatly from investigating mutations that uniquely probe the binding models offered here.

MATERIALS AND METHODS

AlphaFold 3 Docking

The AlphaFold3 online server was used to predict the binding models between A9 and the TLR4 complex.[148] We used sequences for the human TLR4 ectodomain, MD-2, CD14, and A9 which match previous crystallographic studies.[11,152,153] These structures have N- and C-terminal tails removed which are disordered. We included four calcium ions for each A9 homodimer in our structure prediction. Stoichiometries are detailed in Fig. 4.3. We proceeding with further analyses using only top-ranked models.

Mutagenesis and Cloning

Full-length human TLR4, MD-2 and CD14 genes were encoded in pcDNA3.1(+) plasmid vectors under a CMV constitutive promoter. hTLR4 was a gift from Ruslan Medzhitov (Addgene plasmid # 13086; http://n2t.net/addgene:13086). hMD-2_pcDNA3.1+ was purchased from Genscript (LY96_OHu26610C_pcDNA3.1(+)). pcDNA3-CD14 was a gift from Doug Golenbock (Addgene plasmid # 13645 ; http://n2t.net/addgene:13645). S100A9 Mutagenesis of human and single human TLR4 and A9 constructs were done as recommended using the QuikChange Lightning site-directed mutagenesis kit (Agilent). Primers were designed using the online QuikChange Primer Design server (Agilent) and ordered from Eurofins. Linear mutagenized plasmids were re-circularized using KLD enzyme mix (New England Biolabs).

S100A9 Expression and Purification

Recombinant human S100A9 was expressed and purified as described previously.[130] In brief, cysteine-free (C3S) A9 in a pETDUET-1 vector was expressed in Rosetta BL21(DE3) pLysS *E. coli.* Bacteria was grown in 1.5L liquid cultures at 37C, with 225rpm shaking until OD_{600} = 0.8-1.0 was reached. Cultures were induced with 1mM IPTG for 16hrs overnight at 16°C. Cells were harvested using centrifugation at 3,000rpm at 4°C and lysed at 15,000rpm at 4°C. A9 was then purified through three subsequent rounds of column chromatography: HisTrap nickel immobilized metal ion affinity at pH 7.4, HiTrap Q anion exchange at pH 8, and finally HiTrap Q anion exchange at pH 6. Protein purity of >95% was verified via SDS-PAGE. Protein was concentrated and buffer exchanged into 25 mM Tris, 100 mM NaCl at pH 7.4, and flash-frozen dropwise into liquid nitrogen. Proteins were stored at -80 °C until further use. Protein concentration was determined using A280 with an extinction coefficient of 6990 M-1 cm-1 (monomer). A9 protein concentration is reported here as μM dimer.

Functional Transfection Assay

We measured activation of TLR4/MD2 via a well-established NF-kB assay in HEK293T cells as previously described[33,116,127] Cells were passaged up to 30 times and maintained in DMEM with the addition of 10% FBS and Antibiotic-Antimycotic, at 37°C with 5% CO_2. For each experiment, 135 μL of cells at ~25% confluency were seeded in a 96-well plate. Each well of cells were then transiently co-transfected with 65μL aliquots of 100ng plasmids consisting of 10ng TLR4, 1ng MD-2, 1 ng CD14, 20ng pGL3-elam-luc, 1ng Renilla pRL-TK, and 67ng empty pcDNA3.1(+) vector diluted in Opti-MEM (ThermoFisher Scientific). PLUS and lipofectamine were used as recommended to enhance transfection efficiency (ThermoFisher Scientific).

Endotoxin samples were diluted in endotoxin-free water and sonicated in a jewelry ultrasonicator for 15 minutes at room temperature at 15 watts immediately before use in order to uniformly disrupt micelles. Transfected cells were incubated for 20-24 hours. After this, transfection mix was removed and replaced with 100µL per well treatments of 200ng/well LPS or 0-5uM A9 + 200µg/mL polymyxin B (an LPS-chelator, to reduce contaminating signal). Before use in treatments, A9 was buffer exchanged into endotoxin-free PBS, using Pall Microsep concentrator spin columns. Treatments were diluted in 25 µL endotoxin-free PBS and 75 µL serum-free DMEM. Cells were incubated with treatments for 3 hours, then activity for each treated well was assessed using the Dual-Glo® Luciferase Assay System (Promega) with the SpectraMax i3 plate reader. All treatment conditions were tested in technical triplicate wells. Raw luciferase values per well were corrected for cell count by dividing luciferase by renilla values to yield a relative activity value. The relative activity for technical triplicate wells was averaged and considered as one biological replicate. All data presented were done in (at least) biological triplicate. Biological replicates for each species TLR4/MD-2/CD14 set was always treated with triplicate wells of mock treatment (no protein or endotoxin added) for background subtraction. 1.7uM wildtype hA9 + PB was on every plate and used as a normalizing factor, with relative signal for this treatment set to 1.

Native Mass Spectrometry (nMS):

Commercial TLR4/MD-2 (carrier-free) was purchased lyophilized and resuspended directly into 200 mM ammonium acetate at pH 7 (R&D Systems). Purified A9 and resuspended TLR4/MD-2 were further buffer-exchanged into 200 mM ammonium acetate at pH 7 using Micro Bio-spin P6 columns to remove salts or other artifacts from purification (Bio-Rad,

Hercules, CA). Buffer- exchanged A9 was diluted to a working concentration of 6.66 μM; TLR4/MD-2 were diluted to 3.33 μM. All A9 samples had 33 μM CaCl2 spiked in to ensure A9 is in the calcium-bound state. Previous experiments found that calcium-spiking immediately before experimentation results in more calcium-bound A9 versus overnight incubation (data not shown). Protein samples were analyzed with native MS on a Waters Synapt G2-Si quadrupole–ion mobility– time-of-flight instrument (Milford, MA). Samples were introduced to the instrument using borosilicate capillary needles (prepared in-house with emitter i.d. ~450 nm) threaded with a platinum wire to allow nano-electrospray ionization (nESI). The electrospray was operated at capillary voltage of 0.3- 0.5 kV with the sample cone and temperature at 25 V and 25 °C, respectively, in positive mode. The Trap cell was operated with an argon gas flow of 10 mL/min. The Collision Energy for both the Trap and Transfer cells was set to 5 V. For ion mobility separations, the IMS cell was pressurized at ~3.4 mbar nitrogen buffer gas. IM separation was performed with a traveling wave height of 30 V and wave velocity of 600 m/s. All native MS data were collected over the m/z range of 500-8000.

Bridge to Chapter V

In Chapter IV, I discuss work I've done to gain further insight into the mechanism by which S100A9 activates TLR4. I propose several direct binding mechanisms created with AlphaFold3 which are plausible given functional and structural data that is known about this interaction. From here, I functionally characterized a most-likely direct top-binding model where S100A9 bridges the top of two TLR4s to drive dimerization and thus activation of the receptor. It remains inconclusive if this model accurately describes the true binding interaction. In my final Chapter V, I summarize the findings from each of my chapters and provide concluding remarks.

CHAPTER V

CONCLUDING REMARKS AND SUMMARY

This dissertation offers three bodies of work that interplay broadly between the study of protein evolution and specifically in the study of the specificity and function of an innate immune receptor. As for TLR4, in this work we aimed to better understand how this powerful immune receptor recognizes and initiates inflammation with a variety of ligands. We set out to answer two important questions: (1) How did TLR4 evolve species-specific recognition of LPS with different acylation states? And (2) How does TLR4 activate with the endogenous protein S100A9?

Increasing accessibility of computational methods through our ancestral sequence reconstruction pipeline, topiary, allows for more scientists to include useful evolutionary context in their studies of protein structure and function. This bioinformatics tool compiles and streamlines best-practice software and programs to reduce time spent with file preparation and conversion, and by default makes thoughtful choices based on statistical confidence which can be customized or overridden by more expert computational evolutionary scientists. We were able to harness ASR to provide further support for the hypothesis that the TLR4 ancestor from all placental mammals and potentially earlier was antagonized by hypo-acylated LPS, and that agonism for this ligand may be selected for in certain lineages.

For our first question, we characterized several extant and ancestral TLR4s that display antagonism with hypo-acylated LPS. Combining results from a previous molecular dynamics simulation of LPS docking in TLR4/MD-2 with results from positive selection tests, we concluded that TLR4 diversifies primarily to recognize changes to the core and O-antigen of LPS, whereas MD-2 diversifies to recognize changes to the lipid A portion of LPS. After

characterization of species-swapping mutations of MD-2 sites under positive selection, we found that L4 agonism is easy to damage but impossible to improve through single mutations. *In silico* docking experiments in the dimerized mouse TLR4/MD-2 structure showed that this dimerization is weaker when L4 is bound versus when L6 is bound. Introduction of the mouse-to-human MD-2 mutation E122K further destabilized the dimerization interface notably with L4 but minimally with L6. In sum, we conclude that it is biochemically challenging and requires several sequence-level changes to orient L4 into an agonistic position, resulting in L4 antagonism being the default specificity of TLR4/MD-2 unless there is selective pressure to evolve agonism.

For our second question, we found that A9 must have a novel interaction with TLR4/MD-2 that is unique and non-competitive with the canonical ligand, LPS. While we were not able to conclusively resolve the binding stoichiometry between A9 and the TLR4 complex, we were able to generate four direct binding models which are highly plausible given previous literature, homology of binding amongst other TLRs, and intelligent protein-protein structure prediction calculations. We tested a top-binding model where A9 drives dimerization by docking on top of each TLR4/MD-2 heterodimer, by introducing 24 mutations to TLR4 and 6 mutations to A9 that would disrupt this putative interaction. We were not able to identify any single mutation which entirely broke this interaction, although a small selection of mutations singularly affected A9 activation, and another selection of mutations affected both A9 and LPS activation. Even we found that A9 binds non-competitively with the canonical MD-2 binding pocket, is it likely that any TLR4 surface that A9 binds to could also be a surface that interacts productively with the O-antigen of LPS. We do not outright reject the top-binding model, although further

studies probing this and other binding models could help uncover the true binding mechanism between A9 and TLR4/MD-2.

This work highlights the importance of combining evolutionary and structural techniques when studying the role proteins play and how they arrived at their modern day functions.

REFERENCES CITED

1. Mushegian, A. & Medzhitov, R. Evolutionary perspective on innate immune recognition. *J. Cell Biol.* **155**, 705–710 (2001).
2. Ruslan Medzhitov, C. J. Jr. Innate immune recognition: mechanisms and pathways. *Immunol. Rev.* **173**, 89–97.
3. Silva, R. C. M. C. & Gomes, F. M. Evolution of the Major Components of Innate Immunity in Animals. *J. Mol. Evol.* **92**, 3–20 (2024).
4. Pahwa, R., Goyal, A. & Jialal, I. Chronic Inflammation. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2025).
5. Xiao, H., Siddiqui, J. & Remick, D. G. Mechanisms of Mortality in Early and Late Sepsis. *Infect. Immun.* **74**, 5227–5235 (2006).
6. Daviaud, F. *et al.* Timing and causes of death in septic shock. *Ann. Intensive Care* **5**, 16 (2015).
7. Rhee, C. *et al.* Prevalence, Underlying Causes, and Preventability of Sepsis-Associated Mortality in US Acute Care Hospitals. *JAMA Netw. Open* **2**, e187571 (2019).
8. Roach, J. C. *et al.* The evolution of vertebrate Toll-like receptors. *Proc. Natl. Acad. Sci.* **102**, 9577–9582 (2005).
9. Liu, G., Zhang, H., Zhao, C. & Zhang, H. Evolutionary History of the Toll-Like Receptor Gene Family across Vertebrates. *Genome Biol. Evol.* **12**, 3615–3634 (2020).
10. El-Zayat, S. R., Sibaii, H. & Mannaa, F. A. Toll-like receptors activation, signaling, and targeting: an overview. *Bull. Natl. Res. Cent.* **43**, 187 (2019).
11. Park, B. S. *et al.* The structural basis of lipopolysaccharide recognition by the TLR4–MD-2 complex. *Nature* **458**, 1191–1195 (2009).

12. Lauer, S., Kunde, Y. A., Apodaca, T. A., Goldstein, B. & Hong-Geller, E. Soluble MD2 increases TLR4 levels on the epithelial cell surface. *Cell. Immunol.* **255**, 8–16 (2009).
13. Poltorak, A. *et al.* Defective LPS Signaling in C3H/HeJ and C57BL/10ScCr Mice: Mutations in Tlr4 Gene. *Science* **282**, 2085–2088 (1998).
14. Kong, Q. *et al.* Phosphate Groups of Lipid A Are Essential for Salmonella enterica Serovar Typhimurium Virulence and Affect Innate and Adaptive Immunity. *Infect. Immun.* **80**, 3215–3224 (2012).
15. Ittig, S. *et al.* The Lipopolysaccharide from Capnocytophaga canimorsus Reveals an Unexpected Role of the Core-Oligosaccharide in MD-2 Binding. *PLOS Pathog.* **8**, e1002667 (2012).
16. Kawahara, K., Tsukano, H., Watanabe, H., Lindner, B. & Matsuura, M. Modification of the Structure and Activity of Lipid A in Yersinia pestis Lipopolysaccharide by Growth Temperature. *Infect. Immun.* **70**, 4092–4098 (2002).
17. Pandolfi, F., Altamura, S., Frosali, S. & Conti, P. Key Role of DAMP in Inflammation, Cancer, and Tissue Repair. *Clin. Ther.* **38**, 1017–1028 (2016).
18. Piccinini, A. M. & Midwood, K. S. DAMPening Inflammation by Modulating TLR Signalling. *Mediators Inflamm.* **2010**, 672395 (2010).
19. Land, W. G. The Role of Damage-Associated Molecular Patterns (DAMPs) in Human Diseases. *Sultan Qaboos Univ. Med. J.* **15**, e157–e170 (2015).
20. Fukata, M. & Abreu, M. T. TLR4 signalling in the intestine in health and disease. *Biochem. Soc. Trans.* **35**, 1473–1478 (2007).
21. Bartels, Y. L. *et al.* Inhibition of TLR4 signalling to dampen joint inflammation in osteoarthritis. *Rheumatology* **63**, 608–618 (2024).

22. Spirig, R., Tsui, J. & Shaw, S. The Emerging Role of TLR and Innate Immunity in Cardiovascular Disease. *Cardiol. Res. Pract.* **2012**, 181394 (2012).
23. Hydrophobicity: an ancient damage-associated molecular pattern that initiates innate immune responses | Nature Reviews Immunology. https://www.nature.com/articles/nri1372.
24. Calderwood, S. K., Gong, J. & Murshid, A. Extracellular HSPs: The Complicated Roles of Extracellular HSPs in Immunity. *Front. Immunol.* **7**, (2016).
25. Camp, S. M. *et al.* Unique Toll-Like Receptor 4 Activation by NAMPT/PBEF Induces NFκB Signaling and Inflammatory Lung Injury. *Sci. Rep.* **5**, 13135 (2015).
26. Vogl, T. *et al.* Autoinhibitory regulation of S100A8/S100A9 alarmin activity locally restricts sterile inflammation. *J. Clin. Invest.* **128**, 1852–1866 (2018).
27. Romerio, A. & Peri, F. Increasing the Chemical Variety of Small-Molecule-Based TLR4 Modulators: An Overview. *Front. Immunol.* **11**, (2020).
28. Vogl, T. *et al.* MRP8 and MRP14 control microtubule reorganization during transendothelial migration of phagocytes. *Blood* **104**, 4260–4268 (2004).
29. Kerkhoff, C., Klempt, M., Kaever, V. & Sorg, C. The Two Calcium-binding Proteins, S100A8 and S100A9, Are Involved in the Metabolism of Arachidonic acid in Human Neutrophils*. *J. Biol. Chem.* **274**, 32672–32679 (1999).
30. Orlandi, K. N. & Harms, M. J. Zebrafish do not have a calprotectin ortholog. *PLOS ONE* **20**, e0322649 (2025).
31. Zimmer, D. B., Eubanks, J. O., Ramakrishnan, D. & Criscitiello, M. F. Evolution of the S100 family of calcium sensor proteins. *Cell Calcium* **53**, 170–179 (2013).
32. Shang, X., Cheng, H. & Zhou, R. Chromosomal mapping, differential origin and evolution of the S100 gene family. *Genet. Sel. Evol.* **40**, 449 (2008).

33. Loes, A. N., Bridgham, J. T. & Harms, M. J. Coevolution of the Toll-Like Receptor 4 Complex with Calgranulins and Lipopolysaccharide. *Front. Immunol.* **9**, (2018).

34. Aderem, A. & Ulevitch, R. J. Toll-like receptors in the induction of the innate immune response. *Nature* **406**, 782–787 (2000).

35. Candore, G. *et al.* Inflammation, Longevity, and Cardiovascular Diseases. *Ann. N. Y. Acad. Sci.* **1067**, 282–287 (2006).

36. Pabst, S. *et al.* Toll-like receptor (TLR) 4 polymorphisms are associated with a chronic course of sarcoidosis. *Clin. Exp. Immunol.* **143**, 420–426 (2006).

37. Noreen, M. *et al.* TLR4 polymorphisms and disease susceptibility. *Inflamm. Res.* **61**, 177–188 (2012).

38. Mutations in TLR4 signaling that lead to increased susceptibility to infection in humans: an overview - Stefanie N. Vogel, Agnes A. Awomoyi, Prasad Rallabhandi, Andrei E. Medvedev, 2005. https://journals.sagepub.com/doi/abs/10.1177/09680519050110060801.

39. Gruber, A., Manček, M., Wagner, H., Kirschning, C. J. & Jerala, R. Structural Model of MD-2 and Functional Role of Its Basic Amino Acid Clusters Involved in Cellular Lipopolysaccharide Recognition*. *J. Biol. Chem.* **279**, 28475–28482 (2004).

40. Viriyakosol, S., Tobias, P. S. & Kirkland, T. N. Mutational Analysis of Membrane and Soluble Forms of Human MD-2*. *J. Biol. Chem.* **281**, 11955–11964 (2006).

41. Kawasaki, K., Nogawa, H. & Nishijima, M. Identification of Mouse MD-2 Residues Important for Forming the Cell Surface TLR4-MD-2 Complex Recognized by Anti-TLR4-MD-2 Antibodies, and for Conferring LPS and Taxol Responsiveness on Mouse TLR4 by Alanine-Scanning Mutagenesis1. *J. Immunol.* **170**, 413–420 (2003).

42. Meng, J., Drolet, J. R., Monks, B. G. & Golenbock, D. T. MD-2 Residues Tyrosine 42, Arginine 69, Aspartic Acid 122, and Leucine 125 Provide Species Specificity for Lipid IVA. *J. Biol. Chem.* **285**, 27935–27943 (2010).

43. Chisholm, L. O., Orlandi, K. N., Phillips, S. R., Shavlik, M. J. & Harms, M. J. Ancestral Reconstruction and the Evolution of Protein Energy Landscapes. *Annu. Rev. Biophys.* **53**, 127–146 (2024).

44. Worth, C. L., Gong, S. & Blundell, T. L. Structural and functional constraints in the evolution of protein families. *Nat. Rev. Mol. Cell Biol.* **10**, 709–720 (2009).

45. Schweitzer, M. H., Schroeter, E. R., Cleland, T. P. & Zheng, W. Paleoproteomics of Mesozoic Dinosaurs and Other Mesozoic Fossils. *PROTEOMICS* **19**, 1800251 (2019).

46. Harms, M. J. & Thornton, J. W. Analyzing protein structure and function using ancestral gene reconstruction. *Curr. Opin. Struct. Biol.* **20**, 360–366 (2010).

47. Pauling, L., Zuckerkandl, E., Henriksen, T. & Lövstad, R. Chemical Paleogenetics. Molecular 'Restoration Studies' of Extinct Forms of Life. *Acta Chem. Scand.* **17 supl.**, 9–16 (1963).

48. Spence, M. A., Kaczmarski, J. A., Saunders, J. W. & Jackson, C. J. Ancestral sequence reconstruction for protein engineers. *Curr. Opin. Struct. Biol.* **69**, 131–141 (2021).

49. Nicoll, C. R. *et al.* Ancestral-sequence reconstruction unveils the structural basis of function in mammalian FMOs. *Nat. Struct. Mol. Biol.* **27**, 14–24 (2020).

50. Furukawa, R., Toma, W., Yamazaki, K. & Akanuma, S. Ancestral sequence reconstruction produces thermally stable enzymes with mesophilic enzyme-like catalytic properties. *Sci. Rep.* **10**, 15493 (2020).

51. Zakas, P. M. *et al.* Enhancing the pharmaceutical properties of protein drugs by ancestral sequence reconstruction. *Nat. Biotechnol.* **35**, 35–37 (2017).

52. Anderson, D. P. *et al.* Evolution of an ancient protein function involved in organized multicellularity in animals. *eLife* **5**, e10147 (2016).

53. Diez-Hermano, S., Ganfornina, M. D., Skerra, A., Gutiérrez, G. & Sanchez, D. An Evolutionary Perspective of the Lipocalin Protein Family. *Front. Physiol.* **12**, 718983 (2021).

54. Harman, J. L. *et al.* Evolution of multifunctionality through a pleiotropic substitution in the innate immune protein S100A9. *eLife* **9**, e54100 (2020).

55. Mascotti, M. L. Resurrecting Enzymes by Ancestral Sequence Reconstruction. in *Enzyme Engineering* (eds. Magnani, F., Marabelli, C. & Paradisi, F.) vol. 2397 111–136 (Springer US, New York, NY, 2022).

56. Merkl, R. & Sterner, R. Ancestral protein reconstruction: techniques and applications. *Biol. Chem.* **397**, 1–21 (2016).

57. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

58. Rees, J. & Cranston, K. Automated assembly of a reference taxonomy for phylogenetic data synthesis. *Biodivers. Data J.* **5**, e12581 (2017).

59. Vialle, R. A., Tamuri, A. U. & Goldman, N. Alignment Modulates Ancestral Sequence Reconstruction Accuracy. *Mol. Biol. Evol.* **35**, 1783–1797 (2018).

60. Tan, G. *et al.* Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference. *Syst. Biol.* **64**, 778–791 (2015).

61. Tumescheit, C., Firth, A. E. & Brown, K. CIAlign: A highly customisable command line tool to clean, interpret and visualise multiple sequence alignments. *PeerJ* **10**, e12983 (2022).

62. Catanach, T. A. *et al.* 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* **7**, e6142 (2019).

63. Morrison, D. A. Multiple sequence alignment for phylogenetic purposes. *Aust. Syst. Bot.* **19**, 479 (2006).

64. Del Amparo, R. & Arenas, M. Consequences of Substitution Model Selection on Protein Ancestral Sequence Reconstruction. *Mol. Biol. Evol.* **39**, msac144 (2022).

65. Joy, J. B., Liang, R. H., McCloskey, R. M., Nguyen, T. & Poon, A. F. Y. Ancestral Reconstruction. *PLOS Comput. Biol.* **12**, e1004763 (2016).

66. Morel, B., Kozlov, A. M., Stamatakis, A. & Szöllősi, G. J. GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss. *Mol. Biol. Evol.* **37**, 2763–2774 (2020).

67. Groussin, M. *et al.* Toward More Accurate Ancestral Protein Genotype–Phenotype Reconstructions with the Use of Species Tree-Aware Gene Trees. *Mol. Biol. Evol.* **32**, 13–22 (2015).

68. Gogarten, J. P., Doolittle, W. F. & Lawrence, J. G. Prokaryotic Evolution in Light of Gene Transfer. *Mol. Biol. Evol.* **19**, 2226–2238 (2002).

69. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).

70. Kinene, T., Wainaina, J., Maina, S. & Boykin, L. M. Rooting Trees, Methods for. in *Encyclopedia of Evolutionary Biology* 489–493 (Elsevier, 2016). doi:10.1016/B978-0-12-800049-6.00215-8.

71. Yang, Z., Kumar, S. & Nei, M. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* **141**, 1641–1650 (1995).

72. 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.* msw223 (2016) doi:10.1093/molbev/msw223.

73. Akanuma, S. *et al.* Experimental evidence for the thermophilicity of ancestral life. *Proc. Natl. Acad. Sci.* **110**, 11067–11072 (2013).

74. 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.* **10**, e1004058 (2014).

75. McKeown, A. N. *et al.* Evolution of DNA Specificity in a Transcription Factor Family Produced a New Gene Regulatory Module. *Cell* **159**, 58–68 (2014).

76. Wheeler, L. C., Anderson, J. A., Morrison, A. J., Wong, C. E. & Harms, M. J. Conservation of Specificity in Two Low-Specificity Proteins. *Biochemistry* **57**, 684–695 (2018).

77. Edgar, R. C. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat. Commun.* **13**, 6968 (2022).

78. 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* **35**, 4453–4455 (2019).

79. Ishikawa, S. A., Zhukova, A., Iwasaki, W. & Gascuel, O. A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios. *Mol. Biol. Evol.* **36**, 2069–2085 (2019).

80. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).

81. Altschul, S. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

82. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

83. 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. *Syst. Biol.* **70**, 1295–1301 (2021).

84. Eaton, D. A. R. Toytree: A minimalist tree visualization and manipulation library for Python. *Methods Ecol. Evol.* **11**, 187–191 (2020).

85. Frith, M. C. How sequence alignment scores correspond to probability models. *Bioinformatics* btz576 (2019) doi:10.1093/bioinformatics/btz576.

86. Edgar, R. C. *High-Accuracy Alignment Ensembles Enable Unbiased Assessments of Sequence Homology and Phylogeny*. http://biorxiv.org/lookup/doi/10.1101/2021.06.20.449169 (2021) doi:10.1101/2021.06.20.449169.

87. Flouri, T. *et al.* The Phylogenetic Likelihood Library. *Syst. Biol.* **64**, 356–362 (2015).

88. Felsenstein, J. CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP. *Evolution* **39**, 783–791 (1985).

89. Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104–2105 (2005).

90. Maddison, D. R. & Maddison, W. P. MacClade 4. (2000).

91. Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276–3278 (2014).

92. 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* **25**, 1189–1191 (2009).

93. Tamura, K., Stecher, G. & Kumar, S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Mol. Biol. Evol.* **38**, 3022–3027 (2021).

94. Zheng, Y. & Zhang, L. Effect of Incomplete Lineage Sorting On Tree-Reconciliation-Based Inference of Gene Duplication. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **11**, 477–485 (2014).

95. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

96. Colless, D. H. & Wiley, E. O. Phylogenetics: The Theory and Practice of Phylogenetic Systematics. *Syst. Zool.* **31**, 100 (1982).

97. Sukumaran, J. & Holder, M. T. DendroPy: a Python library for phylogenetic computing. *Bioinformatics* **26**, 1569–1571 (2010).

98. Pattengale, N. D., Alipour, M., Bininda-Emonds, O. R. P., Moret, B. M. E. & Stamatakis, A. How Many Bootstrap Replicates Are Necessary? *J. Comput. Biol.* **17**, 337–354 (2010).

99. Kimbrell, D. A. & Beutler, B. The evolution and genetics of innate immunity. *Nat. Rev. Genet.* **2**, 256–267 (2001).

100. Eckburg, P. B. *et al.* Diversity of the Human Intestinal Microbial Flora. *Science* **308**, 1635–1638 (2005).

101. Costello, E. K. *et al.* Bacterial Community Variation in Human Body Habitats Across Space and Time. *Science* **326**, 1694–1697 (2009).

102. Hornung, V. *et al.* Quantitative Expression of Toll-Like Receptor 1–10 mRNA in Cellular Subsets of Human Peripheral Blood Mononuclear Cells and Sensitivity to CpG Oligodeoxynucleotides1. *J. Immunol.* **168**, 4531–4537 (2002).

103. Calil, I. L. *et al.* Lipopolysaccharide Induces Inflammatory Hyperalgesia Triggering a TLR4/MyD88-Dependent Cytokine Cascade in the Mice Paw. *PLoS ONE* **9**, e90013 (2014).

104. Wyns, H., Plessers, E., De Backer, P., Meyer, E. & Croubels, S. In vivo porcine lipopolysaccharide inflammation models to study immunomodulation of drugs. *Vet. Immunol. Immunopathol.* **166**, 58–69 (2015).

105. Liang, Y. *et al.* Protective Effect of Resveratrol Improves Systemic Inflammation Responses in LPS-Injected Lambs. *Animals* **9**, 872 (2019).

106. El-Radhi, A. S. Fever in Common Infectious Diseases. in *Clinical Manual of Fever in Children* (ed. El-Radhi, A. S.) 85–140 (Springer International Publishing, Cham, 2018). doi:10.1007/978-3-319-92336-9_5.

107. Mc Chlery, S., Ramage, G. & Bagg, J. Respiratory tract infections and pneumonia. *Periodontol. 2000* **49**, 151–165 (2009).

108. Liu, V. *et al.* Hospital Deaths in Patients With Sepsis From 2 Independent Cohorts. *JAMA* **312**, 90–92 (2014).

109. Kabanov, D. S. & Prokhorenko, I. R. Structural analysis of lipopolysaccharides from Gram-negative bacteria. *Biochem. Mosc.* **75**, 383–404 (2010).

110. (PDF) Effects of bacterial lipopolysaccharide on thermoregulation in green anole lizards (Anolis carolinensis). *ResearchGate* doi:10.1016/j.vetimm.2008.04.014.

111. Alber, A., Stevens ,Mark P. & and Vervelde, L. The bird's immune response to avian pathogenic Escherichia coli. *Avian Pathol.* **50**, 382–391 (2021).

112. LPS-Induced Lung Inflammation in Marmoset Monkeys – An Acute Model for Anti-Inflammatory Drug Testing | PLOS One. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0043709.

113. Fux, A. C. *et al.* Heterogeneity of Lipopolysaccharide as Source of Variability in Bioassays and LPS-Binding Proteins as Remedy. *Int. J. Mol. Sci.* **24**, 8395 (2023).

114. d'Hennezel, E., Abubucker, S., Murphy, L. O. & Cullen, T. W. Total Lipopolysaccharide from the Human Gut Microbiome Silences Toll-Like Receptor Signaling. *mSystems* **2**, 10.1128/msystems.00046-17 (2017).

115. Angers, A. *et al.* The Human Gut Microbiota: Overview and analysis of the current scientific knowledge and possible impact on healthcare and well-being. *JRC Publications Repository* https://publications.jrc.ec.europa.eu/repository/handle/JRC112042 (2018) doi:10.2760/17381.

116. 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 Sci. Publ. Protein Soc.* **28**, 1350–1358 (2019).

117. Lohmann, K. L., Vandenplas, M., Barton, M. H. & Moore, J. N. Lipopolysaccharide from Rhodobacter sphaeroides is an agonist in equine cells. *J. Endotoxin Res.* **9**, 33–37 (2003).

118. Aal-Aaboda, M., Abu Raghif, A. R. & Hadi, N. R. Effect of Lipopolysaccharide from Rhodobacter sphaeroides on Inflammatory Pathway and Oxidative Stress in Renal Ischemia/Reperfusion Injury in Male Rats. *Arch. Razi Inst.* **76**, 1013–1024 (2021).

119. 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.* **109**, 7421–7426 (2012).

120. Oblak, A. & Jerala, R. The molecular mechanism of species-specific recognition of lipopolysaccharides by the MD-2/TLR4 receptor complex. *Mol. Immunol.* **63**, 134–142 (2015).

121. Wlasiuk, G. & Nachman, M. W. Adaptation and Constraint at Toll-Like Receptors in Primates. *Mol. Biol. Evol.* **27**, 2172–2186 (2010).

122. Li, H. *et al.* The redefinition of Helicobacter pylori lipopolysaccharide O-antigen and core-oligosaccharide domains. *PLoS Pathog.* **13**, e1006280 (2017).

123. Vašl, J., Oblak, A., Gioannini, T. L., Weiss, J. P. & Jerala, R. Novel Roles of Lysines 122, 125, and 58 in Functional Differences between Human and Murine MD-2. *J. Immunol. Baltim. Md 1950* **183**, 5138–5145 (2009).

124. Mattis, D. M. *et al.* Studies of the TLR4-associated protein MD-2 using yeast-display and mutational analyses. *Mol. Immunol.* **68**, 203–212 (2015).

125. Kim, H. M. *et al.* Crystal Structure of the TLR4-MD-2 Complex with Bound Endotoxin Antagonist Eritoran. *Cell* **130**, 906–917 (2007).

126. Sijmons, D., Guy, A. J., Walduck, A. K. & Ramsland, P. A. Helicobacter pylori and the Role of Lipopolysaccharide Variation in Innate Immune Evasion. *Front. Immunol.* **13**, (2022).

127. Chisholm, L. O., Jaeger, N. M., Murawsky, H. E. & Harms, M. J. S100A9 interacts with a dynamic region on CD14 to activate Toll-like receptor 4. 2024.05.15.594416 Preprint at https://doi.org/10.1101/2024.05.15.594416 (2024).

128. Goutelle, S. *et al.* The Hill equation: a review of its capabilities in pharmacological modelling. *Fundam. Clin. Pharmacol.* **22**, 633–648 (2008).

129. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

130. Harman, J. L. *et al.* Evolution of multifunctionality through a pleiotropic substitution in the innate immune protein S100A9. *eLife* **9**, e54100 (2020).

131. Baker, E. P. *et al.* Evolution of host-microbe cell adherence by receptor domain shuffling. *eLife* https://elifesciences.org/articles/73330 (2022) doi:10.7554/eLife.73330.

132. Detecting the Signatures of Adaptive Evolution in Protein-Coding Genes - Bielawski - 2013 - Current Protocols in Molecular Biology - Wiley Online Library. https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/0471142727.mb1901s101.

133. Kosakovsky Pond, S. L. *et al.* HyPhy 2.5—A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. *Mol. Biol. Evol.* **37**, 295–299 (2020).

134. Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H. & Frost, S. D. W. Automated Phylogenetic Detection of Recombination Using a Genetic Algorithm. *Mol. Biol. Evol.* **23**, 1891–1901 (2006).

135. Tierney, L. The R Statistical Computing Environment. in *Statistical Challenges in Modern Astronomy V* (eds. Feigelson, E. D. & Babu, G. J.) 435–447 (Springer, New York, NY, 2012). doi:10.1007/978-1-4614-3520-4_41.

136. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).

137. BEKKER, H. *et al.* GROMACS - A PARALLEL COMPUTER FOR MOLECULAR-DYNAMICS SIMULATIONS: 4th International Conference on Computational Physics (PC 92). *Phys. Comput. 92* 252–256 (1993).

138. Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: A web-based graphical user interface for CHARMM. *J. Comput. Chem.* **29**, 1859–1865 (2008).

139. Gowers, R. J. *et al.* MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. *scipy* (2016) doi:10.25080/Majora-629e541a-00e.

140. Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* **32**, 2319–2327 (2011).

141. Wu, D. *et al.* Global, regional, and national incidence of six major immune-mediated inflammatory diseases: findings from the global burden of disease study 2019. *eClinicalMedicine* **64**, 102193 (2023).

142. The top 10 causes of death. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.

143. Tabas, I. & Glass, C. K. Anti-Inflammatory Therapy in Chronic Disease: Challenges and Opportunities. *Science* **339**, 166–172 (2013).

144. Pelletier, M., Simard, J.-C., Girard, D. & Tessier, P. A. Quinoline-3-carboxamides such as tasquinimod are not specific inhibitors of S100A9. *Blood Adv.* **2**, 1170–1171 (2018).

145. CD14 Is a Co-Receptor for TLR4 in the S100A9-Induced Pro-Inflammatory Response in Monocytes | PLOS One. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0156377.

146. Björk, P. *et al.* Identification of Human S100A9 as a Novel Target for Treatment of Autoimmune Disease via Binding to Quinoline-3-Carboxamides. *PLOS Biol.* **7**, e1000097 (2009).

147. Da Silva Correia, J. & Ulevitch, R. J. MD-2 and TLR4 N-Linked Glycosylations Are Important for a Functional Lipopolysaccharide Receptor. *J. Biol. Chem.* **277**, 1845–1854 (2002).

148. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

149. Liu, L. *et al.* Structural basis of toll-like receptor 3 signaling with double-stranded RNA. *Science* **320**, 379–381 (2008).

150. Jin, M. S. *et al.* Crystal structure of the TLR1-TLR2 heterodimer induced by binding of a tri-acylated lipopeptide. *Cell* **130**, 1071–1082 (2007).

151. Lemmon, G. & Meiler, J. Rosetta Ligand docking with flexible XML protocols. *Methods Mol. Biol. Clifton NJ* **819**, 143–155 (2012).

152. Kim, J.-I. *et al.* Crystal Structure of CD14 and Its Implications for Lipopolysaccharide Signaling. *J. Biol. Chem.* **280**, 11347–11351 (2005).

153. Chang, C.-C. *et al.* Blocking the interaction between S100A9 and RAGE V domain using CHAPS molecule: A novel route to drug development against cell proliferation. *Biochim. Biophys. Acta BBA - Proteins Proteomics* **1864**, 1558–1569 (2016).