Diversity and evolution of flagellins: insights from free-living and host-associated microbial environments with a focus on the human gut

Dissertation

Der Mathematisch-Naturwisenschaftlichen Fakultät

Der Eberhard Karls Universität Tübingen

Zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Andrea Borbón-García

aus Bogotá, Kolumbien

Tübingen

2024

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	12.04.2024
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatterin:	Prof. Dr. Ruth E. Ley
2. Berichterstatter:	Prof. Dr. Daniel Huson

Zusammenfassung

Flagellin ist eine Schlüsselkomponente des bakteriellen Flagellums. Es ist eine molekulare Struktur, die für die bakterielle Motilität wesentlich ist. In Wirten wird Flagellin von Rezeptoren des angeborenen Immunsystems erkannt. In Tieren erfüllen Toll-like receptor 5 (TLR5) und NAIP5/NLRC4 diese Funktion, während bei Pflanzen Flagellin-Sensing receptor 2 (FLS2) und Flagellin-Sensing receptor 3 (FLS3) für die Erkennung von Flagellin zuständig sind.

Bislang haben sich Studien des Flagellins vor allem auf Pathogene konzentriert und weitgehend die Vielfalt von Flagellin bei Interaktionen zwischen Wirt-Mikrobe und den kommensalen Bakterien vernachlässigt. Meine Forschungsgruppe hat festgestellt, dass das Flagellin FlaB von *Roseburia hominis*, einem Kommensalen des Darms, trotz starker TLR5-Bindung ein schwacher TLR5-Agonist ist. Dieses phänotypische Merkmal nennen wir „*stille Erkennung*“ durch TLR5. Diese Entdeckung deutet auf eine bisher unerforschte Variationsbreite von Flagellinen und deren Interaktionen mit dem Wirt hin. Darüber hinaus sind die Verteilung und Entwicklung der Diversität von Flagellin in natürlichen mikrobiellen Umgebungen nach wie vor unklar. Ziel dieser Arbeit ist es daher, die Muster der Vielfalt und Evolution von Flagellinen in freilebenden und wirtsassoziierten Lebensräumen zu identifizieren.

Für diese Forschung habe ich öffentlich verfügbare Flagellin-Sequenzen verwendet und profilierte Flagellin-Gemeinschaften in 785 metagenomischen Proben, die freilebende und wirtsassoziierte Umgebungen umfassen. Zusätzlich habe ich Anreicherungsanalysen durchgeführt und Codon-Evolutionsmodelle implementiert, um allgegenwärtige und episodische positive Selektion zu erkennen. Folglich entdeckte ich, dass freilebende Umgebungen eine größere Vielfalt an Flagellin-Genen aufweisen als wirtsassoziierte Umgebungen. Außerdem unterscheidet die Flagellin-Zusammensetzung Biome, insbesondere in freilebenden und tierassoziierten Proben. Darüber hinaus entdeckte ich verschiedene Flagellin-Kategorien, die in jedem Biom angereichert waren. Zudem fand ich Hinweise auf eine allgegenwärtige negative Selektion und eine episodische positive Selektion, die auf den konservierten Domänen dieser angereicherten Flagelline stattfand. Diese Selektionssignaturen waren überall in den Regionen vorhanden, die mit Wirtsrezeptoren interagieren. Ihr Standort war jedoch nicht mit dem Lebensraum verbunden.

Aufbauend auf der Charakterisierung freilebender und wirtsassoziierter Umgebungen strebte ich zunächst Flagellinen in menschlichen Darmmetagenomen zu charakterisieren, um die Beziehung zwischen ihrer Evolutionsgeschichte und ihrer Art der Interaktion mit TLR5 zu untersuchen. Unter Verwendung öffentlich verfügbarer Flagellinsequenzen habe ich Flagellingemeinschaften für 270

menschliche Darmmetagenome profiliert und Homologiesuchen durchgeführt. Dementsprechend konnte ich zwischen vorhergesagten stimulierenden und stillen Flagellinen unterscheiden. Meine Ergebnisse zeigten, dass der Großteil des Flagellins im menschlichen Darm von Mitgliedern der Familie *Lachnospiraceae (Firmicutes)* kodiert wird. Zudem weisen 45 % der mutmaßlichen stillen Flagellinen den vorhergesagten Phänotyp auf und die stille Art der Interaktion bei nicht verwandten Flagellenbakterien weit verbreitet ist.

Diese Arbeit erweitert unser Verständnis der Flagellin-Ökologie in freilebenden und wirtsassoziierten Umgebungen. Darüber hinaus liefert es Einblicke in die Rolle der natürlichen Selektion bei der Diversifizierung von Flagellin zwischen Bakterien. Schließlich wird ein strukturierter Rahmen für die Formulierung wichtiger Hypothesen geschaffen, die zukünftige Untersuchungen von Flagellin-vermittelten Wirt-Mikroben-Interaktionen vorantreiben werden.

Abstract

Flagellin is a key component of the bacterial flagellum, a molecular structure essential for bacterial motility. In hosts, flagellin is recognized by receptors of the innate immune system: Toll-like receptor 5 (TLR5), and NAIP5/NLRC4 in animals, and Flagellin-Sensing receptor 2 (FLS2), Flagellin-Sensing receptor 3 (FLS3) in plants. To date, the study of flagellins has primarily focused on pathogens, largely neglecting flagellin diversity in the context of host-microbe interactions with commensal bacteria. My research group recently revealed that flagellin FlaB from *Roseburia hominis,* a gut commensal, is a weak TLR5 agonist, despite strong TLR5 binding, a phenotype we named 'silent recognition' by TLR5. This discovery hints at an unexplored diversity of flagellins and their mode of interaction with the host. Moreover, the distribution and evolution of flagellin diversity throughout natural microbial environments remains unclear. Thus, this work aims to identify patterns of diversity and evolution in flagellins associated with free-living and host-associated habitats.

I used publicly available flagellin sequences to profile flagellin communities in 785 metagenomic samples encompassing free-living and host-associated environments, performed enrichment analyses, and implemented codon models of evolution to detect pervasive and episodic positive selection. I found that free-living environments comprise a higher diversity of flagellin genes than host-associated environments, and that flagellin composition differentiates biomes, particularly in free-living and animal-associated samples. Moreover, I discovered different flagellin categories that were enriched in each biome, as well as evidence of pervasive negative selection and episodic positive selection occurring on the conserved domains of these enriched flagellins. These signatures of selection were consistently present around the regions interacting with host receptors. However, their location was not associated with the habitat.

Building upon the characterization of free-living and host-associated environments, I then aimed to characterize flagellins in human gut metagenomes to probe the relationship between their evolutionary history and their mode of interaction with TLR5. Using publicly available flagellin sequences, I profiled flagellin communities for 270 human gut metagenomes and performed homology searches to differentiate between predicted stimulatory and silent flagellins. My results indicated that the majority of flagellin in the human gut is encoded by members of the Lachnospiraceae family (Firmicutes), that 45% of the putative silent flagellins exhibited the predicted phenotype, and that the silent mode of interaction is widespread across unrelated flagellated bacteria.

This work expands our understanding of flagellin ecology in free-living and host-associated environments. Moreover, it provides insights into the role of natural selection in the diversification of

flagellin across bacteria. Finally, it establishes a structured framework for formulating important hypotheses that will drive future explorations of flagellin-mediated host-microbe interactions.

Acknowledgments

I want to thank to:

The Max Planck Society for the financial support provided for developing my research.

My TAC committee: Honour McCann, Daniel Huson, Alexander Tyakht, and Ruth Ley for your time, guidance, and feedback throughout the development of this project. To my supervisor, Ruth Ley, for providing me with the resources to do my research, and an ergonomic workplace to take care of my lower back.

Daniel Huson for reviewing this thesis and for providing me with his feedback as part of my doctoral committee.

My deepest gratitude to Honour McCann for her empathy and humanity. I consider myself fortunate to have had the opportunity to invite you to my thesis committee. You made me feel genuinely heard and less isolated.

I would have not finished this document without the help of certain people that I can't thank enough: Alex Tyakht, for his eagerness to help and enthusiasm to provide feedback. I appreciate your constant help. James Marsh, Alejandra Duque, and Jacobo de la Cuesta-Zuluaga for revising this document and helping me to improve it. To Luisa Pallares, for your time and willingness to discuss my results, doodle contingency tables on pieces of paper, and discuss academic life -in Spanish- with me.

My friends in the Ley Lab made these years more enjoyable: Albanita, Jacobo, Guille, Claudi, Tanja, Jess, Xiaoying, Alina, Meghna, and Chris. My colleagues in the Lab that provided me with their insights, feedback, and countless cakes. Especially thanks to Tony for his always on-point suggestions/solutions when I had data analysis-related issues, and Guille for helping me to deal with conda, and for all the laughs and *tinticos* after lunch.

My office relocated a lot over the years (partially due to covid). But I want to thank Claudi and James for making the original office a safe space for me when I was at my most vulnerable. I missed having you as office mates. Claudi, thanks for your empathy, your kind words, and your hugs when they were highly needed. Covid shaked our routines, and with this we had to relocate to the office in the IS building. I treasure our time in that fancy office with standing desks, and the nice memories we built there: Albanita, Jacobo, Jess & Maggie, Tanja, and Taichi. You are all great people and friends. Thanks for all the coffee breaks, lunch breaks, PhD hat-making sessions, BBQs, memes, stickers, bingos, and the guaranteed fun during virtual meetings. When social distancing measurements were over and the space in IS needed, we relocated back to EBIO, and my office switched one more time. I had the fortune to land with smart, kind and funny colleagues again: Xiaoying, Mirabeau, Yihua, and Michael. Thank you Xiaoying for all the mom advice, the Chinese hotpot and snacks, and for encouraging me when I was on the finish line. I really admire the hard work you put into everything you do and how you see the positive in everything. I will miss you for sure.

Tübingen and the MPI gave me amazing friendships: Abiramchi, thanks for being a great *fren* over all these years, and for introducing me into the world of chiles secos, salsas, and sopes. My belly can't thank you enough. Albanita, I treasure your friendship very much and I feel honored to be your friend: from Franzbrötchen to Insta reels. I admire the big effort you put into finding a not-very-stinky French cheese that I might like, and the frenetic love you have for arepas and empanadas, not to mention your obsession with "Te sorprenderás", I've always enjoyed the time and laughs over the years. Isabella, for our common passion for Mexican food, and food in general, all the Asmara nights, the

trips, and the Stuttgart clubbing night that has yet to happen. Lau and Cari. I enjoyed challenging my mind to be able to tell apart who is Laura and who is Carina. Lau, thanks for helping me translate my abstract into German. Sergito Latorre, and Adri. For all the food, beers, parties and gossip.

Aleja and Jacobo: I'm grateful for crossing paths with you. Thanks for the moral support stickers and the actual support, the advice, the venting sessions, and for always checking up. Thanks, Aleja, for being postdoc-Ale and RST-Ale at the same time, and for helping me to figure out my way out. Thanks for encouraging me to write and reminding me that the best thesis is the finished one.

My beloved friends, Majo, Zai, and Ave have been there for me when I needed them – considering the time zone difference-. Thanks for being an important part of my support network, and for the WhatsApp audios that seem like podcasts. Even though we can't talk very often, every time we do, it feels like having a *tinto*.

My BCEM friends: Zai, Fore, Leda, Chiquilla, Juanse, Ave, Aurelia, Came. How good it felt the few times we were able to find a common and "reasonable" time spanning our 6 time zones. All the Monash-sponsored zoom meetings to have beers and gossip were worth the *trasnochada*. I love you all and feel grateful to have you in my life and count on your friendship.

My Colombian and Latin American friends in Tübingen: Diegui, Roger, Sebas, Cami, Martín, Gabi, Eli, Ili, Nacho, Denise, Sebas-Denise, Aaron. They made the experience of moving to a small town feel like home, with lots of parties, memes, and dancing (white people call it social dancing, as I recently learned). Diegui and Roger, thanks for all the Asmara nights until our age allowed it, and for your sincere friendship.

Mi familia: su apoyo y amor incondicional me ha mantenido siempre. He tenido la fortuna de llenarme de amor y comida deliciosa cada vez que visitaba la casa, y de verlos por videollamada cuando estamos lejos. Quedarme atrapada en Colombia por la pandemia es algo que le agradezco al covid, me regaló mucho tiempo con ustedes. A mi Susie-chorizo, que ya no está, pero ha sido el amor más lindo de mi vida.

My Leito precioso. I know it would have been much harder to finish without your love, support and delicious food. Thank you for bringing me back down to earth every time I needed it, for discussing bioinformatics (and non-bioinformatics) problems with me, for laughing at my ridiculous dances, for reminding me to celebrate the small victories, for cheering me on during the races, and for taking care of me every time I used my "frequent patient card" at the hospital.

My Krankenkasse and the Universitätsklinikum of Tübingen. I have never used my health insurance as much as I have here.

A big shoutout to my electric blanket, my supply of Colombian coffee, and all the teas that fueled this writing process and kept me warm while I embarked on the adventure of writing a dissertation during a German winter.

And last, to myself: for embracing both my strengths and vulnerabilities, and for not being afraid to speak up and seek for help when I needed it.

Table of contents

Introduction

Bacteria exhibit remarkable ubiquity in the natural world. They thrive in an array of environments, from free-living habitats like aquatic ecosystems and soils, to niches as diverse as plant rhizospheres and animal gastrointestinal tracts. In the process, they have had to adapt to highly dynamic and challenging conditions. A key structure that has enabled them to move, colonize, and interact with their surroundings is the flagellum, a motility structure chiefly composed of hundreds of units of flagellin [1]. Thus, flagellin plays a pivotal role in the interactions of bacteria with the environment, raising questions about the ecology and evolution of flagellin diversity. How is flagellin diversity distributed across bacterial lifestyles? And what role do these lifestyles play in shaping flagellin evolution?

Flagellated bacteria constitute a substantial portion of the bacterial domain [2], and flagellum-mediated motility provides bacteria with adaptive advantages, enabling them to migrate to environments with more favorable conditions [1,3]. The flagellum, primarily composed of the flagellin protein, is a helical structure involved not only in motility but also in other processes such as colonization and adhesion [3]. Although it is known that bacteria can change their flagellation patterns and flagellin gene expression in response to environmental conditions [4,5], the influence of such environments on the genetic diversity and evolution of flagellins is not yet clear. Evidence suggests that flagellin genes are prevalent among free-living bacteria, but are often lost in many host-associated bacteria [6]. Yet, the relationship between bacterial phylogeny and flagellin diversity, and the mechanisms underlying its diversification remain elusive.

Flagellin belongs to the family of Microbial Associated Molecular Patterns (MAMPs), which are recognized by Pattern Recognition Receptors (PRRs). These receptors have evolved convergently in plants and animals [7], underscoring their significance. Traditionally studied in the context of pathogenic interactions, the ecology and evolution of flagellin in non-pathogenic bacteria now represent an unexplored realm. In this context, two primary questions arise: 1) What are the patterns of molecular diversity, ecology, and evolution of flagellin among commensal bacteria and those living in free-living ecosystems, and 2) what is the intricate relationship between the evolutionary trajectories of flagellins in commensal bacteria and their interaction with host receptors. Leveraging publicly available metagenomic data offers a valuable resource for exploring and addressing these questions, providing answers that will yield insights into the intricate relationship between microbes and their environment.

To this end, the overarching aim of this research was to uncover patterns of diversity and evolution in flagellins associated with both free-living and host-associated habitats. This exploration of flagellin diversity with commensal bacteria across diverse environments lays the foundation for generating hypotheses crucial to advancing the study of host-microbe interactions in the future.

Research Objectives

In this section, I present the thesis structure and discuss the contributions I have made.

In the first chapter of my thesis, I investigate flagellin communities and their diversity patterns among free-living and host-associated environments. I hypothesized that selective pressures on flagellins might differ between free-living and host-associated lifestyles, thus shaping the associated diversity in bacterial flagellins. To address this, I used shotgun metagenomic data available in public repositories spanning several environmental and host-related sources, using a custom flagellin database. The outcome of this chapter provides a comprehensive framework to further study flagellin-mediated host-microbe interactions.

In the second chapter of my thesis, I explore the relationship between human gut commensal flagellins and their mode of interaction with TLR5. This required the development of a method for homology-based prediction of silent flagellins from publicly available sequences, based on *in-silico* and experimental evidence from pathogenic and symbiont-derived flagellins, with a particular focus on FlaB from *Roseburia hominis*. It also involved an extensive exploration of publicly available flagellin sequences to characterize flagellin communities among samples of gut metagenomes from healthy humans and profile the predicted silent flagellins across these samples. The study associated with this chapter has been published in [8]. In addition, I investigated further relationships between the evolutionary trajectories of silent flagellins and bacterial taxonomy. My findings shed light on the complex evolutionary history of a novel mechanism of immune regulation in the gut, which likely plays a role in maintaining the balance between host immunity and gut microbiota.

Background

1. Structure, assembly, and functions of the bacterial flagellum

The bacterial flagellum is a remarkable nanomachine that plays a pivotal role in the life and survival of many bacterial species. It is a filamentous organelle constructed of up to 30 different proteins [3]. Despite an impressive structural diversity of flagella and flagellation patterns in bacteria, they all share a common architecture, comprising at least three parts: the flagellar basal body (FBB), containing a rotary nanomachine called the flagellar motor; the hook, extending from the basal body and acting as a joint to transmit torque energy produced by the flagellar motor to the filament [1]; and a filament that can extend up to several cell lengths, acting as a helical propeller when rotated [1,3].

The flagellar basal body (FBB), as observed in models such as *Escherichia coli* and *Salmonella enterica*, features a rod and several ring structures: the L ring, P ring, MS ring, and C ring. The L ring incorporates into the outer membrane, and the P ring becomes part of the peptidoglycan (PG) layer, both constituted by proteins FlgH and FlgI, respectively. Importantly, the LP ring is missing in gram-positive bacteria such as *Bacillus subtilis* [9], while the MS and C rings are conserved among bacterial species. The MS ring contains the protein FliF and is a component of the rotor, while the C ring contains the proteins FliG, FliM, and FliN and is located in the cytoplasmic space. The C ring is a central part of the torque generation and serves as a switch to alter the direction of motor rotation in E. coli and Salmonella species [1,9].

The flagellar motor is formed by a rotor ring complex and up to a dozen stators around the rotor [1,10–12]. The stators are formed by multiple copies of proteins MotA and MotB [11]. These structures produce torque by sequential rotor-stator interactions coupled with ion flux through the ion channels across the cytoplasmic membrane, converting the ion flux into the mechanical power required for the motor rotation [1,11,12].

The hook is formed by multiple copies of the hook protein FlgL. In *Salmonella*, the hook contains about 120 copies of such protein. The hook is a universal joint connecting the flagellar basal body and the flagellar filament. Its length determines the bending flexibility of the hook structure, as well as the

stability of the flagellar bundle. These flexibility properties determine the ability of the cell to change swimming conditions, together with the direction of rotation of the flagellar motor.

The flagellar filament contains thousands of units of flagellin, FliC. For example, *E. coli's* flagellum contains about 30.000 FliC units. The flagellar filament extends at the cell-distal tip of the growing structure via the cytoplasmic component of the type III export apparatus [3] and behaves like a helical propeller. Although the filament in Enterobacteriaceae species such as *E. coli* and *Salmonella* are formed by copies of the same flagellin subunit (FliC), there are instances of many bacterial species encoding multiple flagellin subunit types, such as the case of *Caulobacter crescentus*, whose flagellar filament is composed of six flagellin types [1]

The assembly of the flagellum is a multi-stage and sequential process starting with the assembly of the type III secretion system (T3SS). This process is followed by the assembly of the flagellar basal body and flagellar motor. The active T3SS then recruits, processes, and translocates the axial components of the flagellum structure from the cytoplasm to the distal end of the growing flagellar structure, to assemble the flagellar hook and filament [1,3,11].

Beyond its role in bacterial motility, increasing research indicates that the flagellum is involved in various cellular processes, including adhesion, biofilm formation, and even host-microbe interactions [3,13].

2. Ecological relevance of the flagellum

The flagellum holds significant ecological relevance by influencing various aspects of microbial life. Its involvement in adhesion plays a crucial role in forming biofilms, complex microbial communities that adhere to surfaces. Biofilms contribute to nutrient cycling, water purification, and the breakdown of organic matter in ecosystems [14–16]. Additionally, the flagellum's participation in host-microbe interactions highlights its impact on symbiotic relationships and pathogenicity [17]. Understanding the ecological implications of the flagellum provides insights into the web of interactions within microbial communities, shedding light on the broader dynamics that shape ecosystems and influence environmental processes.

2.1. Motility, nutrient acquisition, and pathogenesis

Bacterial motility exists in a variety of forms, which include, swimming, swarming, gliding, and twitching. These mechanisms are facilitated by cell appendages such as flagella that rotate [3,18]. Indeed, flagella are among the most widespread motility machines in bacteria and the best-understood prokaryotic motility structure [3,18]. Flagellated bacteria possess a competitive advantage over aflagellate bacteria occupying the same niche, as they can actively navigate and explore their environments rather than depending solely on Brownian movement [3]. The significance of nutrient acquisition is thus intimately tied to the flagellum and its role in motility and exploration. The bacterial flagellum enables bacteria not only to actively move but also to navigate towards nutrient-rich environments. This dynamic interaction is crucial for the survival and proliferation of bacterial species. The ability of the flagellum to propel bacteria toward optimal nutritional sources is a strategic adaptation, enhancing the capacity to thrive in diverse ecological niches [1,3,15].

Efficient nutrient acquisition influences bacterial physiology, affecting cell growth and division, gene expression, and the production of virulence factors [19,20]. Therefore, the connection between flagellar motility and nutrient acquisition affects all stages of the bacterial life cycle and cell dynamics. In the context of pathogenic bacteria, the flagellum's role in motility becomes intertwined with their ability to navigate and access host-derived nutrients, contributing to their pathogenicity and success in colonizing host environments [3]. In commensal bacteria, colonization and accessibility to nutrients lead to bacteria being able to convert unavailable nutrients and make them useful for the host [21–24]. For instance, soil microbes can acquire insoluble nutrients for plants by their extensive flagella, and they can convert these into nutrients available for the host [25].

Host tissues are an excellent source of nutrients for bacteria. These diverse sources of nutrients are part of the symbiotic and pathogenic relationships established with bacteria. Both pathogens and commensals have evolved mechanisms to access host nutrients and to propel themselves towards nutrient-rich niches [20]. Among these, flagellar motility provides a clear advantage for bacteria to reach different internal tissues and subsequently thrive in the host environment. Plant pathogens colonize plant surfaces such as the phyllosphere and rhizosphere, and most of them access internal tissues of the plants, such as the vascular elements to acquire more nutrients and avoid unfavorable abiotic conditions [25]. For example, the bacterium *Ralstonia solanacearum*, a plant pathogen that thrives in

xylem vessels of host plants -a low-nutrient and high-flow environment- uses flagellar motility and chemotaxis to locate host roots and migrate towards the xylem [26].

2.2. Biofilm formation and adhesion

Bacteria exhibit two distinct lifestyles: either as independent planktonic cells or as part of organized surface-attached microbial communities, called biofilms [14]. The flagellum also plays a pivotal role in biofilm development, leveraging its mechanosensor and adhesin properties to initiate biofilm formation and facilitate surface attachment [14,15]. In aquatic oligotrophic environments, where nutrients are scarce, bacteria accelerate their flagellar motility to scavenge nutrients for survival. This strategy improved the foraging efficiency and thereof early-stage biofilm formation [15].

Extending beyond its adhesive function, the flagellum continues to influence the maturation phase of biofilms, impacting their architectural integrity [27]. Biofilms are composed of multicellular nonmotile aggregates typically formed by motile bacteria [28]. The process of bacterial biofilm development initiates with a cell adhering to a surface, initiating the transition from a free-swimming state to adhesion, a process often referred to as a 'swim-or-stick' switch. This switch is regulated by a sensory transduction mechanism known as surface sensing, which involves the rotating flagellum [14]. In *Caulobacter crescentus* as well as other Alphaproteobacteria such as *Asticcacaulis biprosthecum* and *Agrobacterium tumefaciens* [29], the inhibition of the flagellar motor is involved in biofilm development: this transition results from the encounter of the cell with a surface that inhibits the flagellar rotation, then the bacterial cell immediately produces holdfasts and attaches to the surface [14,29].

Advancements in understanding the molecular mechanism by which the flagellum orchestrates biofilm formation and adhesion offer insights into the adaptive strategies bacteria employ in diverse ecological niches. Early findings in *E. coli* showed that mutants deficient in biofilm formation had defective flagellar functions [30]. Likewise, Wood et al. 2006 found *E. coli* strains with impaired flagellar motility to form less stable and flatter biofilms [27]. The flagellum has been reported to function as adhesins in pathogens, including *E. coli*, *Pseudomonas aeruginosa,* and *Clostridium difficile* [31]. Moreover, recent findings show methylation in surface-exposed lysine residues of flagellin in *Salmonella* Typhimurium, which increased their hydrophobicity, thus promoting the bacterial invasion of epithelial cells, by enhancing flagella-dependent adhesion [32]. Together, these findings highlight the role of the flagellum in adhesion and biofilm formation.

2.3. Environmental sensing

Motile bacteria possess the ability to detect and respond to environmental changes, adjusting their movement towards more favorable conditions. This directed movement along chemical gradients, known as chemotaxis, is one of the most extensively studied bacterial behaviors [33]. As a result, many biochemical and biophysics models exist to describe the environmental stimuli and the underlying molecular and genetic mechanisms associated with chemotaxis in models such as *Escherichia coli* and several other model bacteria [33]. Therefore, the flagellar gene regulation stands as a response to the need of bacteria to find sources of nutrients and locate niches that are optimal for bacterial growth. Likewise, chemotaxis also improves the efficiency of environmental colonization by motile bacteria compared to their aflagellate counterparts. For example, bacteria that have evolved mechanisms to utilize pollutants as a source of nutrients have also evolved chemotaxis mechanisms to efficiently locate and move toward these niches [34].

Swimming motility and chemotaxis become crucial in bacteria living in low-nutrient environments. These mechanisms are common in soil bacteria, particularly those living in the rhizosphere (the area surrounding plant roots), which are enriched with root exudates and leaf-derived metabolites such as organic acids, carbohydrates, sugar alcohols, amino acids, plant hormones, and even flavonoids, and phenolic compounds [35,36]. Chemotaxis and flagellar systems assist pathogens in localizing to their preferred sites of infection, such as wounds or open stomata [33,36], or accumulate near nutrient-rich areas around plant roots [33,36]. Accordingly, plant pathogens and symbionts are nearly all motile and possess nearly twice the number of chemoreceptors as motile/animal human pathogens [33]. The same advantages affect symbiotic communities in the context of nutrient acquisition, not only facilitating colonization but also the transmission of microbial symbionts between hosts, as observed in marine animals such as squids and cuttlefish [37]. Therefore, flagellar motility and chemotaxis deeply influence processes associated with host infection and symbiosis, as well as the community dynamics in free-living systems.

3. Evolution of flagellar systems in bacteria

Flagella are widely distributed in many phylogenetically unrelated phyla in the Bacteria domain. The broad distribution of flagellar systems suggests that this type of motility is robust and adaptable.

Indeed, several variations of flagellar motility such as spirochetes swimming, swarming, and swimming in magnetobacteria have been observed [2].

The mechanisms and the nature of the evolutionary trajectories of bacterial flagellar systems are the subject of ongoing debate. Nevertheless, advances in genomics data and analysis, and technologies in various fields combining structural biology, imaging, and molecular genetics have exposed the vast diversity of flagellar systems in bacteria and greatly increased our understanding of their biology [2]. The ancestral flagellar system likely originated from simpler structures that served different purposes, such as the secretion systems used for injecting proteins into host cells [38,39].

One hypothesis suggests that bacterial flagella share a common ancestor with the injectisome of the type III secretion system (T3SS), and have since evolved differently from each other [38,39]. This system is involved in injecting proteins into host cells during infection [38]. The components of the type III secretion system share homology with the components of the basal body of bacterial flagella, supporting the idea that these structures have a common evolutionary ancestor [38,40]. Over time, gene duplications and modifications have led to the development of sophisticated and diverse flagellar systems observed in different bacterial species today [41].

Bacterial flagellar systems are a good example of the evolution of complex systems from simpler components. Liu & Ochman identified an ancient core set of 24 structural flagellar genes that were present in the complete genomes of 41 flagellated species from 11 bacterial phyla, which they proposed to be present in the common ancestor of all Bacteria and suggested a stepwise formation of the flagellar system through duplication and modification of precursor genes [41]. Although this hypothesis has been highly criticized [42], the results showcase the immense diversity of flagellar genes across bacterial phyla.

The evolution of bacterial flagellar systems also involves lateral gene transfer, a process where bacteria acquire genes from other bacteria in their environment [42]. This horizontal transfer of genetic material has played a significant role in shaping the diversity of the T3SS [39] and flagellar systems across bacterial species [42]. As bacteria encounter different ecological niches, the selective pressures they face drive the adaptation and modification of their flagellar systems to optimize their survival and fitness in specific environments. Overall, the evolution of bacterial flagellar systems is a dynamic and intricate process that showcases the remarkable ability of bacteria to evolve and thrive in diverse habitats.

4. Flagellar biosynthesis and regulation

Flagellar motility is expensive for cell economy, therefore the synthesis of flagella is a highly regulated process, affected by external factors including abiotic conditions, growth phase as observed in *Legionella pneumophila* [43], *Enterobacter*, and *Pseudomonas* strains [44], and the interaction of bacteria with their host [45]. The underlying genetic basis of flagellar biosynthesis has been extensively studied in enterobacteria [45], especially in *E. coli* and *S. enterica.* Flagellar genes are grouped in several clusters on bacterial chromosomes, involving nearly 50 genes in *E. coli* and *S. enterica*, clustered in more than 10 operons along three distinct chromosomal regions: clusters I, II, and III [46–48]. The flagellar biosynthesis starts with the transcriptional regulation of flagellar genes, which are organized into hierarchical classes [10,45]. Class I genes initiate the transcription of class II genes, responsible for producing basal body and hook components. The final step in this hierarchical process involves the transcription of gene cluster III, which encompasses the fliC gene that codes for flagellin, along with the flagellum capping protein (fliD) and the flagellum-specific sigma factor (fliA) [45].

Posttranslational modifications, such as methylation and glycosylation may fine-tune flagellin function. The export of flagellin to the exterior environment is facilitated by the flagellar type III secretion system, a complex molecular machinery that spans bacterial membranes and efficiently exports flagellar components into the growing filament [9].

Flagellin biosynthesis involves feedback regulation mechanisms. When a sufficient number of flagella are assembled, feedback loops may down-regulate flagellar gene expression, preventing unnecessary energy expenditure [49]. Environmental cues further modulate this process, allowing bacteria to adapt their flagellar expression based on factors like temperature, nutrient availability, and specific signaling molecules in their surroundings.

5. Diversity of flagellated free-living bacteria

The diversity of flagellated free-living bacteria is extensive, showcasing a remarkable range of adaptations to diverse ecological niches. Across bacterial phyla, flagellation is a widespread mechanism for motility, allowing these microorganisms to actively explore and exploit their environments. More than 80% of known bacterial species are known to produce flagella [50].

Although the basic structure of the flagellum is widely conserved among taxa, one important aspect of its diversity lies in the structural variations of flagella, between and within bacterial lineages, such as the number of flagella per cell, the arrangement of flagella on the cell surface, the number of genes involved in the production and regulation of flagella [50,51]. These variations can be species-specific adaptations that optimize movement in specific environments [52].

Bacterial species exhibit a spectrum of flagellar arrangements that enable movement along surfaces. These arrangements include polar flagella at one pole of the cell (monotrichous), both poles of the cell (amphitrichous), many polar flagella at one or both poles of the cell (lophotrichous), or several polar flagella at both poles of the cell (peritrichous)[4,48]. Dual flagellar systems have been also described in species of the genera *Vibrio* and *Aeromonas*, enabling them to carry out different modes of motility, thus increasing their competitive advantages at the expense of a high energetic cost [45,50,53].

Moreover, the diversity of flagellated free-living bacteria extends beyond physical characteristics to encompass genetic and metabolic diversity. Different bacterial species may possess unique sets of genes related to flagellar biosynthesis and regulation, allowing them to respond to environmental cues with specific motility strategies [54]. Different flagellar regulatory cascades exist, classified into lateral and polar types. The occurrence of these cascades in diverse bacteria seems to be intricately linked to the functional characteristics of a particular flagellar insertion type. These functional characteristics are primarily determined by the environmental niche rather than the phylogenetic relationships between bacteria. Lateral and polar flagellar systems are widely distributed among Proteobacteria and may even simultaneously function in some bacteria, as observed in *A. brasiliense* and *Vibrio parahemolyticus* [45]

The diverse flagellar repertoires enable bacteria to thrive in various habitats, from aquatic environments to soil and within host organisms. For example, members in the order Enterobacterales have been isolated from a wide range of environments, including air, soil, and water. Likewise, they include some well-known plant and animal pathogens [50]. This ecological success can be partially attributed to the flagellar motility and the diversity of flagellar systems [48,50].

6. Flagellin: the building block of the flagellum

Flagellin serves as the fundamental building block of the bacterial flagellum, comprising the major structural component of the flagellar filament, together with the FliD protein at the tip of the filament [55,56]. The flagellar filament is composed of between 100-20,000 flagellin monomers [7]. Flagellin is a highly conserved protein that exhibits remarkable structural stability and versatility. Its primary role is to polymerize into a helical structure, forming the long, rigid filament that extends from the bacterial cell surface [7,51,55]. Beyond its structural role, flagellin also plays a crucial role in the host immune response, as it is recognized by pattern recognition receptors, such as Toll-like receptors, triggering innate immune reactions [55,57,58]. The versatility of flagellin across bacterial species underscores its significance as a key molecular player in both bacterial motility and host-microbe interactions.

6.1. Flagellin structure and molecular diversity

Flagellins are elongated structural proteins forming the flagellar filament, exhibiting a conserved yet diverse molecular architecture [56]. Structurally, the flagellin monomer is organized into four distinct domains: D0, D1, D2, and D3 domains. Domains D0 and D1 are located in the C-terminus and N-terminus ends of the protein and are highly homologous among species. In the folded protein, they build the core of the filament. whereas D2 and D3 -forming the hypervariable region (HVR)- which are located in the center of the polypeptide, form the most peripheral area on the flagellum surface [59,60]. Evidence on the flagellin FliC from *Salmonella* Typhimurium shows a protein folded in the shape of the upper-case Greek letter gamma (Γ). The vertical arm is formed by the coiled-coil domains D0 and D1, which are separated from each other by a 'spoke region'. Domains D2 and D3 consist mostly of B-strands forming the horizontal arm of the molecule [52]. The HVR makes the outer domains significantly different in composition and length, even within strains of the same species [52,56,59,61]. On the other hand, D0 and D1 are highly conserved among bacterial clades and contain specific motifs that mediate the binding and activation of host innate immune receptors [55,62,63].

6.2. Flagellin size variability

The size of flagellin exhibits significant variability among bacterial species. Their molecular weight ranges from 26 kDa in *Bacillus cereus* to 115 kDa in *Desulfotalea psychrophila* [52,56]. This vast size range is explained by the differences in the size of the HVR, which is almost absent in *Bacillus cereus* and contains nearly 1000 residues in *Desulfotalea psychrophila.* The D0 and D1 domains facilitate numerous

intra and inter-subunit interactions, thereby limiting their diversity. In contrast, the surface-exposed HVR is more flexible to exhibit a broader range of variations. Evidence from 202 flagellin sequences underlines the remarkable conservation of the D0 and D1 domains, where only the loop between ND1a and ND1b and the B-hairpin -connecting the ND1 to the HVR- shows variations from two to nine residues between species [52]. Based on this flagellin diversity screen, several generalizations have been made regarding flagellin primary structure: 1) the minimum length for any flagellin is about 250 residues, 2) all flagellins contain a conserved block of nearly 140 residues from the start codon to the ND0, ND1 subdomains and the B-turn, and 3) all flagellins contain a conserved block of ~90 residues in the C terminus, corresponding to the CD1 and CD0 subdomains [52].

6.3. Functional significance of flagellin's structural diversity

The N-terminal and C-terminal regions display a great proportion of conserved hydrophobic residues, highlighting their importance in maintaining the coiled-coil interaction. In addition, these hydrophobic residues are also important to enable hydrophobic interactions with the D0 domains of adjacent flagellins, thus maintaining the structure of the flagellar filament [52]. The residues in the spoke region, separating the D0 and D1 domains, show the highest degree of conservation among all flagellins. This conservation is likely due to their role in maintaining the integrity of the inner tube within the filament structure. This is achieved by tightly packaging adjacent monomers through inter-domain interactions [52].

Residues universally conserved in flagellins play a critical role in influencing the overall conformation of the filament. Moreover, residues that are exposed in the inner channel of the filament can have variable yet conserved hydrophilic residues, thus facilitating the translocation of flagellin monomers during the flagellar filament assembly [52].

6.4. Functional roles of flagellin

Flagellin's role is primarily associated with bacterial motility providing structural integrity to the flagellar filament. It forms the helical structure of the filament, allowing the flagellum to function as a whip-like appendage for motility [1,3]. As discussed previously, this type of motility is crucial for the survival and adaptability of bacteria in diverse ecological niches. However, flagellin's role also extends

to host-microbe interactions, particularly with the innate immunity of plants and animals, both vertebrates and invertebrates, as extensively investigated in various studies [51,64].

In animals, the interaction between flagellin and host cells involves the recognition of conserved regions of extracellular flagellin monomers through Toll-like receptor 5 (TLR5), and intracellular flagellin through NAIP5-NLRC4 [65]. Similarly, plants recognize flagellin through the Flagellin sensitive 2 (FLS2) receptor [66], Flagellin sensitive 3 (FLS3) receptor [67], and FLSx [68], triggering proinflammatory and adaptive immune responses [55,62,69]. These interactions with innate immunity are considered important for priming and regulating adaptive immune responses [55,70], impacting not only host-pathogen but also host-symbiont interactions.

One example of flagellin-mediated host-pathogen interactions comes from recent research on *Salmonella,* showing that the methylation of surface-exposed lysine residues of flagellin monomers enhances adhesion, thus increasing gut colonization in a gastroenteritis mouse model and cell invasion in an *in-vitro* model [32]. In contrast, evidence from the human gut commensal *Roseburia hominis* shows an immunomodulatory effect of their flagellins using murine and *in vitro* models [71]. The monocolonization of mice with *R. hominis* correlated with the upregulation of motility and chemotaxis-related genes in the bacteria and immunity-related genes in the host. In addition, it showed a protective effect in DSS-induced colitis mice models. Due to the flagellin's immunomodulatory role, there is growing interest in using it as an adjuvant in human vaccines to stimulate humoral and adaptive immune responses. [72]. Therefore, flagellin has a major significance in the host-microbe interaction for both pathogen and commensal bacteria.

6.5. Flagellin evolution and microbial fitness

Given the functional roles of flagellin in bacterial motility and its interaction with the surroundings, its evolution might have played a crucial role in shaping microbial fitness and adaptability. Flagellin evolution is also intertwined with host-pathogen interactions, as changes in its composition and structure can influence immune recognition, thus enabling bacteria to evade or manipulate the host's immune responses [48,73].

Evasion of immune recognition is crucial in pathogenic bacteria, as observed in α-, and ε – Proteobacteria [74]. Examples within these taxa, including human pathogens such as *Helicobacter pylori, Campylobacter jejun*i, and *Bartonella bacilliformis,* whose flagellar motility is essential for efficiently

infecting mammalian hosts; they exhibit mutations in the TLR5 epitope that allow them to evade the innate immune system of the host. These mutations alone would ameliorate motility, as observed in *Salmonella* mutants, therefore these bacteria counterbalance them with additional mutations throughout the flagellin to ensure the maintenance of motility. Therefore, a balance between flagellin variation and the maintenance of motility is required to efficiently colonize mammalian hosts [74].

The evolution of flagellin is shaped by the bacterial requirement to adapt to changing environmental conditions. This is exemplified by *Salmonella,* where flagellin variation is influenced by factors such as recombination prevention and the presence of a second flagellin locus [75]. In essence, the evolution of flagellin reflects the dynamic interplay between microorganisms and their environments, contributing significantly to microbial fitness and survival strategies.

In Enterobacteriaceae species, amino acid substitutions in the hypervariable region (HVR) of the flagellin generate antigenic diversity, while the ends of the flagellin polypeptide are conserved and more stable markers of flagellin evolution [75]. The antigenic diversity generated by flagellin variation in the hypervariable region has been widely used to identify serotypes, and diagnose and track infections [59,75]. *Escherichia coli* has 53 recognized antigens, while *Salmonella enterica* encodes 114 different serotypes [51,76,77].

While flagellin sequence variation enables bacteria to manipulate the host response, it comes at a functional cost to motility, affecting navigation within host environments. Therefore, the evolutionary dynamics of flagellin diversity involves a trade-off, where overcoming host recognition may negatively impact bacterial motility, requiring additional mutations to counterbalance such limitations and maintain virulence, as observed in Proteobacteria [74]. Understanding these fitness trade-offs is crucial for unraveling the complex strategies bacteria employ to navigate the dynamic interplay between host immunity and their survival.

Dalong and Reeves (2020) conducted a comprehensive analysis revealing significant flagellin diversity within the hypervariable region (HVR) of various prokaryotic organisms [59]. Their research identified areas within the HVR that lacked homology across different sequences, indicating a high degree of heterogeneity. Furthermore, a detailed comparison of flagellin sequences from *Escherichia coli* and *Salmonella enterica* uncovered two distinct layers of variation within the HVR: firstly, sequence variations that delineate the H-antigen groups specific to *E. coli* or *S. enterica*, and secondly, variations within these groups themselves. The unusual diversity observed in the evolution of flagellins, as noted

by Dalong and Reeves, is partly attributed to horizontal gene transfers across different phyla, highlighting the enormous diversity and evolutionary complexity of flagellins among bacterial species [59]. This finding emphasizes the challenges and intricacies involved in understanding the evolutionary dynamics of flagellins.

7. Flagellin as a virulence factor in pathogens

The success of pathogenic bacteria in low-nutrient, physically and biochemically dynamic niches depends on strategic adaptations and metabolic specialization, such as enhanced adhesion and the production of virulence factors [26]. Among these, flagellin plays a crucial role in various pathogens targeting plant and animal hosts, providing them with clear fitness advantages. However, these advantages come at the cost of disrupting host homeostasis by invading epithelial cells, translocating across epithelial barriers, and forming biofilms, thus presenting a complex interplay in the host-pathogen relationship [17].

In plant hosts, pathogens exploit pre-existing openings like stomata, nectarthodes, hydathodes, and lenticels, as well as tissue abrasions, to gain access to the plant's interior, where they can reach nutrient-rich conditions and proliferate rapidly [25]. In the gastrointestinal tract of animals, pathogens utilize flagellar motility to access cells, translocate to nutrient-rich tissues, and resist clearance by fluid flow [17]. In humans mucosa, for example, flagellar motility enables counterbalancing the upward flow of mucus from the bronchial epithelia or the peristalsis in the intestine, to achieve colonization. Motility coupled with chemotaxis mechanisms successfully allows pathogenic bacteria to target specific mucosal tissues, as observed in *Helicobacter pylori* and *Pseudomonas aeruginosa*, which can find and colonize the stomach and lungs, respectively [78–80]. Similar examples are evidenced in *Vibrio cholerae* and *Salmonella Typhimurium*, which rely on flagellar motility to colonize host tissues [7].

Flagellin's role in bacterial pathogenesis is not restricted to motility and localization of target tissues. Emerging evidence from pathogenic bacteria, including *Clostridium difficile* in mice cecal tissue, *P. aeruginosa* in the airway lumen, and enteropathogenic *E. coli* in the intestinal mucosa shows their flagellins to bear adhesin-like properties. These adhesive properties are suggested to be located in the hypervariable region of the flagellin protein, suggesting to be a serotype-dependant phenotype [7].

Beyond its role in motility and adhesion, flagellin directly interacts with the host's innate immune system, triggering inflammatory responses. Pattern-recognition receptors (PRRs), including Toll-like

receptor 5 (TLR5), and Flagellin-sensitive 2 receptor (FLS2), present on the surfaces of certain host cells in animals and plants, respectively, specifically recognize flagellin. This recognition activates signaling cascades leading to the production of pro-inflammatory cytokines and systemic defense mechanisms [17]. While the immune response aims to eliminate the pathogen, some bacterial species have evolved mechanisms to manipulate or evade this recognition, allowing them to establish persistent infections [7,17,81]. The ability of flagellin to simultaneously contribute to bacterial motility and provoke immune responses underscores its significance as a multifaceted virulence factor in the context of pathogenicity [17].

8. Flagellin perception systems in eukaryotes

Host-associated, flagellated bacteria face a challenge absent in many other free-living environments: a host actively sensing and restricting flagellar motility [17]. Hosts have a repertoire of mechanisms to avoid microbial colonization. First, physical barriers, renewal of epithelial cells, and the presence of antimicrobial peptides such as lysozymes and defensins protect the host against microbial colonization. Microbes on the other hand have evolved mechanisms to overcome such barriers. For this reason, several types of sentinel cells play an important role in detecting potential pathogens and inducing an innate immune response against them. They can sense pathogens by pattern-recognition receptors (PRRs), which detect conserved microbe-associated molecular patterns (MAMPs). These molecular patterns include flagellin, present in both pathogenic and non-pathogenic bacteria of both plant and animal hosts [7,55].

8.1. Flagellin perception systems in vertebrates

As a part of the innate immunity of animals, the Toll-like receptor (TLRs) family, a group of leucine-rich repeats receptors, can detect intra- and extracellular MAMPs. This is a family of transmembrane proteins functioning as homo- and heterodimers. They consist of three subdomains: a leucine-rich repeat (LRR) directly involved in MAMP recognition, a membrane-spanning motif, and a Toll-Interleukin (IL)-1 receptor domain (TIR) required for signal transduction upon MAMP recognition [7]. Among TLRs, TLR5 specifically recognizes bacterial flagellin. Upon flagellin recognition, TLR5 dimerizes and induces MyD88-dependent signaling leading to the activation of proinflammatory transcription factor NF-kB in epithelial cells, monocytes, and dendritic cells, contributing to the immediate clearance of pathogens from the host [7,55,63,64,82]. TLR5 detects a specific conformation of the flagellin domain D1 [7].

Studies on the structural and mechanistic basis of TLR5-flagellin interaction found that the residues in a protruding loop of LRR9 of the TLR5 interact with three helices corresponding to a conserved motif of the flagellin D1 domain [63]. This surface in the TLR5 is conserved across vertebrates and structurally homologous to the region where other LRRs bind their ligands [82]. On the other hand, studies have shown residues scattered across the N- and C-terminal parts of the flagellin, which are involved in TLR5 binding and signaling. Among these, a conserved motif in the D1 (LQRIRELAVQ) seems to be responsible for TLR5 high-affinity binding and signaling, whereas the conserved motif in the D0 (LGAIQN) contributes to signaling but has little effect on binding [51,58,63]. The conserved motif in the D1 is present in diverse flagellins from y-Proteobacteria and Firmicutes bacteria [55].

Similarly, the intracellular human receptor NLRC4 (also known as IPAF), a member of the NOD-like receptor (NLR) family, enables sensing cytosolic flagellin. Flagellin can gain access to the host cytosol during bacterial infection, through protein transport systems, such as T3SS in *Salmonella* and T4SS in *Legionella* [83]. Upon sensing flagellin via the neuronal apoptosis inhibitory protein (NAIP), the NAIP-NLRC4 inflammasome assembles, later inducing a proinflammatory response through the activation of Caspase-1 and the release of proinflammatory cytokines [82,84,85]. Highly conserved regions of 35 residues and 52 amino acids in the C-terminus and N-terminus of the flagellin, respectively, are necessary and sufficient to induce the NLRC4-mediated immune response [86–88]. The cytosolic recognition of flagellin is a crucial mechanism for detecting cell invasion and mobilizing a rapid and localized response.

The importance of the animal host perception of flagellin is exemplified by TLR5/NLRC4 deficient mice, which are unable to efficiently detect flagellin, thus being more susceptible to chronic inflammation and infection by enteric pathogens [17,89,90]. Likewise, humans with a defective TLR5, produced by a stop codon ($TLR5^{329STOP}$) are more susceptible to Legionnaire's disease, caused by the flagellated bacteria *Legionella pneumophila* [91] and exhibit inflammatory phenotypes as shown by Merx and colleagues [92]

8.2. Flagellin perception systems in plants

In plants, the PRRs flagellin-sensitive receptor 2 (FLS2) which is present in many species of angiosperms and a gymnosperm [93,94], and flagellin-sensitive receptor 3 (FLS3), present in certain solanaceous plants, including tomato, potato, and pepper [93], are activated upon the recognition of

flagellin motifs flg-22 and flg-28, respectively. FLS2 is a transmembrane protein composed of a leucine-rich repeat (LRR) domain and an intracellular signaling domain, which functions as a serine-threonine kinase. The recognition of flg22, a conserved N-terminal 22 amino acid peptide in the D0 domain (RLSTGSRINSAKDDAAGLQIA) [51], triggers the association of FLS2 with the BRI1-associated receptor kinase (BAK1) to activate the Pattern-triggered immunity (PTI). Likewise, FLS3 associates with BAK1 after infection, thus activating the PTI. This immune response leads to the generation of numerous systemic defense mechanisms, including mitogen-activated protein kinase (MAPK) cascades, the production of reactive oxygen species (ROS) and nitric oxide, together with extensive transcriptional changes that hinder the infection process [67,81,95].

While both FLS2 and FLS3 depend on BAK1 for initiating PTI activation, the evidence of an extended production of ROS following FLS3 activation suggests that FLS2 and FLS3 utilize distinct molecular mechanisms to trigger PTI in response to flagellin recognition [96]. In addition to FLS2 and FLS3, an as-yet-unknown receptor (FLSx) has been reported in monocotyledons, which recognizes another flagellin epitope of about 35 amino acid residues in the flagellin C-terminal region [68,97–99]. Unlike mammals, plants lack cytoplasmic flagellin receptors, as has been demonstrated by artificial infections into the cell cytosol of *Nicotiana benthamiana* [100]. It is hypothesized that this lack of cytoplasmic flagellin receptors evolved as a defense mechanism of plants against necrotrophic pathogens [51].

Interestingly, although TLR5 and FLS2 possess LRRs, these are not homologous. Moreover, their flagellin epitopes are also different, suggesting an independent evolution of the detection of flagellin from potentially harmful bacteria, thus highlighting their importance in the innate immunity of plants and animals [7].

9. Overcoming host recognition of flagellin

Host tissues are crucial for microbial survival. A key step in the success of bacterial colonization depends on their ability to avoid and manipulate the host's immune response. Since flagellin is a potent elicitor of animal and plant immune responses, host-associated bacteria have developed mechanisms to overcome adaptive immune responses and therefore promote successful colonization [81].

Over the last few years, our understanding of the mechanisms employed by bacteria to evade and manipulate immune recognition and overcome flagellin-triggered immunity has notably advanced

[81,101]. Post-translational modifications, modulation of flagellin transcription, enzyme-mediated flagellin degradation, and flagellin sequence polymorphisms are among the main mechanisms that bacteria utilize to overcome flagellin-triggered immunity [51,81,101].

Polymorphisms in MAMP sequences are a frequently used strategy to avoid detection. Several examples of plant-associated bacteria suggest a spectrum of flagellin-triggered immune responses as a consequence of flagellin sequence variation. *Xanthomonas campestris pv campestris* exhibit intra-species flagellin variation that impacts the strength of the FLS2-dependent responses [102]. Cheng and colleagues (2021) found that most of the ε-, δ-, and α-Proteobacteria induced moderate, weak and no responses, respectively, in *Arabidopsis thaliana*. However, most members of the γ and β-Proteobacteria induced strong flagellin-induced responses [103].

Recent studies have shown the crucial role of flagellin polymorphisms in plant-associated bacteria, as a mechanism to modulate the PAMP-triggered immunity upon flagellin recognition by the host. Instead of evading FLS2 recognition, many bacteria produce polymorphic forms of flg22. These flg22 variants possess a conserved N-terminal motif, allowing flagellin to bind to the FLS2. However, their last five residues in the C-terminus are polymorphic, inhibiting the interaction of FLS2 with BAK1 and thus abolishing the FLS2-BAK1-mediated signaling [73,104]. As a result, diverse stimulatory phenotypes are observed in root commensal communities, thus highlighting the importance of flagellin polymorphisms in modulating the plant immune response and microbial community assembly and structure [73,104].

Evasion of recognition caused by flagellin polymorphisms has also been described in animal pathogens. Members of the ε-Proteobacteria such as *Helicobacter pylori, Campylobacter jejuni*, and *Bartonella bacilliformis* exhibit mutated TLR5 epitopes, completely abolishing TLR5 recognition. To counterbalance the negative impact on motility, they present additional mutations in the HVR [74]. In addition, the packaging of the D1 domains in the flagellin protofilament of *C. jejuni* differs from the packaging in *S.* Typhimurium, providing them with an advantage to evade the TLR5 recognition [105].

Another example of flagellin polymorphisms is observed in *S. enterica*. The majority of the species encode two flagellin types (phase 1 and phase 2), and experiments with mice models show controlled levels of infection when *S*. Typhimurium expressed phase 2 flagellin [51]. Other reports have also shown host discrimination between commensal and pathogenic bacteria, as observed between pathogenic and commensal *E. coli* strains in mice models [61], where differences in the HVR composition dictate

commensal properties of *E. coli* strains. This was also observed between nonpathogenic *E. coli* K12 and pathogenic strains of *S.* Typhimurium, where nonpathogenic strains elicited a less pronounced flagellin response [106].

Mechanisms independent of polymorphic flagellin are also observed across a spectrum of pathogenic and non-pathogenic bacteria. These mechanisms specifically prevent flagellin detection, by targeting host immune receptors or masking the flagellin itself [51,81]. *Pseudomonas aeruginosa* strategically secretes alkaline protease (AprA) to degrade monomeric flagellin, effectively preventing immune activation in both animals and plant hosts [107]. The endophyte *Bacillus subtilis* produces the lantibiotic subtilomycin to reduce the plant's defensive response [108]. This compound binds self-produced flagellin, preventing the plant's immunity from detecting it. This mechanism may be relevant for endophyte adaptation while providing insights into the fine-tuning mechanisms of beneficial endophyte bacteria with host plants, contributing to the understanding of plant microbiota assembly [108].

10. Role of flagellin in modulating adaptive immunity in animals

Flagellin's role in immunity goes beyond innate immunity. In animals, the adaptive immune response triggered upon recognition of flagellin produces flagellin-specific antibodies and T-cells to effectively clear potential pathogens [17]. The loss of innate immune recognition of flagellin is associated with a decrease in levels of anti-flagellin antibodies, as previously observed [109,110]. Moreover, TLR5-deficient mice exhibit elevated levels of proinflammatory cytokines compared to wild-type mice, while having increased abundances of Proteobacteria, specifically enterobacteria species [90]. Cullender and colleagues also observed that TLR5-deficient mice are unable to produce an adaptive immune response to flagellin. [89].

Pioneering studies with *S. adelaide* showed its power to stimulate antibody responses in rats [111]. Flagellin induces the maturation of TLR5-expressing dendritic cells (DCs) and the upregulation of co-stimulatory molecular and antigen-presenting capacity in a MyD88-dependent manner [17,111]. Therefore, the presence of a flagellin-specific antibody repertoire is largely dependent on TLR5 expression. As shown by Cullender and colleagues, TLR5 facilitates the production of flagellin-specific antibodies and is inversely correlated with the expression of flagellar genes in commensals from the gut microbiota [89]. Accordingly, flagellin priming of innate immunity is crucial for the proper

maturation of adaptive immunity, which can modulate the microbiome's production of flagella and maintain the integrity and homeostasis of the mucosal barrier [89].

In a healthy human gastrointestinal tract (GI), plasma cells constantly secrete commensal-specific IgA and IgG into the intestinal lumen, and a fraction of these antibodies are flagellin-specific [112]. Metatranscriptomics data show that only a small fraction of flagellin-encoding bacteria in the healthy human gut express flagellar genes [89]. This observation suggests that host mechanisms exist to sense and control flagellar motility from gut commensals.

Studies on *Roseburia hominis*, a human flagellated symbiont, show bidirectional regulations of gene expression. Patterson and colleagues observed that monocolonization of mice with *R. hominis* resulted in the upregulation of motility genes in *R. hominis*, and upregulation of TLR signaling -including TLR5-, and T cell-related genes in the host, indicating that TLR5/flagellin signaling is an important mediator of T cell expansion. The immunomodulatory effects of *R. hominis* also protected against DSS-induced colitis. They also observed downregulation of the NF-kB pathway, involved in proinflammatory cascades, thus contributing to the immune homeostasis [71]. This evidence on the cross-talk between host and commensals illustrates the coevolution of bacterial motility and flagellin-perception systems, where commensals can act as immunomodulatory agents while the host can affect bacterial motility.

11. Coevolution of MAMPs and host's PRRs

Plant and animal hosts must navigate intricate environments, filled with diverse pathogens and symbionts. Simultaneously, microbial communities residing in host environments encounter uncountable challenges to successfully colonize, overcoming host defense mechanisms. The adaptation to these complex environments has been a major aspect of ecological adaptations for both players in host-microbe interactions [113]. Viewing the flagellin-PRR relationship as an ongoing arms race between host and microbes, coevolutionary dynamics likely play a pivotal role in shaping the adaptive strategies of bacterial flagellins and their corresponding receptors in plants and animals, mediating the intricate host-pathogen interactions [113].

Pathogens exert strong selective pressures in host populations. Understanding the mechanisms underlying host-pathogen coevolution has been a major area of research interest in both animal and plant systems [114]. Studies on comparative genomics have shown that immunity-related genes and

virulence factors are among the most polymorphic genes in plant hosts and their pathogens, respectively [114,115]. As PRRs are directly positioned at the host-environment interface and are potentially under coevolutionary dynamics with their pathogenic and commensal counterparts, they provide a suitable model for studying natural selection caused bidirectionally between hosts and their associated bacteria [113,116].

Flagellar motility represents potential risks for the host, which has resulted in the independent evolution of flagellin perception systems in plants and animals [7,51]. Over an evolutionary timescale, this resulted in numerous and often redundant mechanisms that collectively work to inhibit flagellar motility and clear potential flagellated pathogens [17]. Interestingly, as observed in diverse systems, many of the host mechanisms to inhibit flagellar motility are non-lethal, likely reflecting the host's need to reduce the collateral damage for beneficial symbionts. Moreover, the host's active "pruning" or limitation of commensal flagella expression also could represent a coevolved strategy where inhabiting commensal bacteria limit their motility to avoid clearance by the host [17].

The initial stage of plant immunity involves the detection of MAMPs through PRRs, a process that potentially influences the assembly of microbial communities within the plant endosphere. Research indicates that both beneficial and pathogenic microbes can bypass PTI by either modifying MAMPs or secreting effector molecules that enhance virulence [93]. The genetic variation in the flg22 epitope, which triggers varying PTI responses in hosts, appears to be a common strategy among species of plant-associated bacteria [73,102,104,117,118]. In contrast, variation in the flg22 perception and differential responses to the same peptide has been observed across *A. thaliana* accessions [119]. Further studies on plant innate immunity have hypothesized that hosts may have evolved specificity to different microbial elicitors, including flg22 variants, and they in turn are constantly changing in microbial populations as a response to host recognition [120].

In *Pseudomonas syringae* pv. *tomato* (Pto), there is evidence that selection favors variants of the less immunogenic flgII-28 alleles [121], indicating a role of flagellin-triggered defenses in pathogen selection. Furthermore, investigations into the *A. thaliana* microbiota reveal a low occurrence of the flg22 peptide epitopes [93]. This suggests that during the establishment of the microbiota, members are selected either for their reduced ability to trigger corresponding PRRs or for evolving different MAMPs recognized by other PRRs [93].

Studies on the selection landscape of Toll-like receptors across vertebrates show a higher proportion of sites under purifying selection. This signal of selection is significantly higher in the extracellular domain (ECD), which is the region directly interacting with the microbial ligand [113]. This reflects functional limitations inherent to vertebrate TLR genes, a predictable outcome considering their role in structurally recognizing conserved microbial PAMPs [122]. Liu and colleagues screened the modes and strength of natural selection of Toll-like receptors across the Vertebrata clade, finding a range of dN/dS ratios comparable with those of rapidly evolving genes [113]. Among all vertebrate TLRs, TLR5 exhibits a high accumulation of positively selected codons and is situated among the TLRs with stronger selective pressures [113].

In various taxonomic groups of vertebrates, including birds and mammals, there have been multiple instances of recurrent episodic positive selection in TLRs [123–127]. Evidence from a wide screen of mammal TLRs shows positively-selected codons in residues 207 and 400 of the TLR5 [125], these are located within the 228 amino acid region of the ectodomain previously identified to be involved in flagellin recognition [128]. Likewise, evidence from primates consistently shows positively selected codons within the same region of TLR5 [124]. Contrastingly, Sharma and colleagues provided evidence of a convergent loss of TLR5 in four distinct mammalian lineages comprising guinea pigs, Yangtze river dolphins, pinnipeds, and pangolins [129]. Similarly, a stop codon ($TLR5^{329STOP}$) producing an inactive form of TLR5 is observed across human populations, resulting in higher susceptibility to Legionnaire's disease [91]. This highlights the dynamic evolutionary patterns in the immune system components of different vertebrate groups, showing both selective pressures in the adaptation of TLR and parallel evolutionary trends in their loss [129], highlighting their impact on pathogen recognition and ultimately, microbiota assembly.

12. Microbial functional diversity through culture-independent methods

The great majority of microbes in natural environments remain unculturable [130], posing a challenge to our understanding of their genetic diversity and ecological roles across diverse ecosystems [130]. Traditional culture-dependent methods impose several limitations for capturing the full spectrum of microbial diversity and functions [131]. However, the development of high-throughput sequencing and metagenomics methods has revolutionized the characterization and understanding of microbial

functional diversity, driving significant discoveries in microbial ecology and biotechnology [130,132,133]. Instead of relying on cultivating individual microorganisms, metagenomics allows for the direct extraction and sequencing of DNA from environmental samples, providing a comprehensive snapshot of the genetic material within a community [132,133]

Metagenomics has leveraged advances in sequencing technologies [134], algorithms for the analysis of biological sequences [132,135], and the development of statistical frameworks [136–139], including machine learning algorithms for the analysis of massive data [140–142]. They represent a great breakthrough in microbial ecology, allowing us to understand microbial communities at the genetic and metabolic level and boosting our capabilities to examine unexplored environments [132]. In addition to the advances in the generation of massive data, the advances in computing infrastructure, storage capabilities [143], and compression algorithms [144] have allowed researchers to produce and store data, unlocking the potential of collaborative research and empowering researchers to delve into vast amounts of data to answer fundamental questions in microbial ecology and evolution, increasing our capabilities to characterize and investigate yet unknown microbial habitats and the biological processes within.

12.1. Gene-centric mining of metagenomes for targeted functional profiling

A multitude of methods exist for the analysis of metagenomic shotgun data. These can be split into assembly-based methods and mapping-based methods [132,133,135]. Assembly-based methods aim to reconstruct entire genomes from metagenomic data. These methods are a great proxy for retrieving catalogs of microbial genomes within a given sample, despite requiring significant computational power and high sequencing genome coverage (i.e. more than 20x) to recover good-quality Metagenome-Assembled Genomes (MAGs) [135]. On the other hand, read-based approaches offer a cost-effective alternative. These methods rely on aligning metagenomic reads to reference genomes, taxonomic marker genes, databases of annotated genes, proteins, or pathways [132,135]. Both methods have their strengths and limitations.

Read-based profiling of metagenomics data is a suitable approach to investigating biological functions in communities of arbitrary complexity given sufficient sequencing depth and satisfactory reference database coverage. In addition, this strategy requires less computational power, enabling an efficient performance for large meta-analyses [135]. For this approach, translated sequence searches against

functionally characterized protein families can be applied to large metagenomic datasets [135], as implemented in the HUMAnN pipeline [145]. Databases can include combinations of putatively predicted protein families such as KEGG [146] or UniProt [147] along with manually curated databases, allowing the characterization of the functional potential of a microbial community. Additional in-depth characterizations of functions of interest have been applied to analyzing virulence factor genes and antibiotic-resistance genes, using well-curated databases of antibiotic-resistance genes and virulence factors, such as the virulence factor database (VFDB) or the antibiotic resistance genes database (ARDB) [148–150].

12.2. Public Metagenomic Data and Open-Source Science

The availability of large-scale metagenomic datasets through public databases and repositories has democratized access to valuable data and metadata, fostering inclusivity and significantly impacting the pace and scope of scientific advancements. By democratizing data availability and access, scientists worldwide can contribute and leverage a collective pool of information, encouraging innovative approaches by integrating diverse perspectives. Platforms like the Sequence Read Archive (SRA) and the European National Archive (ENA) host an extensive collection of metagenomics datasets contributed by researchers worldwide. This abundance of data not only accelerates scientific discoveries but also encourages interdisciplinary and collaborative research. Community-driven initiatives, such as the development of platforms for archiving and implementing bioinformatics algorithms and pipelines, contribute to the standardization and reproducibility of analyses [151–153]. Open-source science promotes the free exchange of methodologies, tools, and findings, enabling a collective effort to tackle complex scientific questions and fostering innovation: one example of this is MGnify, a platform hosting hundreds of thousands of metagenomic samples together with tools to facilitate the assembly and downstream analysis of microbiome-derived sequences [154].

However, fully leveraging the advantages of open-source science presents significant challenges to the scientific community [155,156]. Firstly, the maintenance of data repositories and computational infrastructure necessary for large-scale data analysis. This aspect is critical, as it ensures the availability and usability of data [155]. Second, the establishment of standardized practices for data deposition, and equally important, their corresponding metadata, which is crucial for correctly interpreting patterns arising from the analysis of their corresponding metagenomic datasets [156]. The standardization of these practices is not only a logistical necessity but also a requirement to ensure the integrity and reproducibility of research are a great need in the path to achieving open-source science [157]. Third,

adherence to Open Data policies and compliance with the ethical standards of data sharing [156,158]. Eckert and colleagues found that around 20% of the metagenomics studies published between 2016 and 2019 did not deposit their datasets into public repositories, or were deposited but not accessible [158]: public data are integral to the validation of new computational tools, data mining and discoveries in the "omics" field [159,160]. Consequently, the research conducted in this thesis greatly benefited from the use of public data available in established repositories.

Chapter 1: Flagellin diversity in free-living and host-associated environments.

This chapter will present the results from analyzing public shotgun metagenomes for samples derived from free-living and host-associated environments, to identify patterns of flagellin composition differentiating these bacterial habitats. Firstly, I aimed to describe the taxonomic diversity of flagellins deposited in public repositories and their phylogenetic relationships. Secondly, I characterized the flagellin communities across metagenomic datasets from diverse bacterial habitats and identified the flagellin types that were enriched in each biome. Finally, I investigated patterns of natural selection occurring in these enriched flagellins and the relationship between these patterns and host environment.

The results of this chapter are being prepared for publication.

Introduction

Flagellar motility, a trait that enhances microbial mobility and adaptability, is widely distributed across diverse phyla. A key structural component of flagellar systems is the flagellin monomer. These monomers are not only crucial for bacterial motility but also play crucial roles in host-microbe interactions, including adhesion and the activation of innate immune responses in both plants and animals [17,51]. The molecular capability for flagellar-based motility is widely distributed across diverse bacterial taxa, underscoring its evolutionary robustness and adaptability, hence highlighting its significance for microbial ecology and evolution [2,54]. At the same time, convergent mechanisms for flagellin recognition have evolved in plants and animals [51], representing an evolutionary arms-race between flagellin and their receptors, and highlighting the importance of understanding this evolution in an ecological context.

In the fields of metagenomics and microbial ecology, the study of flagellins within the microbiome represents a largely unexplored field. Despite their importance, there exists a substantial knowledge gap in understanding flagellin diversity and their evolutionary dynamics between communities originating from different environments. In response, the increasing availability of high throughput DNA sequencing data for diverse microbiomes represents a substantial opportunity to investigate within- and between-community protein diversity and function on an unprecedented scale [161].

Despite these opportunities, the analysis of functions within microbial communities from metagenomic data represents a computationally challenging problem. Typically, the identification and profiling of protein families from metagenomic data is achieved via one of three methods: mapping DNA reads to databases of nucleotide sequences [162,163], translating DNA reads and mapping them to databases of protein sequences [163], or assembling full-length genes from DNA reads using *de novo* assembly techniques before annotating these genes against reference databases [164,165]. These mapping-based methods are susceptible to false positives due to small read lengths that can spuriously map to unrelated genes as a result of local sequence homology. On the other hand, *de novo* assembly approaches require substantial sequencing depths to accurately detect low-abundance genes [135,163] (the *de novo* assembly is also challenged by regions of local homology, likely leading to the assembly of chimeric genes [135,163,166]). In terms of computational cost, all three approaches require large amounts of computational resources and processing time [161]. Therefore, strategies that achieve good sensitivity, specificity, with efficient processing times are necessary to properly characterize under-studied gene families within microbial communities [161].

Understanding the mechanisms underlying microbial adaptation to diverse niches is of fundamental importance from both ecological and evolutionary perspectives [167]. The establishment of an evolutionary framework for the study of flagellin diversity not only enhances our understanding of microbial adaptability but also advances our knowledge of functional evolution at the molecular level. While the diversity of the flagellin domains has been studied for many species of Proteobacteria, particularly species of the Enterobacteriaceae family, and Firmicutes bacteria [48,75,168–170], a broader view of their evolution and the mechanism mediating their adaptation to various environments is still missing.

This research aims to bridge this knowledge gap by comprehensively interrogating the diversity of flagellin genes within and between different microbiomes. By leveraging the power of metagenomics data, I provided an in-depth analysis of flagellin gene sequences derived from diverse environments. The use of metagenomic datasets allows for an extensive exploration of flagellin diversity at a broad scale. I expanded the characterization of flagellin communities across bacterial habitats to the study of natural selection influencing the evolution of biome-enriched and biome-specific flagellins.

This work represents a pioneering effort to elucidate the diverse landscape of flagellin communities within diverse microbiomes. By integrating metagenomic data analysis within an evolutionary framework, this work can shed light on the complex interplay between bacterial flagellins and their environments. This research not only fills a critical gap in our understanding of microbial ecology but also creates a framework to easily profile flagellin communities from shotgun sequencing metagenomics data.

Results

1. Diversity and characterization of public flagellin sequences

The dereplication of the Dalong & Reeves dataset at 100% identifity left 33,051 unique sequences. From these, only 24,915 had a corresponding genome and taxonomic annotation with GTDB. The processing of the database with ShortBRED to obtain unique peptide markers identified a total of 33,010 peptide markers corresponding to a total of 9,963 flagellin clusters. The representative sequence from each cluster was taxonomically annotated with GTDB v202 and used for the downstream analyses

While the flagellin sequence database exhibits a vast diversity of flagellins, flagellated bacteria comprise a relatively small fraction of the entire Genome Taxonomy Database (GTDB). Using the GTDBv202, I analyzed the prevalence of flagellated taxa within the domain Bacteria and found that less than 10% of the bacterial species cataloged in this database encoded at least one flagellin **(Figure 1A)**. These species accounted for 37.8% of the bacterial phyla represented in the GTDB **(Figure 1A-B)**. These results underscore the selective distribution of flagellated bacteria and provide insights into their taxonomic spread, highlighting that while flagellation is a notable feature, it is not ubiquitously present across all bacterial lineages.

I then explored the distribution of flagellins across bacterial taxa, as represented in the flagellin sequence database. The analyses showed an extensive taxonomic diversity of flagellins represented by 24,915 unique flagellin sequences spanning 40 bacterial phyla. The median number of flagellins per phylum was 14.5 (IQR: 72.25), among which Proteobacteria and Firmicutes encoded the highest number of flagellins: 15,510 and 3,218 sequences, respectively. I found that 21 out of 423 families represented in the database contained 74% of the total flagellin sequences **(Figure 1C)**. These families were distributed between the phyla Proteobacteria, Firmicutes, and Campylobacterota, with 13 families belonging to the phylum Proteobacteria. In particular, the family Enterobacteriaceae was the most prevalent flagellated bacterial family, having the highest number of flagellin sequences (n=3,529) which represented 14.16% of the total database. This was followed by the family Vibrionaceae, also from the phylum Proteobacteria, (n=2,098), and the family Campylobacteraceae from the phylum Campylobacterota (n=1,616). Within the phylum Firmicutes_A, the most prevalent families were Lachnospiraceae (n=779) and Clostridiaceae (n=682). Likewise, the most prevalent Firmicutes families were Bacillaceae_G (n=1099) and Paenibacillaceae (n=349). The results indicated a broad taxonomic diversity of flagellin sequences across different bacterial families. Families such as Burkholderiaceae, Rhizobiaceae, and Pseudomonadaceae, among others, contributed to the overall flagellin diversity with counts ranging from 1,100 up to 1,500 sequences per family. Flagellin had a skewed abundance distribution, with a few families like Enterobacteriaceae dominating the flagellin sequence count, whereas other families such as Shewanellaceae were represented with a relatively low count of 191 sequences **(Figure 1C)**.

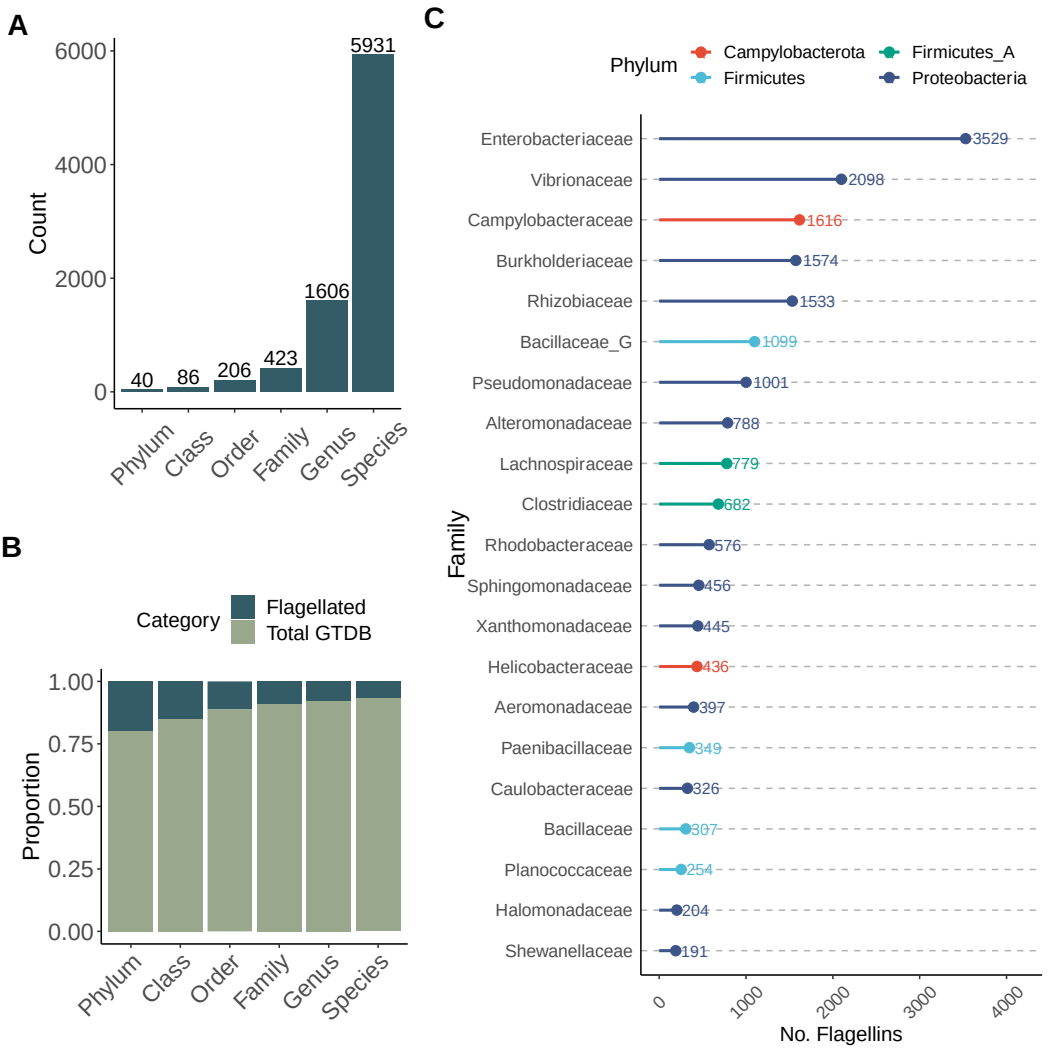

Figure 1. Taxonomic distribution of flagellins from public databases. A) Count of flagellin genes in the database identified at each taxonomic rank from phylum to species. **B)** Proportion of flagellated bacteria compared to the total annotated in the Genome Taxonomy Database (GTDB). **C)** Taxonomic distribution of flagellin in the most abundant families within the database. The panel shows the top 5% of the families with at least one flagellin, highlighting the prevalence of Proteobacteria.

Flagellin sequence length varies across bacterial taxa, including instances of flagellin sequences of over 1000 residues, which were prevalent among several phyla **(Figure 2A, B)**; it is not known if flagellin sequence length is correlated with bacterial genome size. I found a moderate but significant negative correlation between flagellin sequence length and genome size across the bacterial domain (rho: -0.10, p <2.2e-16, Spearman's correlation). However, accounting by phylum showed a few phyla where this correlation was positive: Acidobacteriota (rho: 0.55, p=0.003), Firmicutes_B (rho: 0.4, p=0.0002),

Firmicutes_A (rho: 0.1, p=4.43e-6), and Planctomycetota (rho: 0.49, p=6.25e-6) **(Figure 2B)**. Taken together, these results suggest that, the correlation between genome size and flagellin sequence length is phylum-dependent.

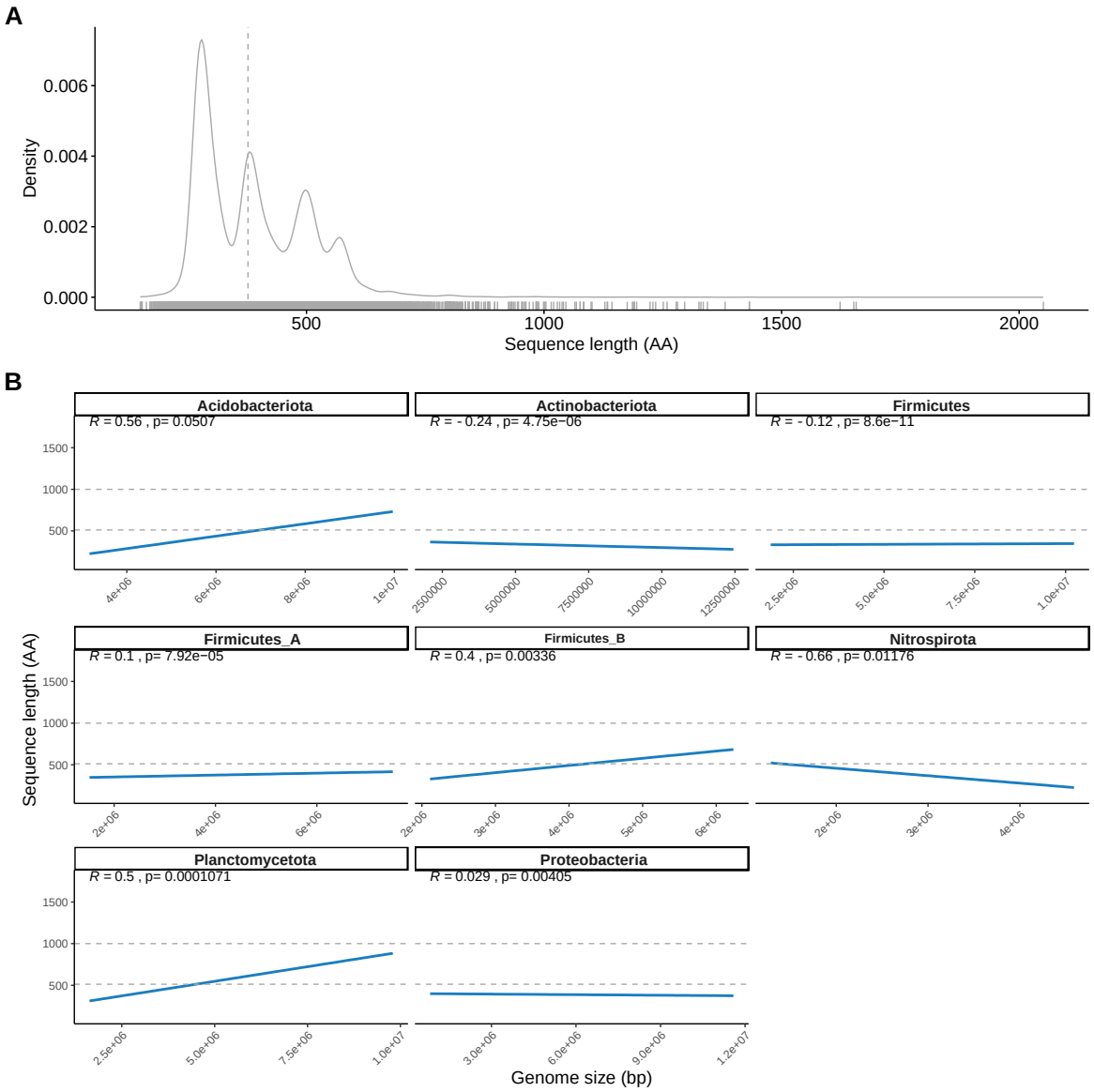

Figure 2. Relationship between flagellin sequence length and genome size across taxa. A) Density plot showing the distribution of the protein sequence length of the flagellins in the database, peaks at different lengths show the variable distribution of protein sizes. B) Phyla for which a correlation between flagellin sequence length and genome size was found to be significant. The effect size and p-value adjusted for multiple testing are shown above each plot. Dashed gray lines show cut-offs at 500 and 1000 amino acids. Sequences highlighted in orange have a protein length higher than 1000 amino

acids

Finally, I investigated the phylogenetic relationships between flagellins in the database. I observed that the phylogenetic relationships between flagellins did not entirely reflect the taxonomy of the bacteria. Although the flagellin phylogeny mostly recapitulated bacterial taxonomy, I observed that Firmicutes flagellins did not form a monophyletic group but were instead spread throughout the phylogeny. I tested whether closely related flagellins displayed similar sequence lengths as a consequence of their phylogenetic proximity. I found that the sequence length of flagellins was correlated with their phylogeny, and found a significant positive correlation (C_{mean}=0.76, p=0.001; Pagel's λ=0.98, p=0.001) **(Figure 3)**.

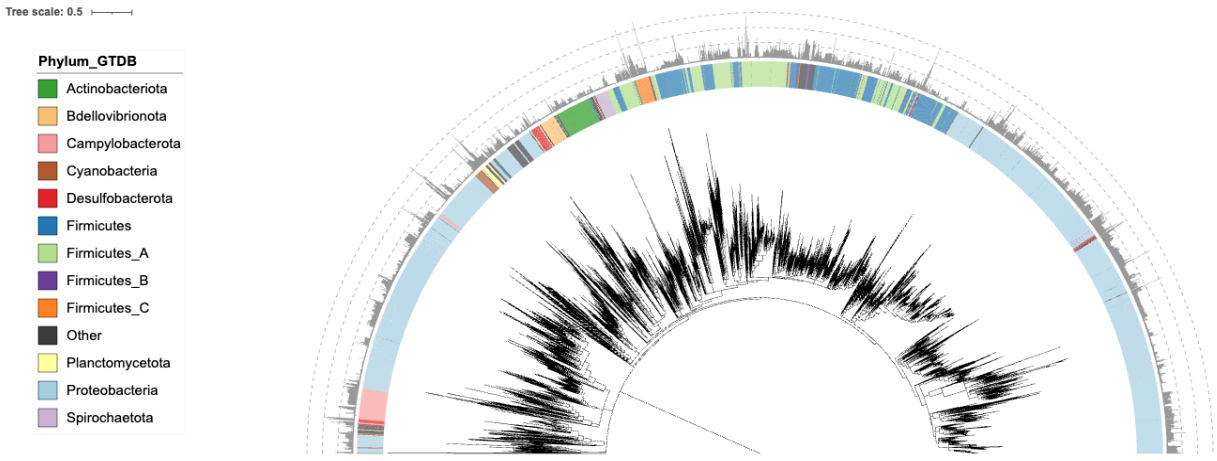

Figure 3. Phylogenetic relationships between flagellins in the database. Phylogenetic reconstruction of flagellin clusters obtained with ShortBRED. The taxonomy was assigned with GTDB for the representative sequence of each family gene. The barplot shows the distribution of sequence length across the tree, dashed gray lines show 500, 1000, and 1500 amino acid sequence length. The tree was generated with FastTree and midpoint rooted in iTOL.

2. Characterization of flagellin communities in free-living and host-associated environments

Our understanding of the ecology and diversification of flagellin across bacteria taxa is limited, in part due to limitations in the use of public sequence data and their corresponding metadata. To overcome

these limitations, I used metagenomic datasets as a proxy to investigate the flagellome, which I defined as the flagellin communities within a sample, using the flagellin database I constructed as the reference to do a read-based profiling of metagenomics data.

The initial set of collected metagenomes contained 785 samples in total. Among these, there were 559 samples from animals, 83 samples from plants, and 133 samples from free-living environments, including terrestrial and aquatic habitats. However, to reduce any potential bias due to sample imbalance between the biomes, I subsampled the initial set of metagenomes by comparing three strategies. The distribution of samples for each subsampling strategy is summarized in **Table 1.** The chosen strategy for downstream analyses (Method 1) contained 178 samples from animals, 75 samples of plants, and 99 samples from free-living environments.

Method	**Actinopterygii**	**Amphibia**	**Aquatic**	**Aves**	**Mammalia**	**Plants**	**Reptilia**	**Terrestrial**	**Total**
OriginalDataset	53	11	102	100	394	87	14	31	792
Method1	53	11	69	53	53	75	14	30	358
Method2	8	6	69	42	88	75	14	30	332
Method3	11	11	11	11	11	10	11	11	87

Table 1. Sample distribution across biome types in the original dataset and the subsampling methods.

Flagellome taxonomic diversity across host-associated and free-living environments.

I characterized the flagellin communities in samples from free-living (FL) and host-associated (HA) environments. To account for sample imbalance between categories, I subsampled the collected metagenomes. The final analysis included 99 samples from FL environments, that include aquatic and terrestrial habitats, 75 samples from plants, and 178 samples from animals. The analysis of alpha diversity showed a higher median diversity in FL environments than in host-associated environments across all metrics **(Figure 4a)**. The observed species and Chao1 indices, which focus on species richness, showed pronounced differences, with FL displaying a significantly higher range of diversity values ($p<0.001$, Kruskal-Wallis rank sum test). The Simpson index, indicative of the probability that two individuals randomly selected from a sample belong to the same species, showed less disparity

between the groups, although the FL samples exhibited higher values than host-associated samples (p=3e-16, Kruskal-Wallis rank sum test).

Within host-associated samples, plants had a significantly higher diversity than animals ($p<0.001$, Wilcoxon-rank sum test), and higher species dominance, as evidenced by significantly higher values in the Simpson index (p=6.6e-13, Wilcoxon-rank sum test). Nevertheless, they were not significantly different from FL samples (p=0.45, Wilcoxon-rank sum test) **(Figure 4b)**. Together, these results show a gradient in flagellin diversity from free-living being the most diverse habitats to animals being the least diverse among the three investigated biomes.

Flagellome compositional differences across host-associated and free-living environments.

I then explored the differences in flagellome composition between host-associated and free-living environments. For this, I calculated several distance metrics based on the number of flagellin reads in metagenome samples. Non-phylogenetic metrics (Jaccard and Bray-Curtis)

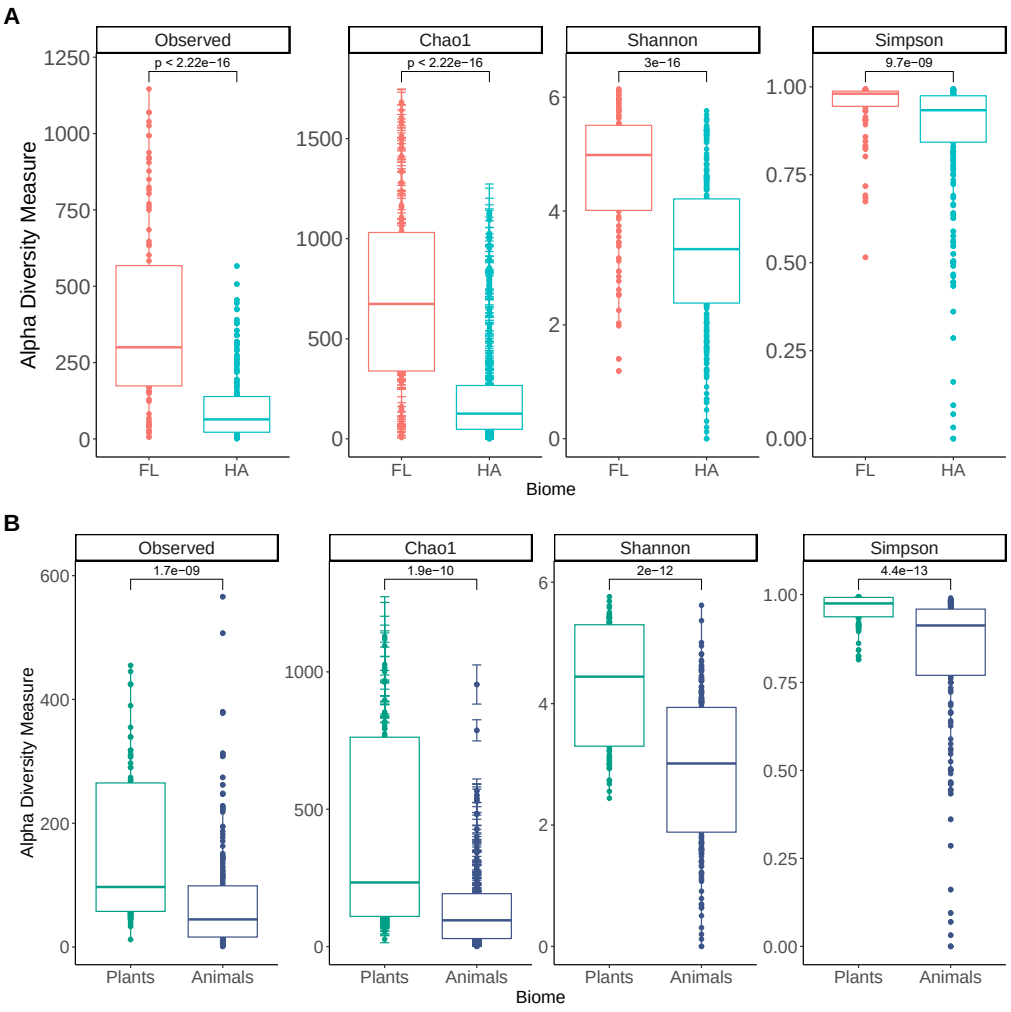

Figure 4. Flagellin alpha diversity measures across different biomes. A) Flagellin alpha diversity comparing Free-Living (FL) and Host-Associated (HA) biomes using: Observed species, Chao1 estimator, Shannon entropy, and Simpson index. B) Flagellin alpha diversity comparing plant and animal hosts. The significance of differences between groups was assessed using a Kruskal-Wallis test.

tended to cluster FL samples separately from host-associated ones, with some overlap between them, indicating some differences in their flagellome composition **(Figure 5A)**. In contrast, phylogenetic-based metrics (weighted and unweighted UniFrac onf flagellin phylogeny) showed higher overlap between the groups **(Figure 5A)**. Nevertheless, the differences in flagellome composition did not allow to differentiate between biome types (R<0.2, ANOSIM).

The phylogenetic metrics of community composition provided a more nuanced understanding of the variances between biomes. In particular, the weighted and unweighted UniFrac metrics captured

33.7% and 19.6% of the variance between biomes along the two first principal coordinates, respectively. In contrast, non-phylogenetic metrics captured only 9.5% and 15.4% of the variance, using binary (Jaccard index) and quantitative (Bray-Curtis dissimilarity) metrics, respectively. This suggests that incorporating both flagellin phylogenetic relationships and flagellin abundances offers a substantial improvement in differentiating biomes based on their flagellin communities. Differences in flagellin composition were greater between biomes than within biomes, and these differences were more evident with phylogenetic measures **(Figure 5B)**. Interestingly, animals showed the highest pairwise dissimilarity, followed by plants and free-living environments **(Figure 5B)**. This indicates less overlap in flagellin community composition and suggests that hosts may play a role in selecting specific flagellin communities.

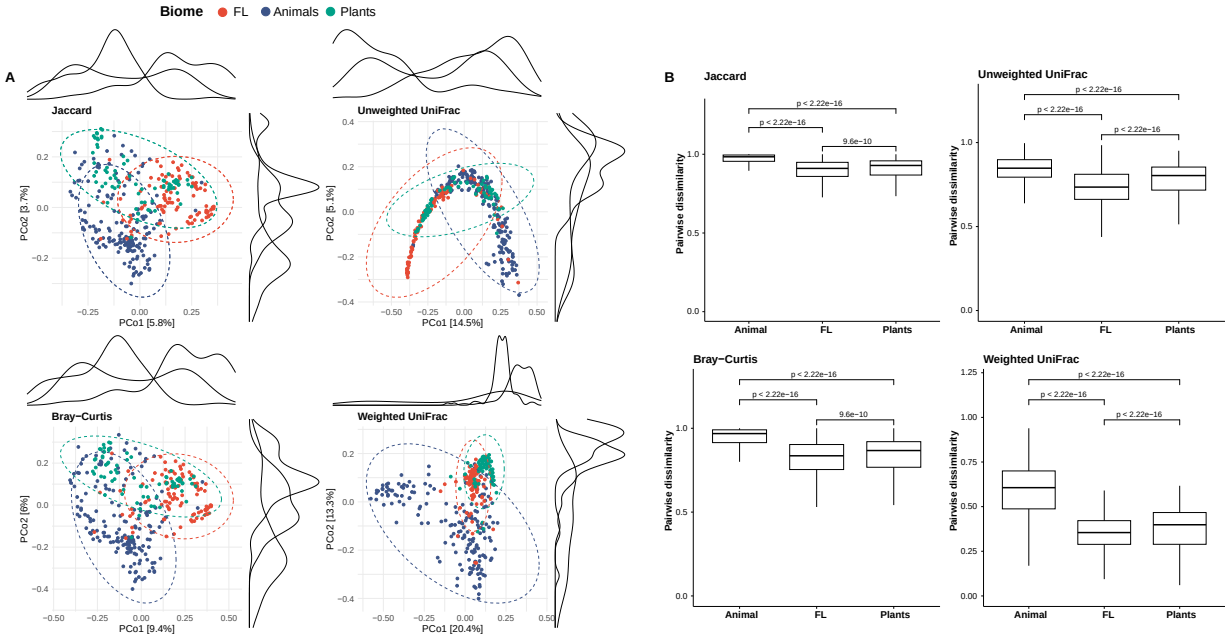

Figure 5. Beta diversity analyses within and between biomes comparing free-living (FL) environments, plants, and animals. A) PCoA plots comparing compositional differences between biomes using non-phylogenetic (Jaccard index, Bray-Curtis index), and phylogenetic (unweighted UniFrac and weighted UniFrac) metrics. Ellipses represent the 95% confidence interval for each biome. Density plots show the distribution of samples in each biome along each axis. The value in the brackets shows the percentage of variation explained by each axis. B) Within-group pairwise dissimilarity based on Jaccard index, Bray-Curtis, unweighted UniFrac, and weighted UniFrac, the significance of differences between groups was tested with a Kruskal-Wallis test.

Integrating alpha diversity metrics into the analysis of flagellin community composition showed a correlation with taxonomic flagellin diversity, as observed with Jaccard, Bray-Curtis, and unweighted UniFrac metrics **(Figure 6)**. Samples with a more similar flagellin composition often shared similar flagellin taxonomic diversity to the species level, indicating that flagellin taxonomic diversity is a significant component of overall community variance. Nevertheless, the strength and nature of the relationship between composition similarities and alpha diversity varied depending on the metric used.

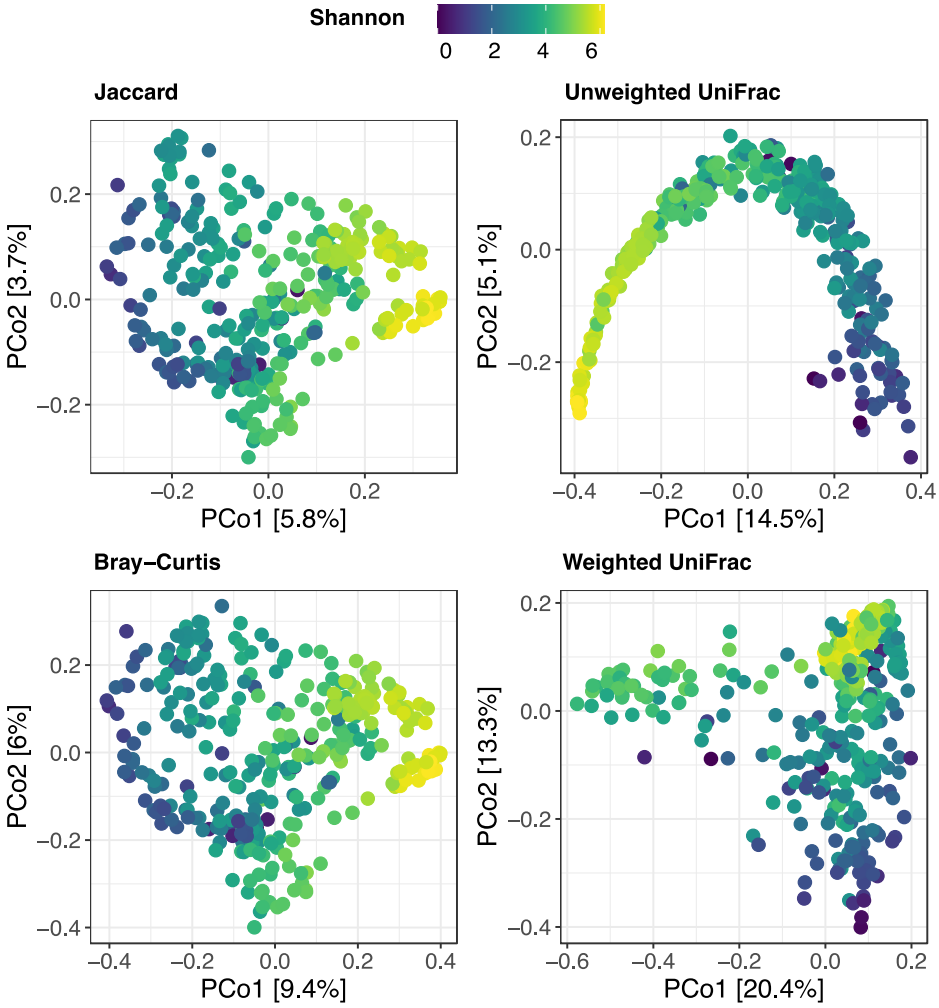

Figure 6. Beta diversity analyses based on species diversity. Principal Component Analysis using four metrics: Jaccard, Bray-Curtis, unweighted, and weighted UniFrac. The PCoA is colored by the Shannon diversity index. The values in squared brackets represent the percentage of variation that is explained by each axis.

Diversity of flagellomes in host-associated environments.

I further explored the diversity of flagellins in HA biomes. The data showed a low degree of differentiation between samples from animals and plants, being weighted UniFrac the metric that better captured the variation in flagellin composition, explaining 38.3% of the variance in the first two principal coordinates **(Figure 7A)**. Flagellin communities associated with mammals were distinct from those associated with Actinopterygii (ray-finned fish), Amphibia, Aves (birds), Reptilia (reptiles), and Plants **(Figure 7B)**. Mammals also had a more homogeneous flagellin community composition relative to the broader spread observed in other taxa **(Figure 7B)**. It becomes evident that animals, particularly mammals, consistently separate from plants and other animal groups, indicating fundamental differences in their flagellin community structure **(Figure 7B)**. In contrast, plants' flagellin community composition was rather heterogeneous among samples.

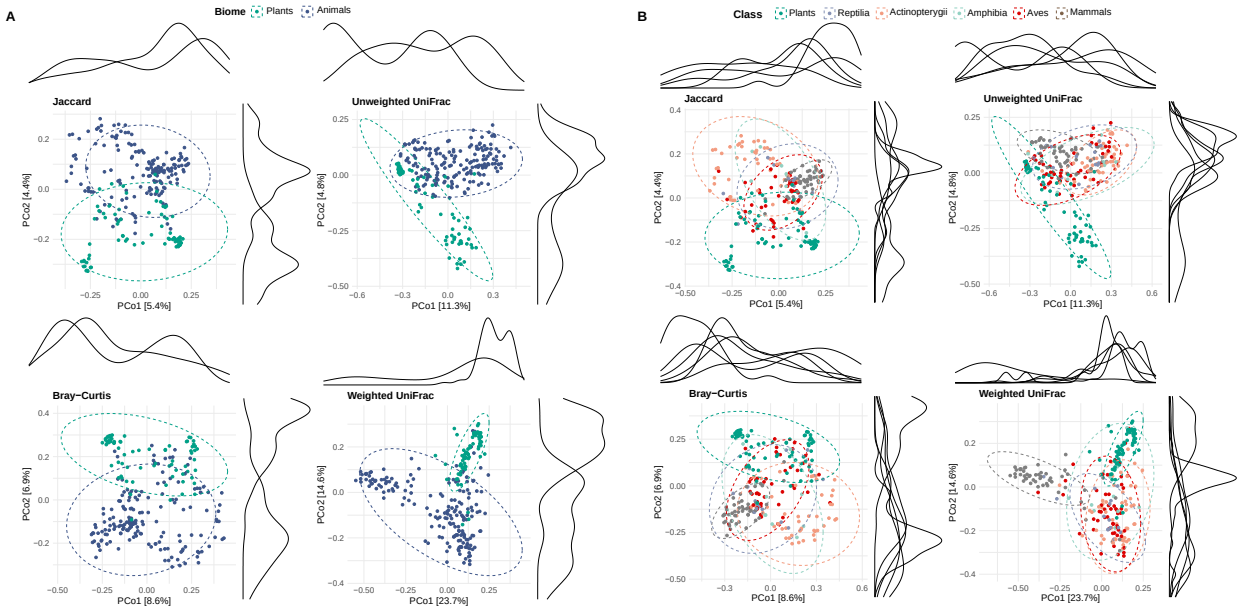

Figure 7. Compositional differences within the host-associated category. A) Principal Component Analysis (PCoA) comparing plants and animals, using non-phylogenetic (Jaccard, Bray-Curtis), and phylogenetic (unweighted, and weighted UniFrac) distances. The density plots show the distribution of samples along each axis. The values inside brackets show the percentage of variation that is explained by each principal component. Ellipses represent the 95% confidence interval for each biome. B) Principal Component Analysis (PCoA) comparing plants and the major clades within animals: fish, reptiles, amphibians, birds, and mammals.

Based on these results, I decided to explore further the differences in animal flagellomes, in particular within Mammals. A total of 178 samples from 119 unique animal taxa were included in the analysis. The flagellome composition of mammals was consistently different from that of the other major groups of vertebrates **(Figure 8A)**. The order Primates was the most prevalent in the animal dataset; within this order, I found that only 19.17% of the total variation in the flagellin composition could be explained by the host family (p=0.001, PERMANOVA). However, based on the ordination analysis, there was a consistent differentiation between old world monkeys (suborder Catarrhini), and new world monkeys (suborder Platyrrhini) in the ordination analysis. The second group, represented only by one family in the dataset (Atelidae) clustered apart from the other suborders **(Figure 8B)**.

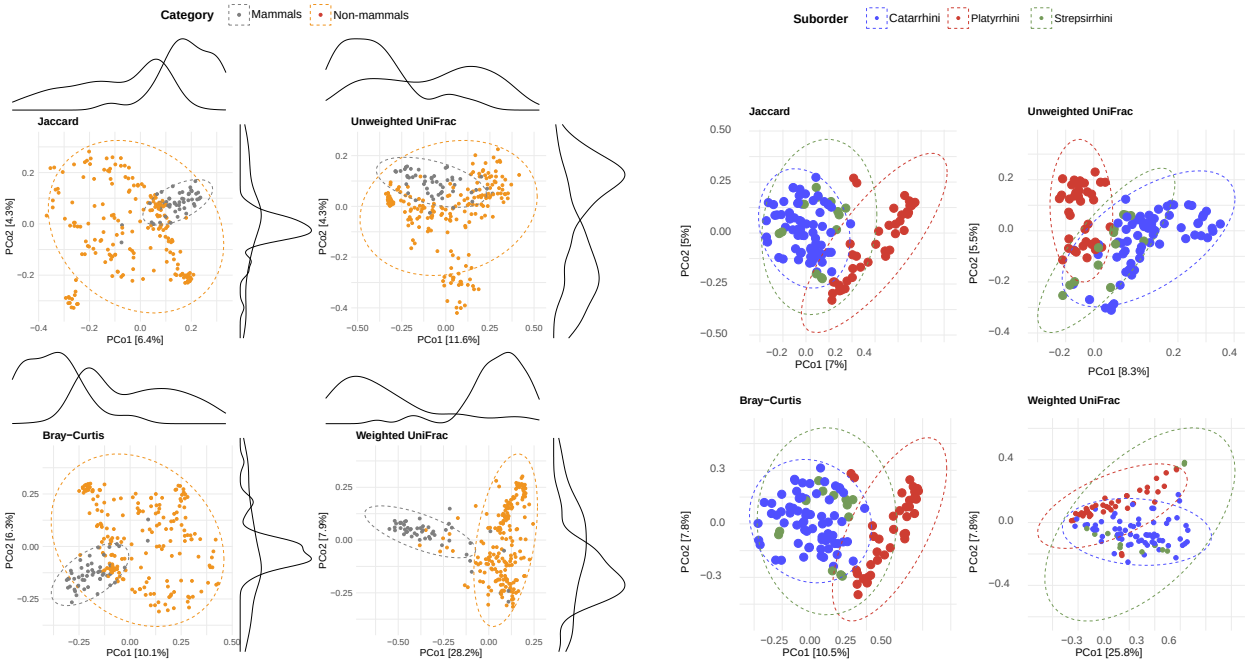

Figure 8. Compositional differences within animals and Primates. A) Principal Component Analysis (PCoA) comparing mammals and non-mammals, using non-phylogenetic (Jaccard, Bray-Curtis), and phylogenetic (unweighted, and weighted UniFrac) distances. The density plots show the distribution of samples along each axis. The values inside brackets show the percentage of variation that is explained by each principal component. Ellipses represent the 95% confidence interval for each biome. B) Principal Component Analysis (PCoA) comparing the suborders of Primates: Catarrhini (old-world Monkeys), Platyrrhini (new-world monkeys), and Strepsirrhini (lemurids and tarsids).

To evaluate the extent to which the overall microbiome composition explained the observed patterns of flagellin community assembly, I profiled the microbiome across all samples using Kraken/Bracken [171,172]. I observed weak to no correlation between the flagellome and microbiome (**Table 2**), supporting that the observations of the flagellome are not confounded by general microbiome compositional patterns.

Dataset	**Mantel's rho**	**p-value**
Youngblut_2021	-0.038	0.866
De la Cuesta-Zuluaga_2020	0.03736	0.077
MGnify_plants	0.5972	0.001
Amato_2019	0.3271	0.001

Table 2. Spearman's correlation comparing the flagellome and microbiome compositional matrix.

3. Enrichment analysis between biomes

Next, I performed pairwise analysis of differential abundance to identify enriched flagellins in either HA or FL samples, and within HA samples, between plants and animals. I found a set of 513 flagellins enriched in free-living samples, belonging to 442 species spanning 24 bacterial phyla **(Figure 9A,C).** Only 110 flagellins were enriched in host-associated environments, comprising 6 bacterial phyla **(Figure 9A)**. The comparison between plants and animals identified 121 and 82 enriched flagellins, respectively (**Figure 9B)**. The enriched set in animals comprised flagellins belonging to 85 different species of 6 phyla, including a great proportion of flagellins from Lachnospiraceae members (phylum Firmicutes_A) and, less commonly, flagellins belonging to the phylum Cyanobacteria. The flagellins enriched in plants corresponded to 71 different species from 5 phyla, belonging mostly to the phylum Proteobacteria, Actinobacteriota, and Desulfobacterota **(Figure 9D).**

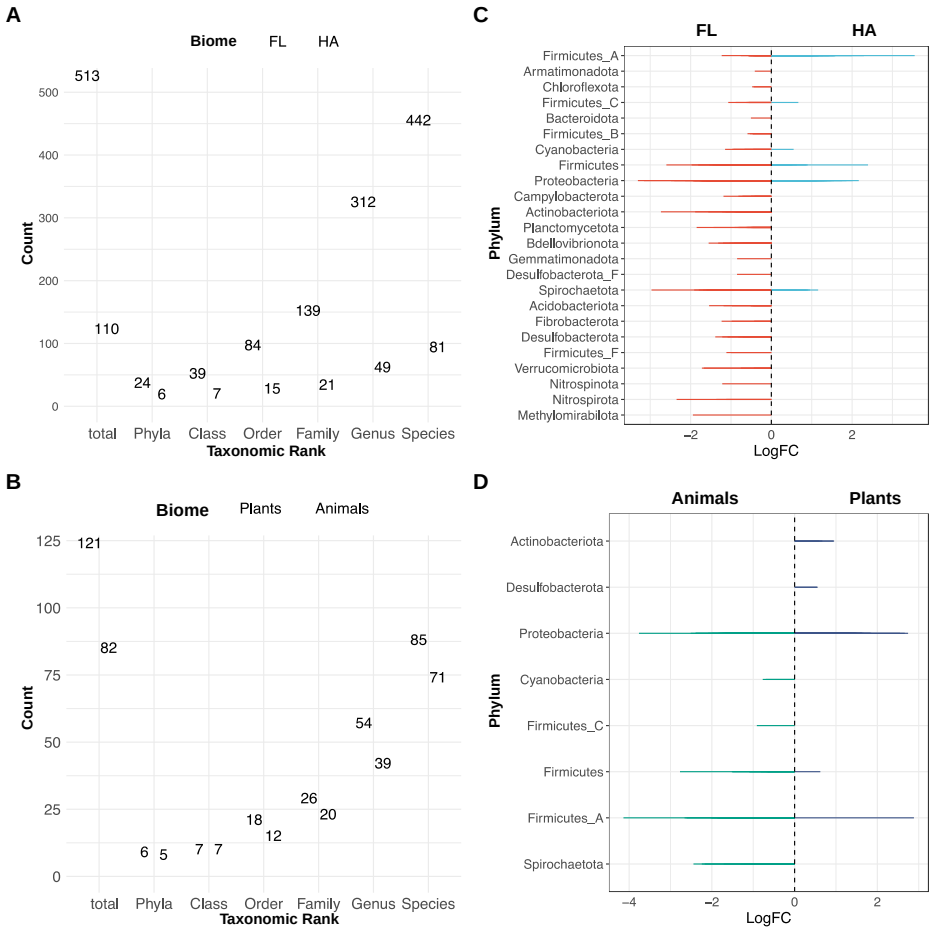

Figure 9. Taxonomic distribution of enriched flagellins across biomes. A-B) Enriched flagellins detected with EdgeR in the pairwise comparison between free-living (FL) and host-associated (HA) environments. C-D) Enriched flagellins detected with EdgeR in the pairwise comparison between plants and animals. A, C) Total number of flagellins detected in each biome summarized by taxonomic rank. B, D) Distribution of enriched flagellins across bacterial phyla. Each dot represents an individual flagellin.

Flagellins from the phyla Apirochaetota, Firmicutes C, and Cyanobacteria are were found enriched only in animals, while flagellins from the phyla Actinobacteriota and Desulfobacterota were found enriched only in plants **(Figure 10A).** A group of flagellins from the phyla Firmicutes, Firmicutes A, and Proteobacteria were found enriched in both plants (n=76 flagellins), and animals (n=106 flagellins) (**Figure 10B)**. Among these, the families Rhizobiaceae, Xanthobacteriaceae, and Caulobacteriaceae (phylum Proteobacteria) accounted for the major proportion of flagellins enriched in plants. In animals, the main families enriched were Enterobacteriaceae and Pseudomonadaceae, from phylum Proteobacteria, and a great proportion of Lachnospiraceae flagellins, from phylum

Firmicutes A, that were overrepresented in animals **(Figure 10B)**. Likewise, 83 and 48 flagellins from 5 families were enriched in both animals and plants, accounting for 53 and 38 species, respectively. From these, only flagellins belonging to Pseudomonas E and Pseudobutyrivibrio were enriched in the two biomes (**Figure 10C**). This indicates that flagellins from these phyla are not confined to a single environment but rather exhibit a dual enrichment, underscoring their versatile presence across both animals and plants.

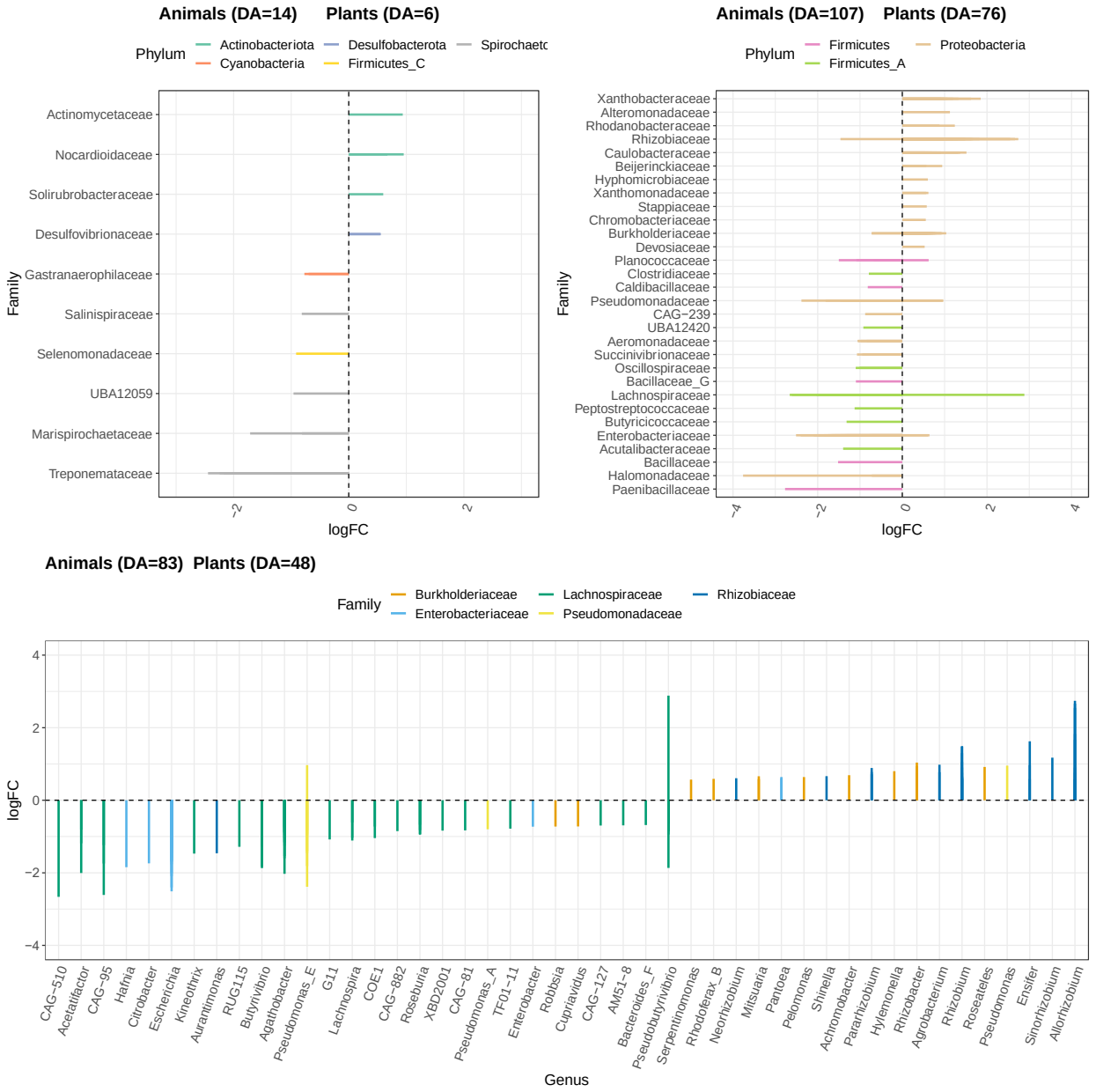

Figure 10. Taxonomic summary of enriched flagellins found between animals and plants. A) Distribution of enriched flagellins across bacterial families from phyla that are not shared between plants and animals, each point represents an individual flagellin. B) Distribution of enriched flagellins

across bacterial families, from phyla shared between plants and animals (Firmicutes, Firmicutes A, and Proteobacteria), showing that most of the flagellins enriched in plants correspond to the Proteobacteria phylum, while flagellins enriched in animals largely span phyla Firmicutes A, and Proteobacteria. C) Distribution of enriched flagellins by genus in the families that were found enriched in both biomes. Values in parenthesis show the number of flagellins enriched in each biome. DA=Differentially Abundant.

Further comparisons between mammals and non-mammals showed a great enrichment of flagellins in mammals. In total, 169 flagellins corresponding to 117 different species were enriched; they represented 9 phyla, among which Firmicutes and Proteobacteria were the most prevalent. On the other hand, non-mammals had 29 enriched flagellins from 20 different species, 93% of them belonging to the Proteobacteria phylum **(Figure 11A)**. The Lachnospiraceae family contained a big proportion of enriched flagellins from multiple genera. Interestingly, none of the Lachnospiraceae members were enriched in non-mammals. A total of 27 genera contain enriched flagellins in mammals, with *Agathobacter, Roseburia, Butyrivibrio, Pseudobutyrivibrio, Lachnospira,* and COE1 being the biggest contributors of enriched flagellins **(Figure 11B)**.

To explore the contribution of mammals to the differentially abundant flagellins observed between animals and other biomes, I performed all-to-all pairwise comparisons within animals. This showed that indeed, mammals exhibit the largest repertoire of flagellins that are differentially abundant compared to the other vertebrate groups: mammals showed the higher number of enriched flagellins compared to Actinopterygii (n=141 flagellins enriched in mammals), followed by Aves (n=132), Reptilia (n=32) and lastly, Amphibia (n=4). Amphibians did not show differentially abundant flagellins compared to the other vertebrate groups, while reptiles showed 24 enriched flagellins compared to Actinopterygii, and 14 compared to Aves. The number of enriched flagellins observed in all pairwise comparisons are summarized in **Table 3.**

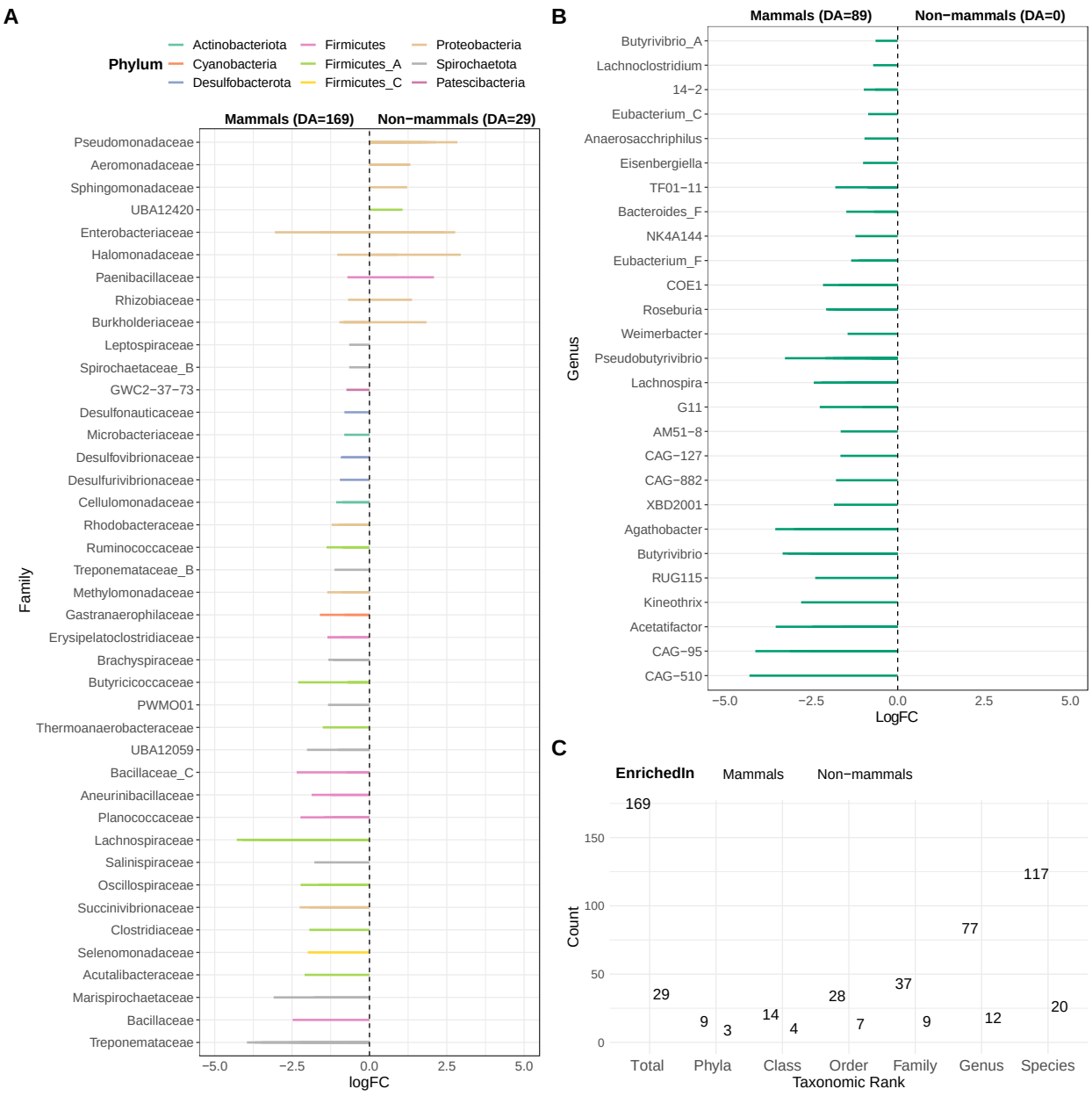

Figure 11. Taxonomic summary of enriched flagellins within animals. **A)** Taxonomic distribution of differentially abundant flagellins between mammals and non-mammals. The majority of mammals-enriched flagellins correspond to the Lachnospiraceae family. B) Subset of Lachnospiraceae-derived flagellins differentially abundant in mammals. No flagellins from the Lachnospiraceae family were found enriched in non-mammals. **C)** Summary of the total number of flagellins detected as enriched in mammals and non-mammals and their corresponding distribution across taxonomic ranks.

Pairwise_comparison	Group	N_flagellins	Phyla	Class	Order	Family	Genus	Species
Aves_Mammals	Aves	46	4	5	10	12	22	30
Aves_Mammals	Mammal	132	7	10	20	28	58	91
Mammals_Fish	Fish	7	2	2	3	3	3	5
Mammals_Fish	Mammal	141	8	12	24	31	63	97
Mammals_Reptiles	Reptiles	27	2	3	5	7	16	23
Mammals_Reptiles	Mammal	32	3	3	5	5	14	25
Mammals_Amphibia	Amphibia	0	0					0
Mammals_Amphibia	Mammal	4	2	2	2	2	2	4
Amphibia_Aves	Amphibia	0	0	0	0	0	0	0
Amphibia_Aves	Aves	0	0	0	0	0	0	0
Amphibia_Fish	Amphia	0	0	0	0	0	0	0
Amphibia_Fish	Fish	0	0	0	0	0	0	0
Amphibia_Reptiles	Amphibia	0	0	0	0	0	0	0
Amphibia_Reptiles	Reptiles	0	0	0	0	0	0	0
Reptiles_Fish	Reptiles	24	4	5	7	8	17	22
Reptiles_Fish	Fish	0	0					0
Ave_Fish	Ave	55	4	5	13	15	28	38
Ave_Fish	Fish	7	2	3	3	3	3	5
Ave_Reptiles	Ave	3	1	1	2	2	2	3
Ave_Reptiles	Reptiles	14	2	2	4	6	9	11

Table 3. Taxonomic distribution of differentially abundant flagellins in all pairwise comparisons within animals.

Phylogenetic relationships of enriched flagellins

One question emerging from these analyses was whether the phylogenetic relationships between flagellins were explained by the bacterial taxonomy or the habitat. The flagellin phylogeny of enriched flagellins between plants and animals showed a division between flagellins enriched in plant hosts and those associated with animal hosts. Interestingly, a few flagellins enriched in plants were more closely related to animal-enriched flagellins than to the clades of plant-enriched flagellins (**Figure 12**).

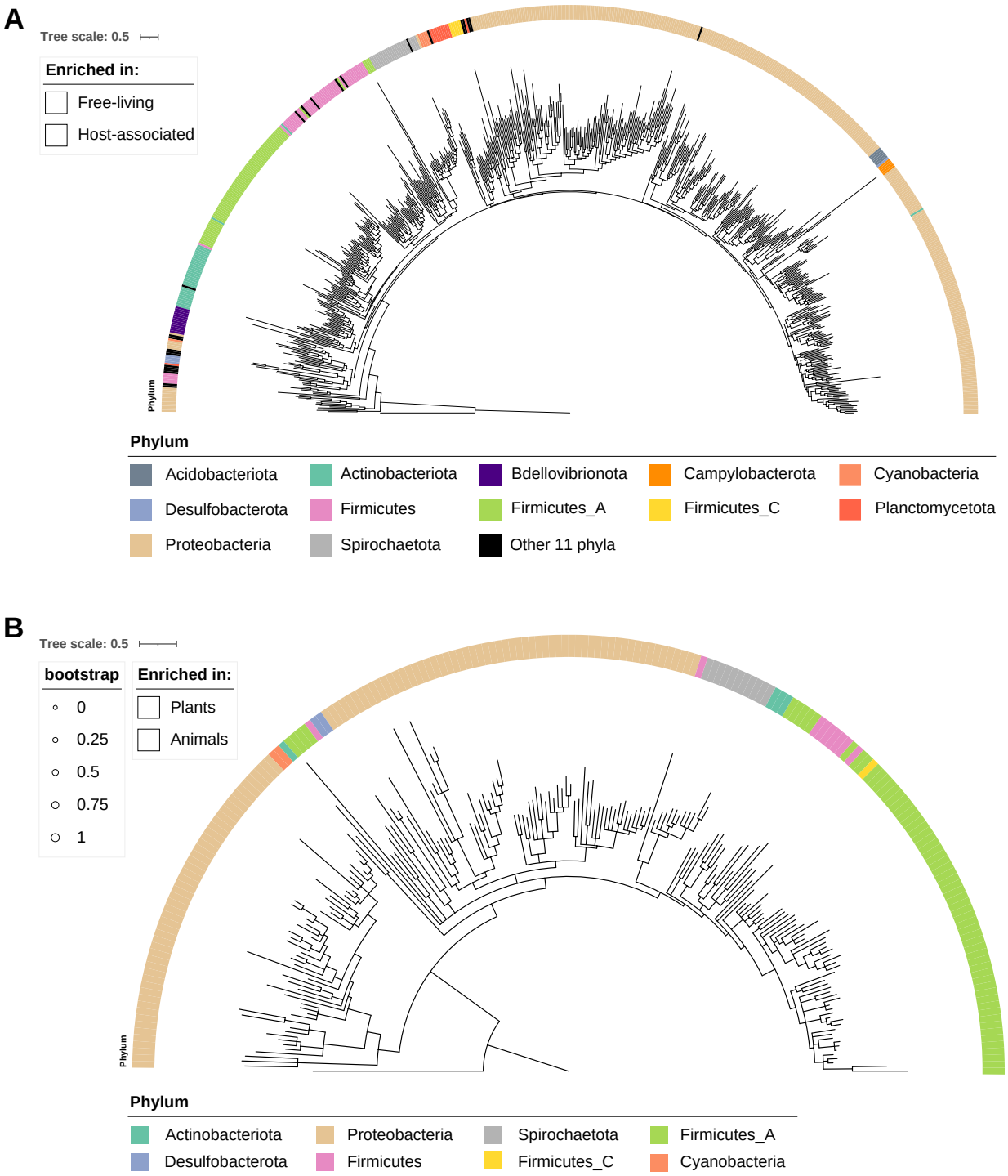

Figure 12. Phylogenetic reconstruction of differentially abundant flagellins detected in pairwise comparisons by EdgeR. A) Phylogenetic reconstruction of flagellins differentially abundant between free-living (FL) and host-associated (HA) environments. Branches with a bootstrap below 0.8 were removed from the tree to improve visualization. The color ranges correspond to the biome where the

flagellins are enriched. The color strip represents the phylum annotation for each individual flagellin. **B)** Phylogenetic reconstruction of differentially abundant flagellins from the pairwise comparison in animals and plants. Visualizations were produced with iTOL.

Biome-specific flagellins

Some flagellins were identified as biome-specific: free-living (n=12), animals (n=18), and plants (n=11). I further characterized these unique sequences in terms of their length, abundance, and taxonomic origin. The sequence length variation among these, and only the flagellins unique to plants showed a narrow range of sequence lengths with a mean sequence length of 314 aa (SD: 40.9), while unique flagellins from animals (mean: 360 aa*, SD: 108 aa) and free-living (mean: 417 aa*, SD: 117.3) exhibited a wide range of sequence lengths **(Figure 13A)**. Nevertheless, no significant differences were observed in the sequence length of unique flagellins among free-living animals and plants (p=0.12, Kruskal-Wallis test). I found that the unique flagellins in animals were dominated by members of the family Lachnospiraceae (phylum Firmicutes_A), while the unique flagellins in plants were predominantly members of the family Rhizobiaceae (phylum Pseudomonadota). The unique families in free-living belonged to numerous families across the Alphaproteobacteria and Gammaproteobacteria classes **(Figure 13B)**.

The phylogenetic analysis showed that biome-specific flagellins from the three major biomes reflected the bacterial taxonomy. Notably, flagellins specific to free-living were more closely related to plant-specific flagellins. Likewise, one flagellin specific to plants was more closely related to free-living-specific flagellins than to other plant-specific flagellins. (**Figure 13B)**. Likewise, two flagellins (phylum Proteobacteria) unique in animals were more closely related to flagellin from other free-living Proteobacteria. A substantial number of flagellins, unique in animals, formed a well-supported clade in the tree.

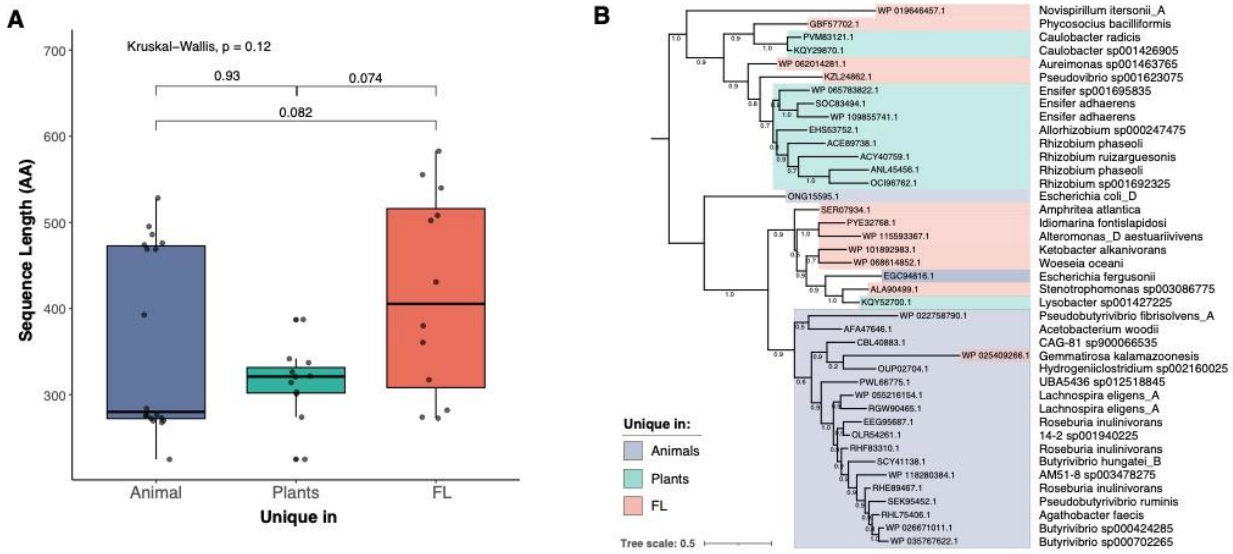

Figure 13. Biome-specific flagellins. A) Distribution of sequence length of biome-specific flagellins found in each biome: FL, animals, and plants. The significance of the difference in sequence length between groups was assessed with a Kruskal-Wallis test. B) Phylogenetic reconstruction of biome-specific flagellins, each color range represents the biome where flagellins are unique.

Phylogenetic structure of enriched flagellins

I performed a comprehensive evaluation of phylogenetic diversity and structure of enriched flagellin communities creating two distinct datasets: a) complete set, and b) biome-specific set. Each dataset offers a unique lens to examine flagellin diversity and further explore natural selection across free-living, plants and animals. The complete set included the entire set of flagellins identified in the analysis of differential abundance. The biome-specific set included the flagellins occurring in only one of each biome.

I evaluated the association between phylogenetic diversity and community structure using the following metrics: Mean Pairwise Distance (MPD), Phylogenetic Diversity (PD), and Mean Nearest Taxon Distance (MNTD)**(Figure 14)**. These metrics were calculated on the flagellin phylogeny for each biome. In the complete set, the MPD, indicative of the overall phylogenetic breadth within a community, was relatively comparable across the three biomes, with a slightly lower value observed in animals. The PD showed a markedly higher phylogenetic richness in free-living environments

compared to those associated with animals and plants. Conversely, the MNTD, a measure of the phylogenetic relatedness at the tips of the tree, showed higher richness for plants, followed by free-living, and lastly by animals **(Figure 14A)**.

In the biome-specific set, the three metrics were consistently lower in animals compared to the other biomes **(Figure 14B)**. This suggests that the animal flagellin communities may be less diverse and more phylogenetically clustered. Free-living showed the highest diversity for all metrics, suggesting higher phylogenetic breadth and richness. Plants showed diversity values similar to free-living.

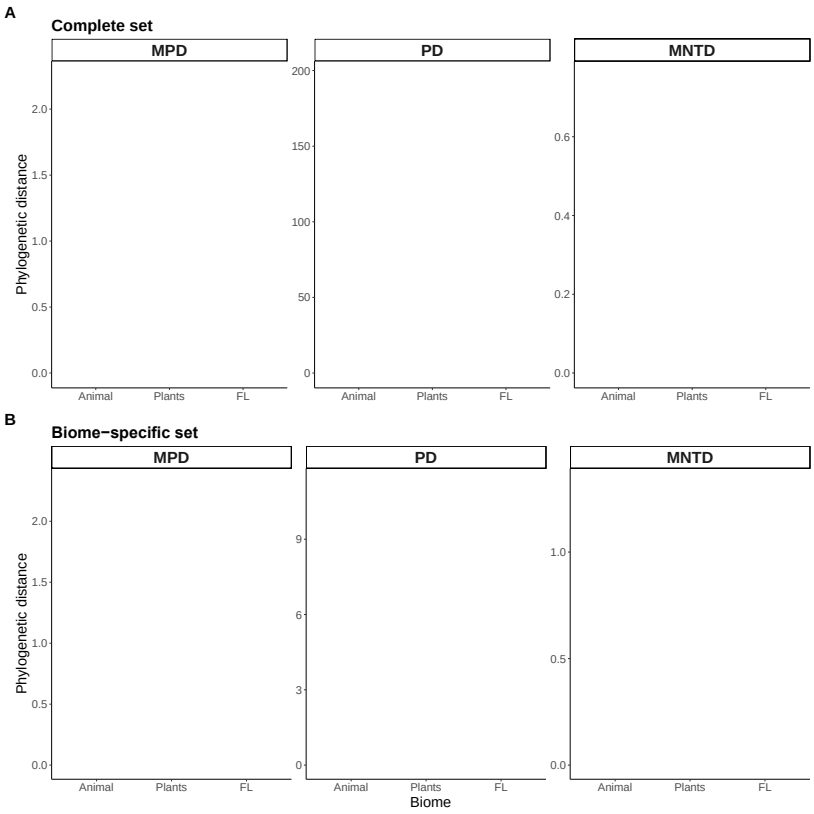

Figure 14. Phylogenetic structure of the community of enriched flagellins found in each biome. A) Phylogenetic diversity in the complete set of enriched flagellins found for each biome. Phylogenetic diversity was measured with three different metrics: Mean Pairwise Distance (MPD), Phylogenetic Diversity (PD), and Mean Nearest Taxon Distance (MNTD). **B)** Phylogenetic diversity in the set of biome-specific flagellins

Across the two datasets, animals consistently showed a pattern of lower phylogenetic diversity and phylogenetic breadth, suggesting a broader and dispersed phylogenetic structure of flagellin communities in the free-living and plant biomes**.**

Random community assembly of enriched flagellins

One way to test the phylogenetic association of enriched flagellins in a specific biome is to test the hypothesis of whether the structure of a community follows a random assembly. For this, I used the standardized effect size (SES) of the above-mentioned phylogenetic measurements: MPD, MNTD, and PD with 1000 bootstrap replicates. The analysis showed that the flagellin community represented in each set of enriched flagellins followed a stochastic assembly process, except for animals, which showed a phylogenetic association in their flagellin communities (**Table 4)**.

Biome	**metric**	**ntaxa**	**obs**	**rand.mean**	**rand.sd**	**obs.rank**	**obs.z**	**obs.p**	**runs**
Animals	MPD	121	1.78	2.19	0.08	1	**-5.3370**	**0.000999**	1000
Plants	MPD	82	2.24	2.19	0.09	695	0.5222	0.694306	1000
FL	MPD	512	2.25	2.19	0.03	987	2.0061	0.986014	1000
Animals	PD	121	37.95	65.98	3.20	1	**-8.7653**	**0.000999**	1000
Plants	PD	82	48.29	49.04	2.73	401	-0.2736	0.400599	1000
FL	PD	512	196.43	192.60	4.77	784	0.8033	0.783217	1000
Animals	MNTD	121	0.38	0.74	0.04	1	**-8.1547**	**0.000999**	1000
Plants	MNTD	82	0.75	0.81	0.06	154	-1.0260	0.153846	1000
FL	MNTD	512	0.50	0.50	0.02	590	0.1763	0.589411	1000

Table 4. Standardized Effect Size (SES) of the phylogenenetic measurements in the group of enriched flagellins. MPD: Mean Pairwise Distance, PD: Phylogenetic Distance, and MNTD: Mean Nearest Taxon Distance.

4. Enriched flagellins as candidates to test natural selection

I further explored the extent of natural selection acting on the flagellin genes. I obtained the corresponding CDS of each enriched flagellin in each biome across the three datasets (complete, and biome-specific). Initially, I tested for recombination, a common source of false positives of selection. I found no evidence of recombination in the evaluated datasets.

Pervasive natural selection in enriched flagellins

To assess pervasive natural selection I used the Fixed-Effects Likelihood Model (FEL) implemented in the Datamonkey server. I used FEL as a site model to investigate selective pressures acting on individual codons across the flagellin sequences enriched in each biome. Across the three datasets, I

found evidence of pervasive negative (purifying) selection widespread across the flagellin gene, as well as sites under neutral evolution **(Figure 15A).**

For the three biomes, I found the majority of codons to be under negative selection ($\omega<1$, pvalue<0.05). Interestingly, negative selection dominated across all codon positions for the three biomes. However, only free-living showed a significant proportion of sites under neutrality, located at the center of the codon alignment (13%). Contrastingly, only 1.2% and 5.7% of sites in flagellins from animals and plants, respectively, showed evidence of neutral evolution (**Figure 15B).**

Episodic natural selection in enriched flagellins

To assess episodic natural selection, I used the Mixed Effects Model of Evolution (MEME) in the Datamonkey server. MEME is a branch-site model that allows the ω parameter to change across branches of the tree, being more sensitive to instances of positive selection that vary temporally throughout the tree [173]. MEME does not directly test for negative selection.

I found evidence of episodic positive selection in the two datasets. The sites with LRT>2 and p-value<0.05 were considered to be under episodic positive selection. The analysis consistently showed positively selected sites (PSS) spanning the D0 and D1 domains in the N- and C-terminal regions of flagellin (**Figure 15C, D)**. The results for the two datasets in each biome consistently exhibited evidence of selection in similar regions, with the complete dataset showing a significantly higher number of sites under selection **(Figure 15C)**.

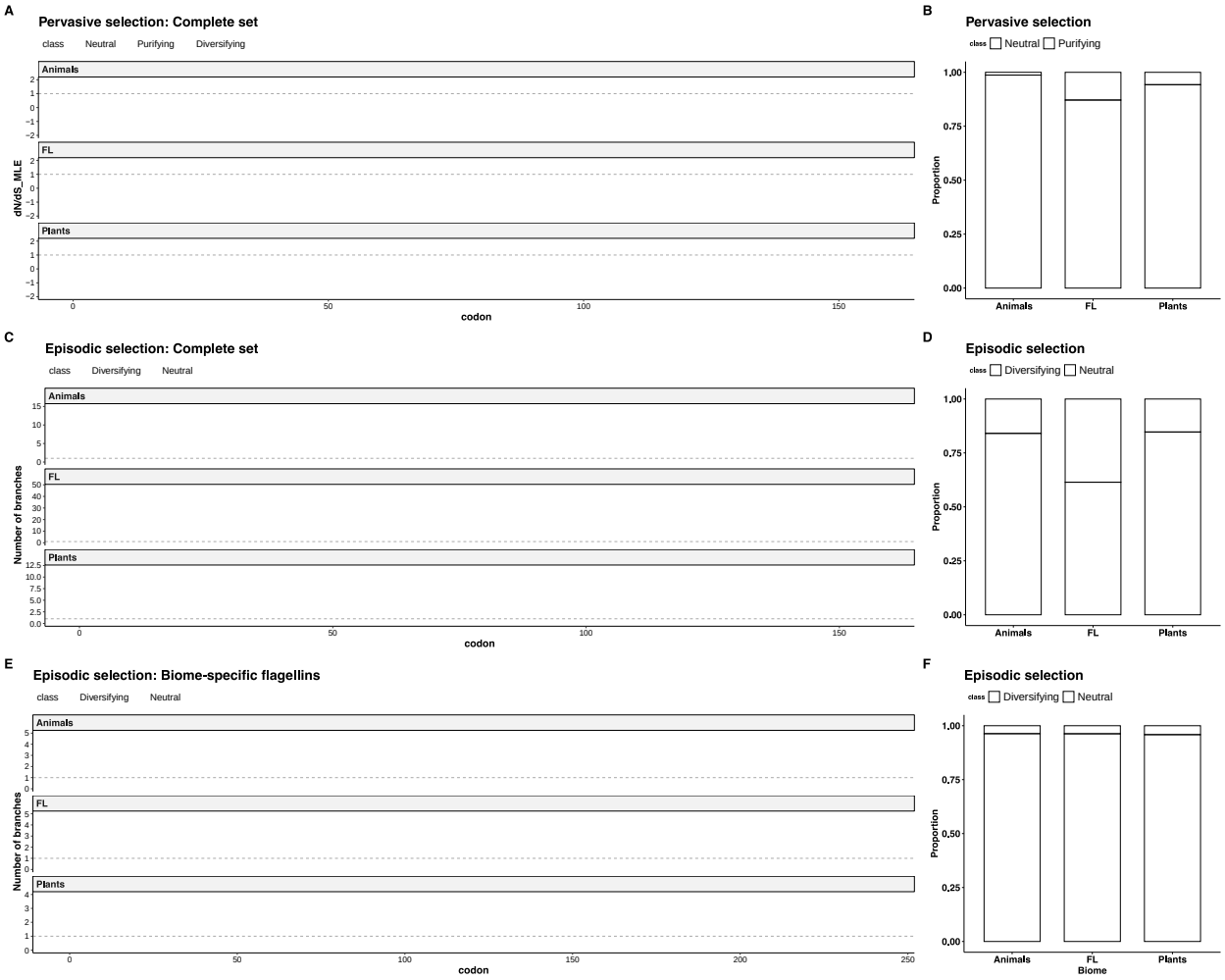

Figure 15. Evidence of pervasive and episodic selection in enriched flagellins. A-B) Maximum-Likelihood estimates of dN/dS for each codon across flagellin sequences. The Fixed-Effects Likelihood (FEL) codon model of evolution was used to assess the effect of pervasive selection in the flagellins from free-living, plants, and animals in the complete set of flagellins. Estimates of $\omega > 1$ and p-value< 0.05 were considered as evidence of negative selection. B) Proportion of sites with evidence of pervasive selection or neutral evolution in the complete set of enriched flagellins. C-F) Number of branches under episodic selection for each codon across flagellin sequences. The Mixed-Effects Model of Evolution (MEME) was used to assess episodic selection. C) Number of branches under positive selection or neutrality for the complete set of enriched flagellins. D) Proportion of sites under episodic selection in each biome. E) Number of branches under positive selection or neutrality for the biome-specific set of flagellins. F) Proportion of sites under episodic selection in each biome.

The complete set of flagellins from free-living samples showed a higher proportion of positively selected sites spanning the conserved domains **(Figure 15C)**. In addition, this was the only dataset

that showed evidence of episodic positive selection in sites belonging to the HVR, although those were not found in the biome-specific dataset (**Figure 15D**). More importantly, the sites under positive selection mapped near the flagellin recognition sites of animals (TLR5 and NAIP5/NLRC4) and plants (FLS2 and FLS3), with a higher proportion of selected sites in the N-terminal region for all biomes and datasets. In particular, plant-associated flagellins showed a higher density of positively selected sites near the flg-22 (FLS2 epitope) **(Figure 15C)**, while animals-associated flagellins presented a higher density of selected sites near the reported NLRC4-interacting region in the N-terminus of the D0 domain **(Figure 15C)**. None of the PSS in animals were localized in the FLS2 epitope in the complete set. Nevertheless, the signatures of positive selection did not reflect the biome where the flagellins are enriched.

Discussion

This chapter focused on examining the diversity of flagellin communities in free-living and host-associated environments. By establishing a detailed framework, I aimed to enhance the study of how flagellin influences host-microbe interactions, providing insights into the ecological and evolutionary roles of flagellated bacteria in diverse habitats.

Flagellated bacteria still are a small portion of known bacterial diversity

Producing a catalog of flagellin diversity lays the groundwork for more refined profiling in metagenomic samples. My findings demonstrate that despite the immense diversity of flagellated bacteria present in public databases, they represent only a fraction of the total bacterial diversity documented to date. Nevertheless, I have developed a method for the sensitive and specific mapping of metagenomic reads to flagellin protein sequences, which may pave the way for more focused investigations into flagellins from particular bacterial taxa or ecological niches. Additionally, the comprehensive database established by this method lays the groundwork for more *de-novo* discovery and characterization of flagellins across bacterial genomes. Methods such as Hidden Markov Models (HMMs) [174] or other classifiers can facilitate these discoveries, continuously enriching this database for a broader representation of bacterial taxa. In summary, the tool I have developed offers a cost-

effective strategy for characterizing flagellin communities, providing an evolutionary framework to explore the influence of natural selection in the diversification mechanisms of these proteins.

Hosts as diversity filters in flagellin communities

Characterizing patterns of composition and diversity of flagellin communities across a wide variety of free-living and host-associated samples supports that hosts act as filters of diversity in the flagellin communities, likely mediated through the flagellin-perception systems present in plants and animals. The higher phylogenetic diversity observed in plant-associated flagellin communities relative to animals could be explained by the ecology and exposed nature of plant-associated surfaces, allowing for the assembly of a broader diversity of bacteria [175]. Contrastingly, the vertebrate gut is characterized by being an isolated niche with very restrictive abiotic conditions such as low oxygen, pH, among others, selecting for very specialized microbial communities and their gene repertoires [175]. In addition to the ecological differences, both plants and animals have evolved specialized flagellin recognition systems that play a role in the assembly and structure of microbial communities [104]. Therefore, the adaptation of flagellins to a specific host ecology might dictate fitness advantages for bacteria thriving in these biomes.

The findings presented here suggest that while the flagellome composition alone may not be sufficient to discriminate between free-living and host-associated environments, a pattern of differentiation emerges when comparing flagellomes from animal and plant hosts, with a further differentiation between mammals and non-mammals. However, I provided evidence of specific flagellins that were enriched across biomes, suggesting a certain degree of host specificity. Remarkably, the pairwise enrichment analyses found flagellins differentially abundant for all pairwise comparisons, except for amphibians, whose flagellin composition lacked distinctiveness from the other animal groups.

One reason behind the lack of distinctiveness between amphibians and the other vertebrate major groups may be attributable to the unique physiological changes they undergo during their lifecycle. Amphibians unlike any other vertebrate experience a vast remodeling of the intestine structure during metamorphosis coupled with changes in feeding behavior [176]. Studies have shown that the microbiota of tadpoles resembles the one of invertebrates and fish, while the adult's microbiota is more similar to amniotes such as mammals, and reptiles, also exhibiting fluctuation of gram-negative bacteria [177,178]. This dual-phase microbial adaptation could account for the absence of differentially abundant flagellin

communities in amphibians, suggesting that the adaptive processes of flagellins are closely related to the host's ecology and life history, needing further research to understand these complex interactions.

Potential for the study of phylosymbiosis in the mammalian gut flagellome

The study of host-microbe interactions, particularly in the context of the mammalian gut microbiome, has been the subject of extensive research [179]. Some studies have focused on the impact of host phylogeny and ecology in microbial assembly [179–181]. Within this context, an interesting finding emerged from the analysis of the flagellomes in Primates, showing a differentiation between new-world monkeys (suborder Plathryrrhini) and old-world monkeys (suborder Catarrhini). Moreover, the order Primates consistently show evidence of positive selection in their TLR5 [124,182], and some studies have shown coevolution with their gut microbiomes [183]. In addition, Muehlbauer and colleagues recently showed that the host gene expression in human colonic cells is species-specific to Hominids microbiomes [184]. My findings extend the observation of host-driven differentiation to the flagellome in Primates and offer an opportunity to investigate phylosymbiotic patterns arising from flagellin composition, as an initial step to test scenarios of host flagellin-TLR5 codiversification [185]. Thus, the study of flagellome-TLR5 codiversification offers a fascinating lens to study the arms race shaping host-microbe interactions.

Further expanding our understanding of mammalian-associated flagellins, the enrichment analysis demonstrated that Lachnospiraceae-derived flagellins are distinctive of mammals compared to the rest of vertebrates. The Lachnospiraceae family is a major representative of the mammalian gut microbiome [186]. This family has members associated with the synthesis of short-chain fatty acids (SCFAs) [187], playing a crucial role in human health [186,188,189]. For these reasons, they have received increased attention in the context of human health. In the particular context of Lachnospiraceae-derived flagellins and their interactions with TLR5, studies have suggested these flagellins have an important role in host-microbiome homeostasis [71]. This evidence supports the relevance of Lachnospiraceae-derived flagellins in the community structure of the mammalian gut microbiome. Flagellins could be important in the maintenance of microbial community structure and symbiosis through the TLR5 interaction.

Flagellin diversity is the result of pervasive negative selection and episodic positive selection

One goal of this chapter was to elucidate the influence of natural selection in the evolution of bacterial flagellins in different ecological contexts, with a particular interest in host-associated flagellins. The finding of widespread pervasive negative selection in the flagellins of free-living, plants and animal biomes underscores the functional constraints in the protein given its role in the maintenance of essential cellular functions [120] such as motility [51]. For this reason, minor mutations might imply structural changes that could abolish motility or impair interactions with their surrounding environment [74]. One example of this is the flagellin FlaA from *H. pylori* and *C. jejuni*, -a well-known TLR5 evader-, that compensates the deleterious effects of mutations in the TLR5 epitope with substitutions across the sequence to maintain motility [74]. Therefore, the maintenance of motility is the result of strong pervasive negative (purifying) selection in the flagellin gene and highlights the adaptive advantage of motile bacteria in any environment.

Nevertheless, although pervasive purifying selection maintains the integrity and functionality of the flagellar filament, I showed that episodic positive selection plays a role at the diversification in the conserved domains of flagellin. This observed pattern might indicate recurrent encounters with flagellin recognition systems, prompting the adaptation of flagellin either for evasion of host defenses or for induce variations in host perception [104,119,120]. Notably, a significant portion of positively selected sites (PSSs) are situated within host epitopes for animals and plants. Although direct evidence for host-driven positive selection was not robustly identified, my findings underscore the significance of episodic positive selection as a mechanism for maintaining a varied repertoire of flagellin forms within microbial communities. This is observed in examples such as the rhizosphere microbiome of *A. thaliana*, where several forms of flg22 are collectively expressed, modulating the Pattern-Triggered Immunity (PTI) [104]

The lack of a clear pattern of host-driven episodic positive selection, is evidenced by a generalized localization of episodic PSSs across the conserved domains in animal and plant-associated flagellins. This pattern suggests that selective pressures on flagellin are not necessarily explained by interactions with the host. Bacteria are very ancient organisms that have evolved millions of years before the evolution of pattern recognition systems in eukaryotes: LUCA and associated prokaryotic cells evolved approximately 3.5 bya and have subsequently diversified into flagellated bacteria and archaea

[40,190–192]. Therefore, the signatures of positive selection might be a reflection of critical residues that maintain the motility and structural stability of the flagellin. Over evolutionary time, these regions might have been coopted by flagellin-recognition systems given their functional constraints.

Limitations and outlook

The lower availability of non-human-associated metagenomes is an important limiting factor in these analyses. Most of the public datasets from non-human organisms and environments correspond to 16S amplicon libraries, thus limiting the functional profiling of such communities. Future explorations in this field can make use of the growing availability of metagenomic datasets and repositories [193–196]. Expanding access to diverse metagenomic datasets holds promise for more comprehensive investigations into the functional flagellin diversity and evolutionary dynamics of flagellin communities across various ecological niches. This will undoubtedly enrich our understanding of host-microbiome interactions and microbial adaptation in diverse environments.

Further analysis focused on known coevolved symbionts of certain clades might shed light on the influence of the host context in the diversity of flagellins. My research focused on the evolution of the diversity of flagellin within microbial communities. However, a further step in looking at the genomic organization of flagellin genes and flagellin operons, as well as the transcriptional landscape of flagellins and interactions between species-specific flagellins will expand our understanding of the evolution of flagellin diversity within the host environment and the biological contexts where these diverse flagellins are expressed. More importantly, to understand the functional implications of substitutions in the residues under positive selection, future directions might focus on the study of the loss-gain function of common substitutions in the sites under positive selection [197].

Recent studies have shown the influence of positive selection in vertebrate TLR5. This pattern is widespread among all major clades of vertebrates. Interestingly, some clades of vertebrates such as river dolphins, pangolins, and guinea pigs have lost their TLR5. Some hypotheses suggest that compensatory recognition mechanisms might allow these species to recognize bacterial flagellin [129]. Unfortunately, our dataset did not include such species. However, it might be relevant to investigate the patterns of flagellin diversity in these species, to understand how compensatory mechanisms of flagellin recognition might shape the evolution of flagellins in such host species.

Methods

1. Flagellin database

I used the flagellin database published by [59]s a reference. First, I performed a dereplication at 100% identity using CD-HIT [198], and annotated the database using InterProScan [199] against the Pfam database [200] . I retained the sequences having both Pfam annotations for the N-terminal domain (PF00669), and the C-terminal domain (PF00700). I removed all sequences named as "FlgL" because this protein do not correspond to the flagellar filament.. I obtained the corresponding taxonomic identifier ('taxid') of each accession in the flagellin database using the R package taxonomizr [201]. Finally, I matched the taxonomic identifier of the flagellins against the bacterial taxonomy of the Genome Taxonomy Database (GTDB) version 202 [202]. I kept only those sequences that passed all the taxonomy controls and had a representative genome in GTDB.

2. Publicly available datasets

For the profiling of flagellin communities, I collected a set of public shotgun metagenomes from different environmental and host-associated sources. The selection criterion was that the datasets were shotgun metagenomes generated with Ilumina. In total, I included 305 samples spanning free-living and animal-associated environments reported in [203]. Additional animal-associated datasets included were: a dataset with 95 samples from wild animals, including Old World monkeys, New World monkeys, Apes, and Lemurids [181], a dataset including 19 samples from Hominids [184], and a dataset of 323 samples spanning the five major groups of vertebrates: Actinopterygii, Reptilia, Amphibia, Aves and Mammals [204]. A dataset of 87 public plant metagenomes was obtained from MGnify [154] and the Sequence Read Archive [205]. The SRA IDs of the samples is summarized in **Table S1.**

Datasets subsampling

There is a biased representation of biomes arising from the overrepresentation of animal samples in the collected metagenomic dataset (i.e. mammals and birds). To mitigate this, I performed a subsampling of the said publicly available datasets, evaluating three distinct strategies to equilibrate the distribution of samples across the biomes. Each strategy was carefully considered to address the

overrepresentation issues and to create a balanced representation of the biomes for subsequent analyses.

The first strategy consisted of a random sampling to obtain the median number of samples per category within animals: Actinopterygii, Reptilia, Amphibia, Aves, and Mammals (median=53 samples/category), plants: rhizosphere, phyllosphere (median=40 samples/biome), and free-living: aquatic, terrestrial (median=30 samples/category). For categories where sample numbers were already below the median, no subsampling was conducted, preserving the original number of samples instead.

The second approach was a more nuanced method within the animal biome, aiming to balance the data by randomly selecting samples up to the median number within each animal family (median=2 samples/family). Simultaneously, for the categories within the plant and free-living biomes, the median number of samples per category was used as the standard for random selection as detailed above.

The third and final strategy adopted a more stringent approach to subsampling, utilizing random selection to normalize all categories within each biome to a uniform baseline: of the minimum global number of samples across all categories(n=11). This stringent rigorous approach guaranteed an equal representation across all categories, thereby eliminating possible overrepresentation biases. However, the reduction of the dataset with this approach also reduces the statistical power of subsequent analyses.

All the diversity analyses were conducted with the three subsampled datasets. Similar patterns of diversity were observed using the three subsampling methods. For this reason, I decided to preserve the first subsampling method for downstream analyses, due to the fair representation of samples while preserving statistical power for the downstream analyses.

3. ShortBRED flagellome profiling

To characterize the flagellin protein communities within the metagenomic samples, I applied the ShortBRED algorithm [161], which reduces the proteins of interest into short and highly representative amino acid sequences known as 'true markers'. The initial step identifies unique peptide markers for

the flagellin sequences in the database I generated. To achieve this, ShortBRED clusters the input sequences into gene families and identifies unique peptide markers of each cluster by finding non-overlapping regions between all consensus sequences and against the Uniref90 database. To streamline this process, the flagellin database was first dereplicated using CD-HIT [198] at a 95% identity threshold to reduce sequence redundancy and increase the efficiency of subsequent steps.

ShortBRED utilizes BLAST for a two-fold purpose: a self-blast to compare consensus sequences within the database, and a reference blast against the Uniref90 database. However, this can be highly time-consuming and computationally intensive given the large volume of reference sequences. To address this challenge, I integrated the DIAMOND algorithm [206] as a substitute of BLAST for the *selfblast* and *refblast* steps. Integrating DIAMOND greatly improved the algorithm's efficiency and reduced the overall runtime. With this optimization, 33,010 unique peptide markers from to 9,963 flagellin gene families were identified. I then used these markers, which were finely tuned to the 9,963 flagellin gene families, to profile their presence and abundance within the metagenomic datasets.

Finally, I used the shortbred_quantify.py script with default parameters to map all metagenomes against identified peptide markers. This mapping works by aligning short metagenomic reads to the said markers, allowing for the quantification of each gene within the metagenome.

4. Flagellin phylogenetic analysis

To examine the phylogenetic relationships between flagellins and their association with the bacterial taxonomy, I reconstructed the phylogeny between the representative sequences of the flagellin gene families obtained with ShortBRED. Initially, I produced a multiple sequence alignment with all the protein sequences using the MAFFT [207] algorithm with parameters '*--ep 0.123 –-auto*', which accurately and efficiently manages large datasets. The alignment was further refined using TrimAI [208] with the parameter '*--gappyout'* to mask poorly aligned regions, thereby enhancing the phylogenetic inference. The resulting alignment was used to construct the phylogenetic tree using FastTree [209] with the Le-Gascuel 2008 (LG) model of amino acid evolution, which infers approximately-maximum-likelihood trees. I selected this method for its ability to handle large alignments rapidly while still producing phylogenies with a high level of accuracy. Finally, I visualized the resulting phylogeny in iTOL [210] and mapped the phyla of origin of each flagellin into the tree. This offers a detailed and up-to-date taxonomic framework, allowing for precise classification of the flagellin sequences in the context of current genomic taxonomy.

5. Flagellin compositional analysis

The flagellin quantification results obtained withshortBRED were processed with the dplyr [211] and phyloseq [212] packages in R to generate compositional matrices, integrating them with the samples metadata and the flagellin phylogeny and taxonomy. I then filtered the data to reduce its sparsity, using a threshold of minimum 5% prevalence OR mean abundance of 10 reads/sample.

I evaluated the alpha diversity of the flagellin taxonomic diversity across biomes using the metrics Chao1, Observed OTUs, Shannon Entropy, and Simpson's Index included in the phyloseq [212] R package. These measures provided a comprehensive view of species richness and evenness within each biome. To compare the diversity across different categories, the Kruskal-Wallis test was applied.

For beta diversity analyses, I incorporated both phylogenetic and non-phylogenetic metrics. This included weighted and unweighted UniFrac, which considers phylogenetic relationships between flagellins, as well as Bray-Curtis and Jaccard distances, which focus on flagellin taxonomic composition. To assess the significance of the differences observed between biomes, two statistical tests were used: Analysis of Similarities (ANOSIM) and Permutational Multivariate Analysis of Variance (PERMANOVA). Additionally, I performed a profiling with Kraken and Bracken [171,172] to obtain the overall microbiome composition, and used a Mantel test to calculate the correlation between the flagellin composition and the microbiome composition.

All the statistical analyses and data processing were performed using dplyr [211], tidyr [213], stringr [214], phyloseq [212], microViz [215], and metagMisc [216] packages in R 4.2.1.

6. Enrichment analysis

To find flagellins that were differentially abundant in each biome, I conducted enrichment analyses using EdgeR [217]. EdgeR effectively handles count-based compositional data and has robust performance even with small sample sizes. It uses an empirical Bayes estimation and exact tests based on the negative binomial distribution. Initially, I performed pairwise comparisons between free-living and host-associated, free-living and plant, and free-living and animal samples. Considering the large sample size within the animal category, I extended the analysis to include all-to-all pairwise comparisons between the five major vertebrate groups: fish, reptiles, amphibians, mammals, and birds. The results were visualized in R 4.2.1 using the ggplot [218] package.

I investigated the role of biome in the phylogenetic relationships between flagellins that were found to be enriched in each of the pairwise comparisons following the phylogenetic analysis approach described above. I obtained the flagellins sequences enriched for all pairwise comparisons and reconstructed their phylogenies. I mapped the specific biome where each flagellin was found to be enriched into the phylogenetic tree.

7. Biome-specific flagellins

I identified flagellins that were uniquely present in one biome but absent in the other two, thus suggesting ecological specialization. Using a presence/absence approach failed to find any candidates. For this reason, I employed a prevalence-based criterion instead. A flagellin was considered uniquely present in animals if it was totally absent in free-living and in a proportion below the 25th percentile in plants. For uniquely present in plants, I selected flagellins totally absent in animals and present in free-living with a proportion below the 10th percentile. For free-living, I selected the flagellins with a proportion in both plants and animals below the 10th percentile. This method ensured that only flagellins with both significant representation in a particular biome and lowly prevalent in the other two were identified as unique to said biome.

Additionally, I compared the distribution of flagellin sequence lengths in the unique flagellins between biomes using a Kruskal-Wallis test. This comparison aimed to identify biome-specific trends in flagellin size that could suggest adaptations to particular environmental conditions.

8. Phylogenetic structure and random community assembly

I evaluated two sets of enriched flagellins. The first group comprised all differentially abundant flagellins based on the pairwise comparisons between biomes (see Enrichment analysis subsection). The second group consisted of biome-specific flagellins identified using a prevalence-based criterion (see Biome-specific flagellins subsection)

To delve into the phylogenetic diversity of these enriched flagellin communities and assess the impact of phylogeny in each community's composition, I employed three metrics: Mean Nearest Taxon Distance (MNTD), Mean Pairwise Distance (MPD), and Phylogenetic Diversity (PD). These metrics together provided an integrated view of the phylogenetic diversity within each flagellin community, providing insights into their phylogenetic breadth. To test if the assembly of these flagellin

communities was stochastic, I computed the Standard Effect Size (SES) for each metric using 1000 bootstrap replicates. This provided a robust statistical framework to discern phylogenetic patterns of community assembly and to infer underlying processes shaping the diversity and distribution of flagellins across the different biomes.

9. Selection analyses on enriched flagellins

I tested the extent of natural selection acting on the flagellins enriched in each biome. I specifically employed the set of flagellins enriched in free-living environments from the pairwise comparisons between free-living and host-associated environments. Similarly, for the animal and plant categories, the selection was based on flagellins that were differentially abundant in the comparison between these two groups.

For each identified flagellin protein, I retrieved the corresponding Coding Sequence (CDS) from the NCBI database using the Batch Entrez tool [219]. Any protein accession that lacked a matching CDS was excluded from further analysis. The alignment of the protein sequences was conducted using CLUSTALW [220] in the GeneiousPrime 2022 suite. Subsequently, I produced the corresponding codon alignments using PAL2NAL [221], applying the codon table 11 for Bacteria and removing all gaps in the alignment.

To assess the potential impact of recombination on the flagellin's evolutionary history, I initially performed a recombination analysis using the GARD analysis [222] on the Datamonkey server [223]. If recombination was detected, I utilized the partitioned data from this analysis in subsequent selection tests to avoid recombination-derived false positives of selection.

I investigated pervasive selection pressures acting on the flagellin sequences through the Fixed Effects Likelihood (FEL) analysis [224] on the Datamonkey server, using 100 bootstrap replicates to ensure robustness. Additionally, episodic positive selection was tested using the Mixed-Effects Model of Evolution (MEME) [173]. Finally, I visualized the positions with evidence of any type of selection or neutrality using FliC of *S.* Typhimurium as a reference sequence, using Geneious Prime 2022. I mapped in this sequence the epitopes and residues that have been reported in the literature to be involved in the interaction with the receptors TLR5, NLRC4, FLS2 and FLS3 to find any pattern of host-driven selection.

Chapter 2: Characterization of flagellin communities and silent flagellins from the human gut microbiome.

This chapter explores the relationship between human gut commensal flagellins and their mode of interaction with TLR5. This required the development of a method to do a homology-based prediction of silent flagellins from publicly available sequences, based on *in-silico* and experimental evidence from pathogenic and symbiont-derived flagellins, with a particular focus on FlaB from *Roseburia hominis*. It also involved an extensive exploration of publicly available flagellin sequences to characterize flagellin communities among samples of gut metagenomes from healthy humans and profile the predicted silent flagellins across these samples. In addition, I investigated further relationships between the evolutionary trajectories of silent flagellins and bacterial taxonomy. My findings shed light on the complex evolutionary history of a novel mechanism of immune regulation in the gut, which likely plays a role in maintaining the balance between host immunity and gut microbiota.

Parts of this chapter were originally published in (Clasen et al., 2023, Appendix 3) [8]. I developed the method for the prediction of silent flagellins and conducted all the flagellome profiling of samples from healthy humans that led to the selection of flagellins for experimental validation. The biochemical experiments to assess the TLR5 binding and signaling activation were performed by Sara Clasen, Michael E. Bell, and Du-Hwa Lee. Data analysis and figures presented in this chapter were done by me alone.

Introduction

The human microbiota is one of the most densely populated ecosystems on earth, with microbial cells vastly outnumbering the number of human cells [225]. About 10:1 microbial cells per human cell are represented in the human gut ecosystem. One of the mechanisms mediating cross-talk between the host and the resident gut microbiota is via the innate immune system [225]. Among the innate immunity, Pattern-recognition receptors (PRRs) were traditionally thought to be activated solely by microbial–associated molecular patterns (MAMPs) from pathogens. However, it is becoming increasingly evident that commensal bacteria also produce such ligands and play a role in the maintenance of intestinal epithelial homeostasis [226–228]. Thus, this interplay may bear an important role in the microbial community structure within host tissues [228].

There are several lines of evidence that highlight the critical role for flagellated bacteria in priming TLR5, which influences the maturation of adaptive immunity [70,229]. Deficiency in Toll-like receptor 5 (TLR5) is linked to conditions such as Crohn's disease (CD) [70,230]. CD patients exhibit increased levels of adaptive immunity to commensal microbes, which were observed through higher levels of flagellin-specific IgG and IgA, as well as circulating flagellin-specific CD^{+4} T cells against Lachnospiraceae-derived flagellins [230]. In contrast, it has been observed that humans with a non-functional $TLR5^{392STOP}$ mutant correlates with CD. Likewise, mice lacking TLR5 develop spontaneous colitis and metabolic abnormalities, while exhibiting shifts in the composition of the microbiota compared to wild-type mice [89,231]. This evidence suggests that feedback loops between innate immunity and microbial ligands are important to maintain the host homeostasis, highlighting the broad relevance of flagellin within the gut environment. Thus, a deeper exploration into the biology of commensal flagellins will provide insights into the intricate TLR5-flagellin interactions in the context of commensal bacteria.

Among human gut commensals, *Roseburia hominis* stands as a highly prevalent member. Using *R. hominis* as a model, studies have shown that commensal-derived flagellins are not only recognized by innate immune receptors but also are important for the modulation of the adaptive immune response [71]. As shown by Patterson and colleagues, the colonization of mice with *R. hominis* has a protective effect on DSS-induced colitis. Moreover, *R. hominis* is found depleted in patients with Crohn's Disease (CD) [71]. This evidence underscores the potential immunomodulatory role of commensal flagellins and highlights our need to understand the host-microbe cross-talk dynamics, which are pivotal for health and disease.

Further insights into flagellins of human gut commensals from the family Lachnospiraceae have highlighted a continuum of pro-inflammatory responses, challenging the notion of TLR5 stimulation as a binary phenomenon. Nonetheless, the molecular underpinnings of this gradient remain elusive [232]. Recent studies have identified critical residues in the CD0 domain, including R467 and the hydrophobic motif VLSLL, which greatly influence TLR5 activation and lose functionality upon mutation [58], Notably, my research group and I introduced a novel concept termed 'silent recognition' by TLR5, characterized by strong receptor binding but weak activation [8].

In this chapter, I aimed to formulate and develop a predictive strategy to identify putative silent flagellins based on sequence information alone. Additionally, I investigated the relationship between the phylogenetic potential of microbiome-associated flagellins and their corresponding taxonomy. Finally, I sought to characterize the flagellomes in both healthy individuals and Inflammatory Bowel Disease (IBD) patients. The findings of this work provided insights into the complex evolutionary history of a novel mechanism of flagellin-TLR5 interaction, which likely plays a role in gut microbial community assembly and host homeostasis. Furthermore, this work offers a novel methodological framework for the analysis of flagellins in the human gut using metagenomic data.

Results

1. Characterizing the human gut flagellome

Several evidence shows increased levels of adaptive immunity against commensal microbes in IBD patients, likely in a TLR5-mediated manner. For this reason, I characterized the flagellin communities of human gut metagenomes from two datasets including samples of healthy individuals (n=27), and irritable bowel disease (IBD) patients (n=105), which I obtained from public repositories [233,234]. I mapped 5131 flagellin sequences across the two datasets. I found the proportion of reads mapped to flagellins to be significantly higher in samples from healthy patients compared to IBD patients (Wilcoxon, p=0.019) **(Figure 16A)**. However, the composition of both flagellomes was similar, both were dominated by Lachnospiraceae-derived flagellins **(Figure 16B-C)**

I then focused on the dataset of [233] because of the availability of matching metagenomes and transcriptomes. In total, 4505 flagellins were identified in the complete dataset. From these, 2579 were

found among metagenomic samples from 27 healthy adults. These flagellins belonged to 583 species spanning 18 phyla. Nevertheless, only five phyla encoded 85% of the total flagellins across all subjects: Desulfobacterota, Cyanobacteria, Firmicutes, Firmicutes A, and Proteobacteria. Flagellins from *Lachnospira* sp. and *Roseburia inulinivorans* were the most abundant among samples of healthy subjects **(Figure 16B)**. In contrast, I found 3911 flagellins among samples from 105 IBD patients. I found that 48% of the flagellins in IBD subjects overlap with the flagellins identified in healthy subjects. The flagellin community among the IBD samples was dominated by members of the Lachnospiraceae family, where the most abundant flagellins belonged to the *Lachnospira*, *Roseburia,* and *Acetatifactor* genera **(Figure 16C)**. Whilst Lachnospiraceae dominated the flagellome composition, I also found flagellins from Cyanobacteria (genus CAG_196) and Proteobacteria (genus CAG_495) to be highly abundant.

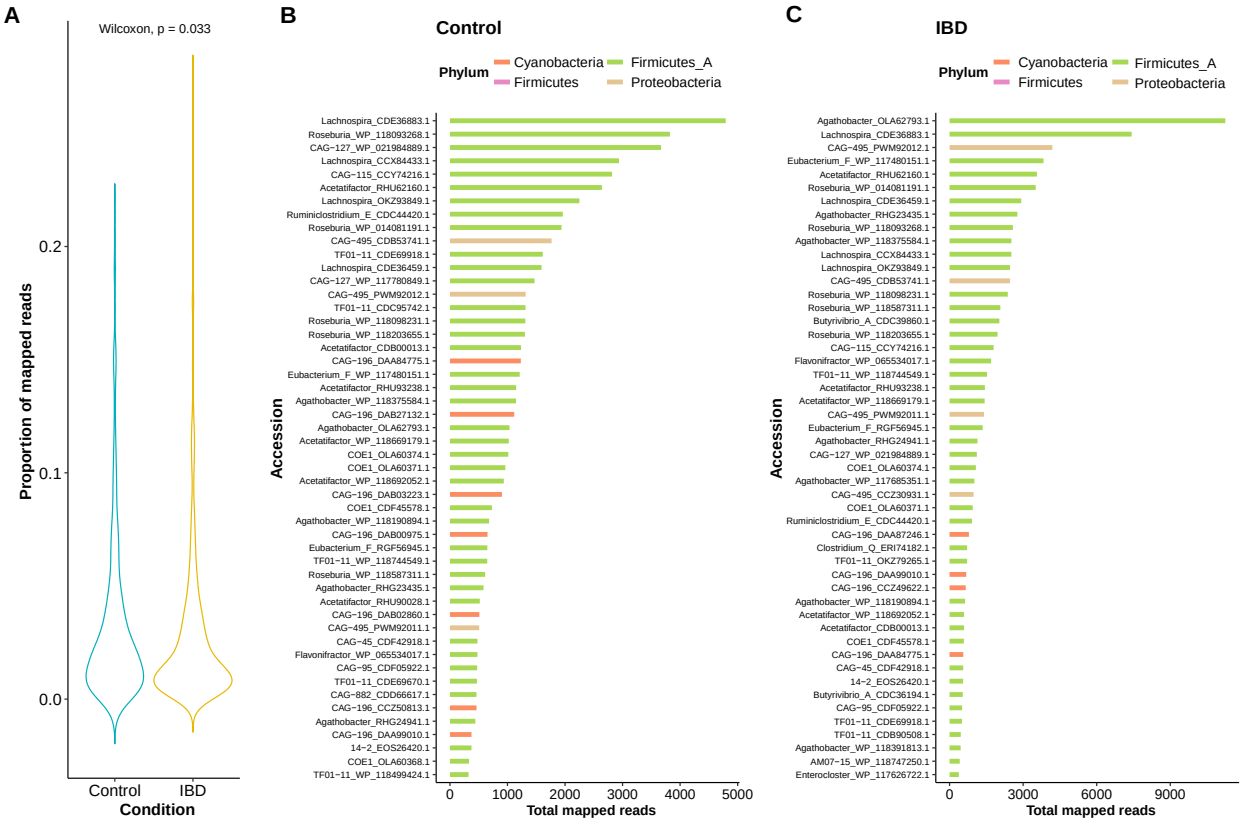

Figure 16. Characterization of the human gut flagellome. **A)** Proportion of reads relative to the total number of reads per sample mapped in samples from control and IBD patients. **B)** Top 50 flagellins mapped in metagenomes of samples from healthy subjects, ranked by abundance and summarized by genus. **C)** Top 50 flagellins mapped in metagenomes of samples from IBD patients, ranked by abundance and summarized by genus.

2. Characterization of silent flagellins

Silent flagellins were defined as flagellins that retain binding to TLR5 but poorly activate TLR5 signaling (Clasen et al., 2023). Based on previous observations from comparisons between pathogens and commensals residue composition, I differentiated between silent and non-silent flagellins based on their sequences. From the complete flagellin database, I considered as silent flagellins those containing an N-terminal S. Typhimurium FliC-like composition, an *R. hominis* C-terminal FlaB-like composition, and a fixed non-polar residue in the position 478. I found a total of 919 flagellin sequences with the FliC-FlaB-like composition. Restricting the candidates to those without a Lysin (K) or Arginine (R) in position 478 reduced the dataset to 195 candidates. The intersection of the predicted RhFlaB-like flagellins with the flagellins mapped in the human gut metagenomes produced a list of 145 RhFlaB-like flagellins present in the human gut. A final dereplication step at 99% identity to reduce sequence redundancy narrowed down the list to 81 silent flagellin candidates. The final list comprised flagellins belonging to 7 families spanning the phyla Proteobacteria and Firmicutes, where Lachnospiraceae are the most prevalent **(Figure 17A**). To increase the taxonomic diversity in the final set of flagellins, I generated an additional random set of flagellins abundant in the human gut, that did not meet the criteria for classification as silent (**Figure 17B**). To do this, I obtained 44 flagellins whose abundance was above an abundance cutoff of the mean of reads/flagellin, and which had an existing GTDB taxonomy. Together, they formed a final set of 125 flagellins that were screened to characterize their phenotypes of binding to TLR5 and activation of TLR5.

The experimental validation by other members in my team allowed us to classify the TLR5-related phenotypes of the final set of flagellins in three types: silent, evader, and stimulatory (Clasen et al., 2023). Among the set of predicted silent flagellins, 46% were experimentally confirmed to have a silent phenotype, 35% were classified as evaders and 19% were stimulatory **(Figure 17D)**. The majority of flagellins classified as silent belonged to 31 species from the family Lachnospiraceae, from which 9 belonged to species of Roseburia **(Figure 17C**).

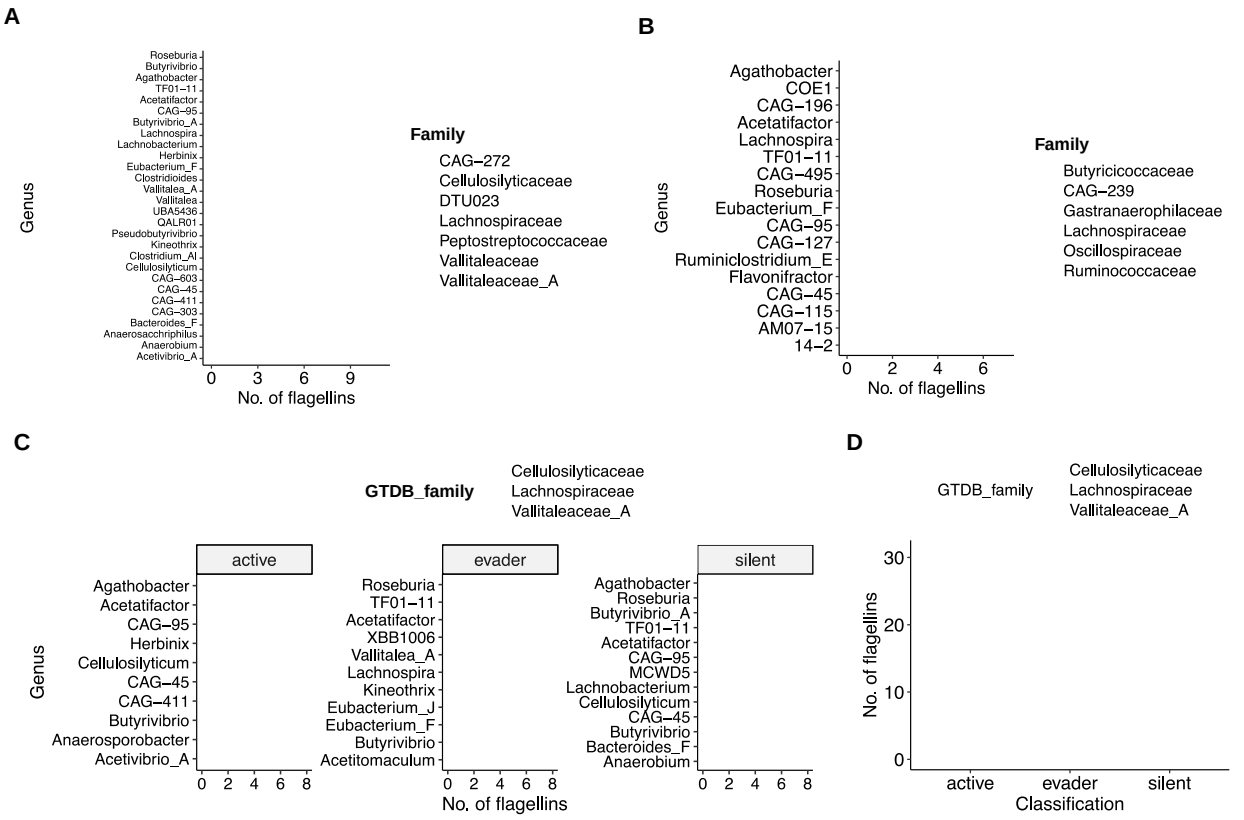

Figure 17. Characterization of RhFlaB-like flagellins. A) A random set of 44 flagellins from the healthy human gut were selected. **B)** A set of RhFlaB-like flagellins were identified by homology searches against the truncated N-terminal and C-terminal regions of S. Typhimurium, and R. hominis, respectively. An additional manual curation and residue content criteria produced a total of 83 putative silent flagellins. **C)** Observed phenotype of flagellins included in the screen of TLR5-related traits, as detailed in Clasen et al, 2023. **D)** Number of flagellin types summarized by the taxonomic family.

3. Phylogenetic analysis of TLR5-related traits

I reconstructed the phylogenetic relationships between the candidates of the screen including the silent flagellins, the silent FlaB of *Roseburia hominis*, the stimulatory FliC from *Salmonella* Typhimurium, the evader FlaA from *Helicobacter pylori* and the Chron's disease-associated flagellin CBir. Although the silent flagellins are encoded mostly by Lachnospiraceae members, the silent phenotype is widespread across unrelated flagellated bacteria (**Figure 18A**).

I found that 36% of the flagellins showed a significant phylogenetic signal (Moran's Index, $p < 0.05$) for both TLR5 binding and activation and interestingly, the local indicators of phylogenetic association (LIPA) primarily associate with active and evader flagellins, suggesting that the silent phenotype is not explained by the phylogenetic relationships of flagellins. (**Figure 18A**).

The same patterng was observed for the TLR5 binding and TLR5 activation traits in relation with their phylogeny. I observed that the phylogenetic correlation between TLR5 binding and activation was evident primarily among bacteria with small phylogenetic distances (**Figure 18B-C)**. This indicates that the observed similarities in phenotypes occur predominantly between closely related flagellins.

4. Comparison of species tree and genes trees in flagellins from screen candidates

I undertook a comparative analysis between the phylogenetic tree from flagellin sequences and the species tree representing these organisms. I found notable discrepancies in the tree topologies (**Figure 19**), indicating that the evolutionary path of the flagellin gene does not strictly follow the speciation events as captured by the species tree. These incongruencies in the topologies were particularly pronounced within the members of the Lachnospiraceae family, suggesting that the flagellin gene may have been subject to evolutionary pressures that differ from those observed at the species level in this family.

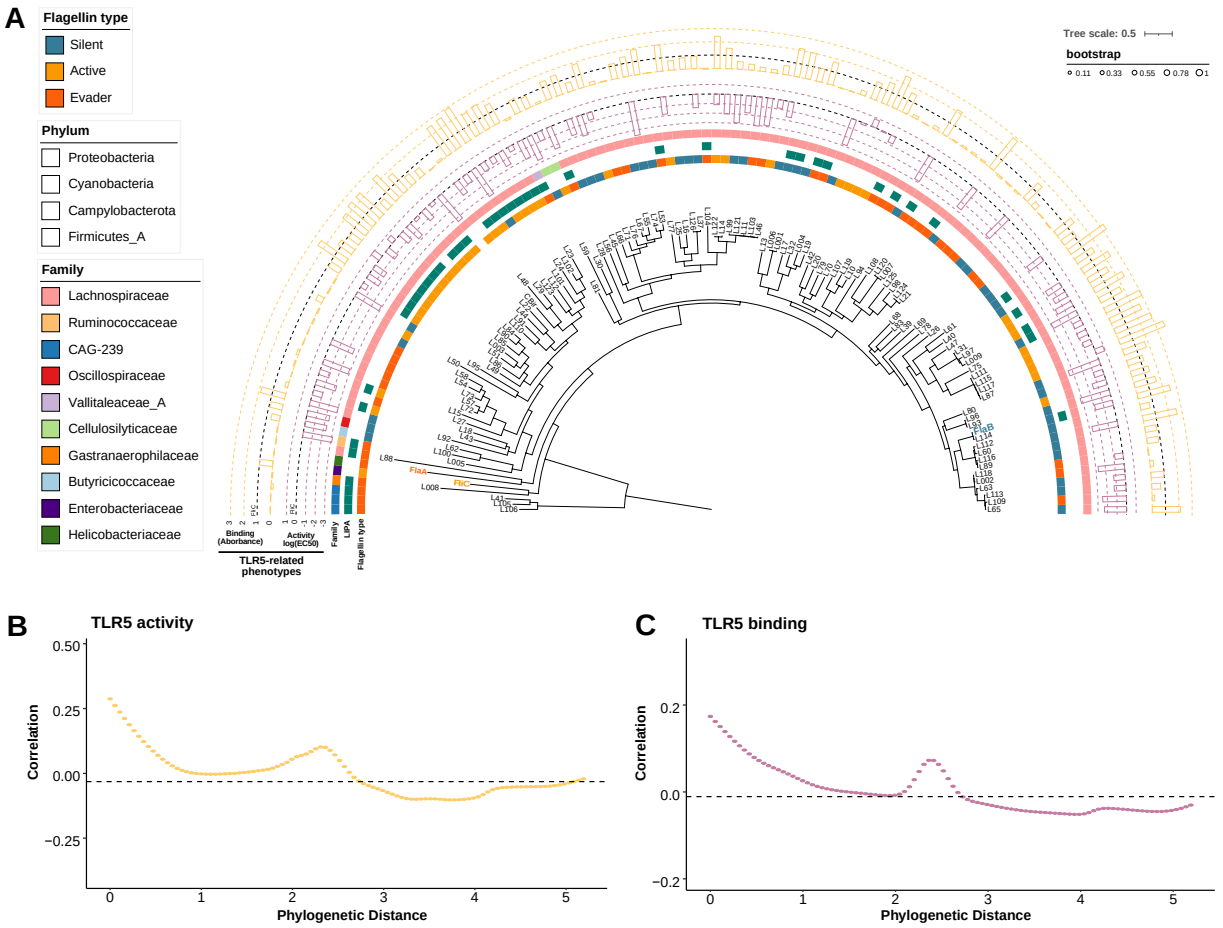

Figure 18. Phylogenetic reconstruction of screened RhFlaB-like flagellins. A) Phylogeny of flagellins, including FliC (S. Typhimurium), FlaA (H. pylori), CBir, and FlaB (R. hominis). The classification of flagellin types is shown as a color strip. The assessed phenotypes are shown as barplots, with values relative to FliC, values for FliC are represented as a black dashed horizontal line. The Local indicators of Phylogenetic Association (LIPA) are highlighted in dark teal. **B-C)** Phylogenetic correlogram using Moran's I and 100 bootstrap replicates of the TLR5 activity **(B)**, and TLR5 binding **(C)**; the vertical lines represent the 95% confidence interval. The red-highlighted values in the horizontal line represent the phylogenetic distance at which a positive phylogenetic correlation is observed. The horizontal dashed line depicts the correlation value expected under the null hypothesis.

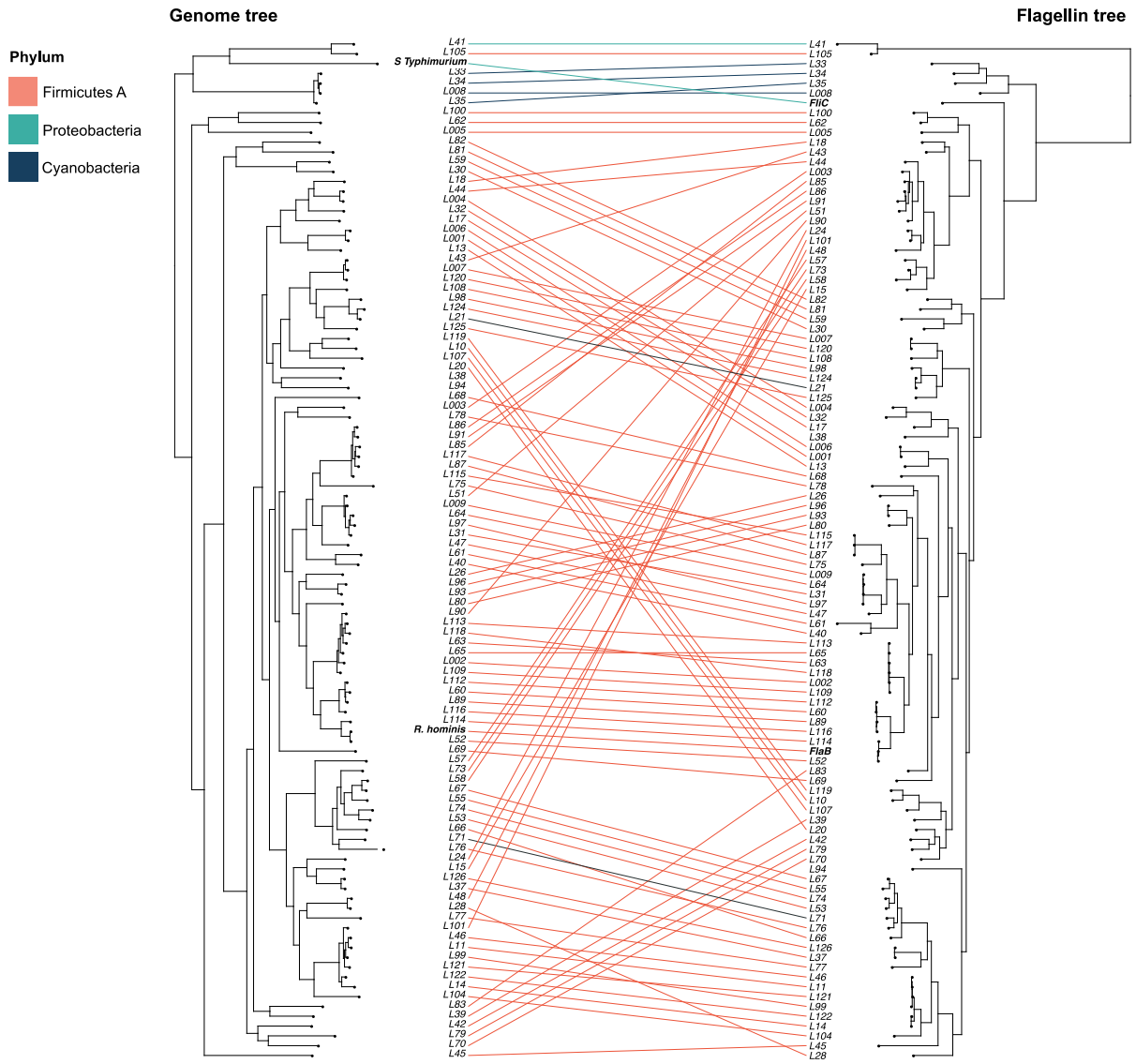

Figure 19. Tree topology comparisons between gene-tree and species-tree for RhFlaB-like flagellins and abundant flagellins from the healthy human gut. Left: Genome tree produced with PhyloPhlAn3. Right: Flagellin tree of the 81 RhFlaB-like candidates and 44 abundant flagellins in metagenomes from healthy humans The tree topology comparison was assessed using phylotools.

Discussion

I focused on flagellins from human gut commensals to unveil their evolutionary dynamics and their relationship with their interactions with TLR5. This resulted in the characterization of flagellin communities from samples of healthy individuals and IBD patients, and the development of a prediction method to identify putative silent flagellins in the human gut microbiome.

The flagellin communities in healthy humans and IBD patients show a dominance of Lachnospiraceae in both conditions, with a dominance of *Agathobacter faecis* in IBD patients and *Lachnospira* in healthy subjects, and a higher overall abundance of flagellin in healthy subjects. The analysis of the transcriptomic data for the same dataset showed that the expression landscape in healthy individuals is consistently dominated by Lachnospiraceae [8]. However, only a few flagellins identified in the ranked metagenomic profiling were also found in the transcriptomics data. One caveat of this alternative profiling approach is the specificity of the classification. This approach only used translated searches of metagenomic reads into protein sequences, without a refined step for the identification of peptide markers, as performed in Chapter 1 (see subsection ShortBRED analyses). Because of the exploratory nature of this analysis, my approach focused on obtaining the ranking of occurrences of flagellins within metagenomic samples. For this reason, further characterization should implement a more specific approach for the comparison of flagellin communities in the context of health and disease.

The *in-vitro* characterization of TLR5-related phenotypes confirmed that 45% of the predicted RhFlaB-like flagellins were classified as silent flagellins, aligning with the prediction. Notably, residue 478 in the C-terminal domain was a critical determinant for the silent phenotype, as it was observed among all flagellins classified as silent. Despite this association, the predictive accuracy of the current methodology requires further refinement. As shown in [8], chimeric variants of silent flagellins, containing a D0 domain from *S.* Typhimurium FliC produced a stimulatory phenotype. This finding underscores the important role of the C-terminal D0 domain in dictating the silent phenotype. Nevertheless, this set of flagellins successfully demonstrated that there is a population of flagellins in the human gut that distribute across phenotypes of stimulatory, active, and silent flagellins, challenging the classic understanding of flagellin-TLR5 interactions as a binary phenotype of activation or evasion [8], and stressing on the need to understand the genetic determinants of the different dynamics of TLR5-flagellin interactions in the context of commensal bacteria, the assembly of flagellin communities and the prevalence of silent flagellins in the host.

The phylogenetic analysis of silent flagellins showed that the silent phenotype is widespread among unrelated bacteria in the tree and does not form a monophyletic group. Two possible hypotheses could explain this pattern. First, the silent phenotype is ancestral to the Lachnospiraceae family and has been lost in some species. Recent research shows that although flagellar motility is lost in many host-associated bacteria, the host exerts control on the remaining flagellated bacteria in the host environment, thus promoting cooperation among commensal bacteria [6]. In this scenario, the host control of flagellin evolution within the human gut might have maintained a balanced flagellin community by favoring silent flagellins that can be recognized but not overstimulate the innate immunity, yet maintaining the capacity to prime the innate immunity for the recognition of flagellin by the adaptive immunity.

Alternatively, the silent phenotype might be a novelty in a few clades of Lachnospiraceae highly adapted to the gut environment. My analysis showed that the phylogenetic signal of TLR5-related traits in flagellins is maintained only among very closely related flagellins (phylogenetic distance < 0.1), and among flagellins with a higher degree of divergence (phylogenetic distance > 2). Most of the silent flagellins correspond to *Roseburia* species, where silent recognition might have evolved, and further mechanisms such as horizontal gene transfer (HGT), which is extensive in the human gut microbiome [235], could have mediated the acquisition of these flagellins by unrelated species within the human gut. This could partially support the incongruencies in tree topologies observed among the set of flagellins that were tested. This raises relevant questions about the evolution of these recognition mechanisms and their relevance in the ecology and adaptability of bacteria to the human gut. It also raises the question of how widespread are these phenotypes outside of human-associated bacteria and whether natural systems of TLR5-silent flagellins are observed in any other animal species. Several studies have shown that the TLR5 activation works in a species-specific manner, as observed in reptiles [236], highlighting the need to understand how widespread is silent recognition in natural systems.

Outlook

This research opens up an exciting field to identify specific residues or motifs in the C-terminal D0 domain that are associated with silent recognition, as well as the evolution of this phenotype in human gut commensals. The development of tools such as AlphaFold2 and powerful methods such as neural networks to predict function from sequence [237–239] can improve the functional classification of the flagellin database I have constructed, and accelerate the generation of hypothesis. In addition, implementation of structure-based methods and can aid sequence homology-based analysis by

predicting how sequence variation from stimulatory flagellins may affect flagellin structure and consequently the interaction with TLR5 [240]. This will be particularly advantageous given the challenge of using homology-based structural prediction alone in flagellins, which D0-D1 domains are highly conserved. Finally, structural analyses can be integrated into the evolutionary context of these bacteria, helping to link structural features to the evolutionary trajectories of these flagellins.

This lays the groundwork for further comprehensive analyses to predict silent flagellins based on sequence composition, while also exploring the evolutionary mechanisms underlying the diversification of silent flagellins in the human gut. From an evolutionary perspective, it is crucial to unravel the evolutionary trajectory of silent flagellins within the gut microbiome and to extend this question to other species. How common are silent flagellins in vertebrate hosts? What are the evolutionary dynamics shaping the diversification of flagellins and their corresponding receptors? Addressing these questions will improve our understanding of host-microbiome dynamics and shed light on the broader ecological significance of silent flagellins in different host species.

Methods

1. Flagellin profiling in human metagenomes

The selection process of candidate flagellins is summarized in Fig. 21. I annotated the flagellin database from Dalong and Reeves [59] with InterProScan [199]. I only retained sequences having both Pfam domain PF00669 (flagellin N-terminal domain, "ND") and PF00700 (flagellin C-terminal domain, "CD"). The filtered database comprised 33,051 flagellin protein sequences. To identify flagellins abundant in the human gut, I filtered the hits from metagenomes of healthy individuals from the IBD multi-omics database of Lloyd-Price *et al.* [233] using the median of read counts as a cutoff. The taxonomy of those hits was assigned using the taxonomizr [201] R package to obtain their taxonomic identifiers (*'taxid'*) from their NCBI protein accession, and a further search of their taxonomic identifiers within the Genome Taxonomy Database (GTDB) [202] release 95. I searched the remaining accessions using Entrez Direct [241] in the National Center for Biotechnology Information (NCBI) Identical Protein Groups database [242] to obtain their assembly accessions, which were further searched within the GTDB taxonomy. I ranked the resulting hits by their read counts and randomly selected across the list to get a broad taxonomic representation of flagellins from the human gut, reducing the

dataset to 44 flagellin candidates. I tested the random community assembly of the final flagellin candidates using different metrics: (i) Cmean and Pagel's Lambda, implemented in the phylosignal R package [243], and the standardized effect size (SES) of (ii) mean pairwise distance (MPD) and mean nearest taxon distance (MNTD), implemented in the picante R package [244]. Finally, I retrieved the coding sequence (CDS) of each protein accession from the NCBI Identical Protein Groups database [242].

2. Identification of RhFlaB-like candidates in the human gut

*Rh*FlaB-like candidates were identified using the flagellin database previously described (see Chapter 1: Flagellin Database). A truncated CD protein sequence of *Rh*FlaB (accession: WP_014081191.1) was mapped against the database using DIAMOND blastp [206] with parameters "*--max-target-seqs 0 --evalue 10-3 --very-sensitive*." The resulting 10,121 hits were filtered using the following thresholds (i) the median of the length of the alignment (100 amino acids), (ii) the sequence identity (39%), (iii) and the number of mismatches (n = 60). I mapped the truncated N-terminal domain of FliC from *Salmonella* Typhimurium (accession: AHA06007.1) against the previous hits with DIAMOND blastp using the same parameters as above, which reduced the list to 919 protein sequences. The resulting list was manually curated by selecting those sequences that did not have either amino acids Arginine or Lysine in position 478 of the alignment, leaving a total of 195 sequences.

To reduce the *Rh*FlaB-like flagellins to only those occurring in the human gut, I mapped gut metagenomes from the IBD multi-omics database [233] and the Franzosa *et al.* [234] dataset to the flagellin database using DIAMOND blastx with parameters "--max-target-seqs 1 --evalue 1e-3 --very-sensitive". The resulting 5,131 protein accessions of the homology search were intersected with the 195 RhFlaB-like candidates from the former step, resulting in 145 flagellins that met the sequence composition criteria and were also present in the human gut. I then dereplicated these flagellins using 99% sequence identity with CD-HIT [245,246] using the parameters "-c 0.99 -M 16000 -n 5", which reduced the dataset to 85 protein accessions. I retrieved their CDSs from the NCBI Identical Protein Groups database [242]. To avoid nucleotide sequence redundancy with the previously identified abundant flagellins, I merged the lists of CDSs, and the final nucleotide sequence list was dereplicated using CD-HIT with parameters "-c 0.99 -M 16000 -n 5", resulting in a final dataset of 85 flagellin sequences. A flowchart summarizing the selection process is illustrated in **Figure S1.**

3. Phylogenetic analysis of TLR5-related traits

To deeper investigate the relationships between flagellin phylogeny and their TLR5-related traits, I conducted a phylogenetic analysis. I created Multiple Sequence Alignments (MSAs) using CLUSTALW in Geneious Pro v2019 and masked the positions with over 70% gaps. The phylogeny was generated with FastTree in Geneious Pro v2019. The tree was visualized in iTOL [210] and quantitative traits (TLR5 binding, TLR5 activation) and discrete variables (taxonomy) were further mapped.

To assess whether the TLR5 binding and activation were randomly distributed across the flagellin phylogeny or were phylogenetically correlated, I tested for phylogenetic signals on the aforementioned traits (TLR5 binding, TLR5 activation) using C-mean and Phagel's Lambda implemented in the phylosignal R package [243]. Then, I obtained the Local Indicators of Phylogenetic Association (LIPA) and mapped them in the flagellin phylogeny using iTOL.

To assess the congruence between the flagellin tree and the species tree, I obtained the genomes associated to each flagellin accession number using Batch Entrez. I generated the genome tree using PhyloPhlan3 [247]. I compared the two phylogenies using a tanglegram generated with the phytools R package [248].

Conclusions

The research presented here serves as an exploratory effort to establish a baseline of knowledge about flagellin diversity, to discern patterns of this diversity across environments, and to determine the degree of natural selection involved in shaping this diversity. By laying this groundwork, I have paved the way for future investigations to apply this analytical framework to specific habitats and host taxa, potentially revealing more intricate details about the evolution of flagellin diversity in natural systems.

In this study, I broadened the scope of flagellin community characterization from simply cataloging bacterial habitats to exploring the influence of natural selection on the evolution of this integral protein. I demonstrated that pervasive negative selection and episodic positive selection maintain flagellin diversity in diverse host-associated and free-living environments. These mechanisms promote the adaptation of flagellins to a variety of niches. Although my research had a broad ecological scope, it provided the baseline information to generate targeted hypotheses. Elucidating the patterns of phylosymbiosis in vertebrate flagellin communities and further investigation of Lachnospiraceae-derived flagellins are relevant topics with the potential to provide a better understanding of TLR5-flagellin-mediated host-microbe interactions.

The methodology employed in this research was carefully designed to identify the signatures of positive selection in flagellins, thereby providing insights into the evolutionary dynamics shaping their diversity. My results suggest that varying environmental conditions and host selective pressures are important determinants in the evolutionary trajectory of flagellin genes. Consequently, these factors have a profound impact on the behavior and adaptive capacity of bacterial populations.

In addition, I have generated a comprehensive catalog delineating the taxonomic diversity of flagellin within the gut microbiota of both healthy individuals and individuals diagnosed with inflammatory bowel disease (IBD). This represents a first effort to delineate silent flagellin variants present within the human gut microbiome using publicly available sequence databases. Future investigations should prioritize the identification of sequence motifs associated with the silent phenotype, thereby shedding light on the underlying molecular mechanisms governing this intriguing phenotype.

In conclusion, this exploration of flagellin diversity and its evolutionary pressures not only broadens our understanding of microbial adaptability, but also underscores the complex interplay between bacteria and their habitats. This approach will allow us to understand how different environmental

conditions and host pressures influence the evolution of flagellin genes and, by extension, the behavior and adaptability of bacterial communities.

Bibliography

1. Nakamura, S., and Minamino, T. (2019). Flagella-Driven Motility of Bacteria. Biomolecules *9*. 10.3390/biom9070279.

2. Miyata, M., Robinson, R.C., Uyeda, T.Q.P., Fukumori, Y., Fukushima, S.-I., Haruta, S., Homma, M., Inaba, K., Ito, M., Kaito, C., et al. (2020). Tree of motility - A proposed history of motility systems in the tree of life. Genes Cells *25*, 6–21. 10.1111/gtc.12737.

3. Chaban B. Hughes HV. Beeby M (2015). The flagellum in bacterial pathogens: For motility and a whole lot more. Semin. Cell Dev. Biol. *46*, 91–103. 10.1016/j.semcdb.2015.10.032.

4. Yang, R.-S., and Chen, Y.-T. (2020). Flagellation of Shewanella oneidensis impacts bacterial fitness in different environments. Curr. Microbiol. *77*, 1790–1799. 10.1007/s00284-020-01999-0.

5. Gibson, K.H., Botting, J.M., Al-Otaibi, N., Maitre, K., Bergeron, J., Starai, V.J., and Hoover, T.R. (2023). Control of the flagellation pattern in Helicobacter pylori by FlhF and FlhG. J. Bacteriol. *205*, e0011023. 10.1128/jb.00110-23.

6. Sharp, C., and Foster, K.R. (2022). Host control and the evolution of cooperation in host microbiomes. Nat. Commun. *13*, 3567. 10.1038/s41467-022-30971-8.

7. Ramos, H.C., Rumbo, M., and Sirard, J.-C. (2004). Bacterial flagellins: mediators of pathogenicity and host immune responses in mucosa. Trends Microbiol. *12*, 509–517. 10.1016/j.tim.2004.09.002.

8. Clasen, S.J., Bell, M.E.W., Borbón, A., Lee, D.-H., Henseler, Z.M., de la Cuesta-Zuluaga, J., Parys, K., Zou, J., Wang, Y., Altmannova, V., et al. (2023). Silent recognition of flagellins from human gut commensal bacteria by Toll-like receptor 5. Sci Immunol *8*, eabq7001. 10.1126/sciimmunol.abq7001.

9. Hiroyuki Terashima, Akihiro Kawamoto, Yusuke V. Morimoto, Katsumi Imada, Tohru Minamino (2017). Structural differences in the bacterial flagellar motor among bacterial species. Biophysics and Physicobiology *14*, 191–198. 10.2142/biophysico.14.0_191.

10. Macnab, R.M. (2003). How bacteria assemble flagella. Annu. Rev. Microbiol. *57*, 77–100. 10.1146/annurev.micro.57.030502.090832.

11. Minamino, T., and Imada, K. (2015). The bacterial flagellar motor and its structural diversity. Trends Microbiol. *23*, 267–274. 10.1016/j.tim.2014.12.011.

12. Wadhwa, N., and Berg, H.C. (2022). Bacterial motility: machinery and mechanisms. Nat. Rev. Microbiol. *20*, 161–173. 10.1038/s41579-021-00626-4.

13. Laganenka, L., López, M.E., Colin, R., and Sourjik, V. (2020). Flagellum-Mediated Mechanosensing and RflP Control Motility State of Pathogenic Escherichia coli. MBio *11*. 10.1128/mBio.02269-19.

14. Belas, R. (2014). Biofilms, flagella, and mechanosensing of surfaces by bacteria. Trends Microbiol. *22*, 517–527. 10.1016/j.tim.2014.05.002.

15. Du, B., Gu, Y., Chen, G., Wang, G., and Liu, L. (2020). Flagellar motility mediates early-stage biofilm formation in oligotrophic aquatic environment. Ecotoxicol. Environ. Saf. *194*, 110340. 10.1016/j.ecoenv.2020.110340.

16. Li, S., Peng, C., Cheng, T., Wang, C., Guo, L., and Li, D. (2019). Nitrogen-cycling microbial community functional potential and enzyme activities in cultured biofilms with response to inorganic nitrogen availability. J. Environ. Sci. (China) *76*, 89–99. 10.1016/j.jes.2018.03.029.

17. Akahoshi, D.T., and Bevins, C.L. (2022). Flagella at the Host-Microbe Interface: Key Functions Intersect With Redundant Responses. Front. Immunol. *13*, 828758. 10.3389/fimmu.2022.828758.

18. Jarrell, K.F., and McBride, M.J. (2008). The surprisingly diverse ways that prokaryotes move. Nat. Rev. Microbiol. *6*, 466–476. 10.1038/nrmicro1900.

19. Wang, J.D., and Levin, P.A. (2009). Metabolism, cell growth and the bacterial cell cycle. Nat. Rev. Microbiol. *7*, 822–827. 10.1038/nrmicro2202.

20. Rohmer, L., Hocquet, D., and Miller, S.I. (2011). Are pathogenic bacteria just looking for food? Metabolism and microbial pathogenesis. Trends Microbiol. *19*, 341–348. 10.1016/j.tim.2011.04.003.

21. Sonnenburg, J.L., Xu, J., Leip, D.D., Chen, C.-H., Westover, B.P., Weatherford, J., Buhler, J.D., and Gordon, J.I. (2005). Glycan foraging in vivo by an intestine-adapted bacterial symbiont. Science *307*, 1955–1959. 10.1126/science.1109051.

22. Xu, J., Bjursell, M.K., Himrod, J., Deng, S., Carmichael, L.K., Chiang, H.C., Hooper, L.V., and Gordon, J.I. (2003). A genomic view of the human-Bacteroides thetaiotaomicron symbiosis. Science *299*, 2074–2076. 10.1126/science.1080029.

23. Corbin, K.D., Carnero, E.A., Dirks, B., Igudesman, D., Yi, F., Marcus, A., Davis, T.L., Pratley, R.E., Rittmann, B.E., Krajmalnik-Brown, R., et al. (2023). Host-diet-gut microbiome interactions influence human energy balance: a randomized clinical trial. Nat. Commun. *14*, 3161. 10.1038/s41467-023-38778-x.

24. Krajmalnik-Brown, R., Ilhan, Z.-E., Kang, D.-W., and DiBaise, J.K. (2012). Effects of gut microbes on nutrient absorption and energy regulation. Nutr. Clin. Pract. *27*, 201–214. 10.1177/0884533611436116.

25. Fatima, U., and Senthil-Kumar, M. (2015). Plant and pathogen nutrient acquisition strategies. Front. Plant Sci. *6*, 750. 10.3389/fpls.2015.00750.

26. Lowe-Power, T.M., Khokhani, D., and Allen, C. (2018). How Ralstonia solanacearum Exploits and Thrives in the Flowing Plant Xylem Environment. Trends Microbiol. *26*, 929–942. 10.1016/j.tim.2018.06.002.

27. Wood, T.K., González Barrios, A.F., Herzberg, M., and Lee, J. (2006). Motility influences biofilm architecture in Escherichia coli. Appl. Microbiol. Biotechnol. *72*, 361–367. 10.1007/s00253-005-0263-8.

28. Guttenplan, S.B., and Kearns, D.B. (2013). Regulation of flagellar motility during biofilm formation. FEMS Microbiol. Rev. *37*, 849–871. 10.1111/1574-6976.12018.

29. Li, G., Brown, P.J.B., Tang, J.X., Xu, J., Quardokus, E.M., Fuqua, C., and Brun, Y.V. (2012). Surface contact stimulates the just-in-time deployment of bacterial adhesins. Mol. Microbiol. *83*, 41–51. 10.1111/j.1365-2958.2011.07909.x.

30. Pratt, L.A., and Kolter, R. (1998). Genetic analysis of Escherichia coli biofilm formation: roles of flagella, motility, chemotaxis and type I pili. Mol. Microbiol. *30*, 285–293. 10.1046/j.1365-2958.1998.01061.x.

31. Haiko, J., and Westerlund-Wikström, B. (2013). The role of the bacterial flagellum in adhesion and virulence. Biology *2*, 1242–1267. 10.3390/biology2041242.

32. Horstmann, J.A., Lunelli, M., Cazzola, H., Heidemann, J., Kühne, C., Steffen, P., Szefs, S., Rossi, C., Lokareddy, R.K., Wang, C., et al. (2020). Methylation of Salmonella Typhimurium flagella promotes bacterial adhesion and host cell invasion. Nat. Commun. *11*, 2013. 10.1038/s41467-020-15738-3.

33. Colin, R., Ni, B., Laganenka, L., and Sourjik, V. (2021). Multiple functions of flagellar motility and chemotaxis in bacterial physiology. FEMS Microbiol. Rev. *45*. 10.1093/femsre/fuab038.

34. Krell, T., Lacal, J., Reyes-Darias, J.A., Jimenez-Sanchez, C., Sungthong, R., and Ortega-Calvo, J.J. (2013). Bioavailability of pollutants and chemotaxis. Curr. Opin. Biotechnol. *24*, 451–456. 10.1016/j.copbio.2012.08.011.

35. Matilla, M.A., and Krell, T. (2018). The effect of bacterial chemotaxis on host infection and pathogenicity. FEMS Microbiol. Rev. *42*. 10.1093/femsre/fux052.

36. Raina, J.-B., Fernandez, V., Lambert, B., Stocker, R., and Seymour, J.R. (2019). The role of microbial motility and chemotaxis in symbiosis. Nat. Rev. Microbiol. *17*, 284–294. 10.1038/s41579-019-0182-9.

37. Wiles, T.J., Schlomann, B.H., Wall, E.S., Betancourt, R., Parthasarathy, R., and Guillemin, K. (2020). Swimming motility of a gut bacterial symbiont promotes resistance to intestinal expulsion and enhances inflammation. PLoS Biol. *18*, e3000661. 10.1371/journal.pbio.3000661.

38. Diepold, A., and Armitage, J.P. (2015). Type III secretion systems: the bacterial flagellum and the injectisome. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*. 10.1098/rstb.2015.0020.

39. Gophna, U., Ron, E.Z., and Graur, D. (2003). Bacterial type III secretion systems are ancient and evolved by multiple horizontal-transfer events. Gene *312*, 151–163. 10.1016/s0378-1119(03)00612-7.

40. Khan, S., and Scholey, J.M. (2018). Assembly, Functions and Evolution of Archaella, Flagella and Cilia. Curr. Biol. *28*, R278–R292. 10.1016/j.cub.2018.01.085.

41. Liu, R., and Ochman, H. (2007). Stepwise formation of the bacterial flagellar system. Proc. Natl. Acad. Sci. U. S. A. *104*, 7116–7121. 10.1073/pnas.0700266104.

42. Snyder, L.A.S., Loman, N.J., Fütterer, K., and Pallen, M.J. (2009). Bacterial flagellar diversity and evolution: seek simplicity and distrust it? Trends Microbiol. *17*, 1–5.

10.1016/j.tim.2008.10.002.

43. Byrne Brenda, and Swanson Michele S. (1998). Expression of Legionella pneumophilaVirulence Traits in Response to Growth Conditions. Infect. Immun. *66*, 3029–3034. 10.1128/iai.66.7.3029-3034.1998.

44. Soutourina, O.A., Semenova, E.A., Parfenova, V.V., Danchin, A., and Bertin, P. (2001). Control of bacterial motility by environmental factors in polarly flagellated and peritrichous bacteria isolated from Lake Baikal. Appl. Environ. Microbiol. *67*, 3852–3859. 10.1128/AEM.67.9.3852-3859.2001.

45. Soutourina, O.A., and Bertin, P.N. (2003). Regulation cascade of flagellar expression in Gram-negative bacteria. FEMS Microbiol. Rev. *27*, 505–523. 10.1016/S0168-6445(03)00064-0.

46. Komeda, Y., Kutsukake, K., and Iino, T. (1980). Definition of additional flagellar genes in Escherichia coli K12. Genetics *94*, 277–290. 10.1093/genetics/94.2.277.

47. Kutsukake, K., Ohya, Y., Yamaguchi, S., and Iino, T. (1988). Operon structure of flagellar genes in Salmonella typhimurium. Mol. Gen. Genet. *214*, 11–15. 10.1007/BF00340172.

48. De Maayer, P., and Cowan, D.A. (2016). Flashy flagella: flagellin modification is relatively common and highly versatile among the Enterobacteriaceae. BMC Genomics *17*, 377. 10.1186/s12864-016-2735-x.

49. Furness, R.B., Fraser, G.M., Hay, N.A., and Hughes, C. (1997). Negative feedback from a Proteus class II flagellum export defect to the flhDC master operon controlling cell division and flagellum assembly. J. Bacteriol. *179*, 5585–5588. 10.1128/jb.179.17.5585-5588.1997.

50. De Maayer, P., Pillay, T., and Coutinho, T.A. (2020). Flagella by numbers: comparative genomic analysis of the supernumerary flagellar systems among the Enterobacterales. BMC Genomics *21*, 670. 10.1186/s12864-020-07085-w.

51. Rossez, Y., Wolfson, E.B., Holmes, A., Gally, D.L., and Holden, N.J. (2015). Bacterial Flagella: Twist and Stick, or Dodge across the Kingdoms. PLoS Pathog. *11*, 1–15. 10.1371/journal.ppat.1004483.

52. Beatson, S., Minamino, T., and Pallen, M. (2006). Variation in bacterial flagellins: from sequence to structure. Trends Microbiol. *14*, 149–151. 10.1016/j.tim.2006.02.001.

53. McCarter, L.L. (2004). Dual flagellar systems enable motility under different circumstances. J. Mol. Microbiol. Biotechnol. *7*, 18–29. 10.1159/000077866.

54. Pallen, M.J., Penn, C.W., and Chaudhuri, R.R. (2005). Bacterial flagellar diversity in the post-genomic era. Trends Microbiol. *13*, 143–149. 10.1016/j.tim.2005.02.008.

55. Song, W.S., Jeon, Y.J., Namgung, B., Hong, M., and Yoon, S.I. (2017). A conserved TLR5 binding and activation hot spot on flagellin. Sci. Rep. *7*, 1–11. 10.1038/srep40878.

56. Nedeljković, M., Sastre, D.E., and Sundberg, E.J. (2021). Bacterial Flagellar Filament: A Supramolecular Multifunctional Nanostructure. Int. J. Mol. Sci. *22*. 10.3390/ijms22147521.

57. Xu, M., Xie, Y., Tan, M., Zheng, K., Xiao, Y., Jiang, C., Zhao, F., Zeng, T., and Wu, Y. (2019). The N-terminal D1 domain of Treponema pallidum flagellin binding to TLR5 is required but not sufficient in activation of TLR5. J. Cell. Mol. Med. *23*, 7490–7504. 10.1111/jcmm.14617.

58. Forstnerič, V., Ivičak-Kocjan, K., Plaper, T., Jerala, R., and Benčina, M. (2017). The role of the C-terminal D0 domain of flagellin in activation of Toll like receptor 5. PLoS Pathog. *13*, 1–20. 10.1371/journal.ppat.1006574.

59. Dalong, H., and Reeves, P. (2020). The remarkable dual-level diversity of prokaryotic flagellins. mSystems *5*, 1–15.

60. Thomson, N.M., Rossmann, F.M., Ferreira, J.L., Matthews-Palmer, T.R., Beeby, M., and Pallen, M.J. (2018). Bacterial Flagellins: Does Size Matter? Trends in Microbiology. 10.1016/j.tim.2017.11.010.

61. Steimle, A., Menz, S., Bender, A., Ball, B., Weber, A.N.R., Hagemann, T., Lange, A., Maerz, J.K., Parusel, R., Michaelis, L., et al. (2019). Flagellin hypervariable region determines symbiotic properties of commensal escherichia coli strains. PLoS Biol. *17*, 1–27. 10.1371/journal.pbio.3000334.

62. Smith, K.D., Andersen-Nissen, E., Hayashi, F., Strobe, K., Bergman, M.A., Rassoulian Barrett, S.L., Cookson, B.T., and Aderem, A. (2003). Toll-like receptor 5 recognizes a conserved site on flagellin required for protofilament formation and bacterial motility. Nat. Immunol. *4*, 1247–1253. 10.1038/ni1011.

63. Yoon, S.-I., Kurnasov, O., Natarajan, V., Hong, M., Gudkov, A.V., Osterman, A.L., and

Wilson, I.A. (2012). Structural Basis of TLR5-Flagellin Recognition and Signaling. Science *335*, 859–865. 10.1126/science.1215584.

64. Lu, Y., and Swartz, J.R. (2016). Functional properties of flagellin as a stimulator of innate immunity. Sci. Rep. *6*, 18379. 10.1038/srep18379.

65. Zhao, Y., Yang, J., Shi, J., Gong, Y.-N., Lu, Q., Xu, H., Liu, L., and Shao, F. (2011). The NLRC4 inflammasome receptors for bacterial flagellin and type III secretion apparatus. Nature *477*, 596–600. 10.1038/nature10510.

66. Gómez-Gómez, L., and Boller, T. (2000). FLS2: an LRR receptor-like kinase involved in the perception of the bacterial elicitor flagellin in Arabidopsis. Mol. Cell *5*, 1003–1011. 10.1016/s1097-2765(00)80265-8.

67. Hind, S.R., Strickler, S.R., Boyle, P.C., Dunham, D.M., Bao, Z., O'Doherty, I.M., Baccile, J.A., Hoki, J.S., Viox, E.G., Clarke, C.R., et al. (2016). Tomato receptor FLAGELLIN-SENSING 3 binds flgII-28 and activates the plant immune system. Nat Plants *2*, 16128. 10.1038/nplants.2016.128.

68. Murakami, T., Katsuragi, Y., Hirai, H., Wataya, K., Kondo, M., and Che, F.-S. (2022). Distribution of flagellin CD2-1, flg22, and flgII-28 recognition systems in plant species and regulation of plant immune responses through these recognition systems. Biosci. Biotechnol. Biochem. *86*, 490–501. 10.1093/bbb/zbac007.

69. Hayashi, F., Hayashi, F., Smith, K.D., Smith, K.D., Ozinsky, A., Ozinsky, A., Hawn, T.R., Hawn, T.R., Yi, E.C., Yi, E.C., et al. (2001). The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. Nature *410*, 1099–1103. 10.1038/35074106.

70. Alexander, K.L., Targan, S.R., and Elson, C.O. (2014). Microbiota activation and regulation of innate and adaptive immunity. Immunol. Rev. *260*, 206–220. 10.1111/imr.12180.

71. Patterson, A.M., Mulder, I.E., Travis, A.J., Lan, A., Cerf-Bensussan, N., Gaboriau-Routhiau, V., Garden, K., Logan, E., Delday, M.I., Coutts, A.G.P., et al. (2017). Human gut symbiont Roseburia hominis promotes and regulates innate immunity. Front. Immunol. *8*, 1–14. 10.3389/fimmu.2017.01166.

72. Bates, J.T., Graff, A.H., Phipps, J.P., Grayson, J.M., and Mizel, S.B. (2011). Enhanced antigen processing of flagellin fusion proteins promotes the antigen-specific CD8+ T cell response

independently of TLR5 and MyD88. J. Immunol. *186*, 6255–6262. 10.4049/jimmunol.1001855.

73. Parys, K., Colaianni, N.R., Lee, H.-S., Hohmann, U., Edelbacher, N., Trgovcevic, A., Blahovska, Z., Lee, D., Mechtler, A., Muhari-Portik, Z., et al. (2021). Signatures of antagonistic pleiotropy in a bacterial flagellin epitope. Cell Host Microbe *29*, 620-634.e9. 10.1016/j.chom.2021.02.008.

74. Andersen-Nissen, E., Smith, K.D., Strobe, K.L., Rassoulian Barrett, S.L., Cookson, B.T., Logan, S.M., and Aderem, A. (2005). Evasion of Toll-like receptor 5 by flagellated bacteria. Proc. Natl. Acad. Sci. U. S. A. *102*, 9247–9252. 10.1073/pnas.0502040102.

75. Mortimer, C.K.B., Gharbia, S.E., Logan, J.M.J., Peters, T.M., and Arnold, C. (2007). Flagellin gene sequence evolution in Salmonella. Infect. Genet. Evol. *7*, 411–415. 10.1016/j.meegid.2006.12.001.

76. Beutin, L., Strauch, E., Zimmermann, S., Kaulfuss, S., Schaudinn, C., Männel, A., and Gelderblom, H.R. (2005). Genetical and functional investigation of fliC genes encoding flagellar serotype H4 in wildtype strains of Escherichia coli and in a laboratory E. coli K-12 strain expressing flagellar antigen type H48. BMC Microbiol. *5*, 4. 10.1186/1471-2180-5-4.

77. McQuiston, J.R., Parrenas, R., Ortiz-Rivera, M., Gheesling, L., Brenner, F., and Fields, P.I. (2004). Sequencing and comparative analysis of flagellin genes fliC, fljB, and flpA from Salmonella. J. Clin. Microbiol. *42*, 1923–1932. 10.1128/JCM.42.5.1923-1932.2004.

78. Scott, D.R., Marcus, E.A., Wen, Y., Oh, J., and Sachs, G. (2007). Gene expression in vivo shows that Helicobacter pylori colonizes an acidic niche on the gastric surface. Proc. Natl. Acad. Sci. U. S. A. *104*, 7235–7240. 10.1073/pnas.0702300104.

79. Matilla, M.A., Velando, F., Tajuelo, A., Martín-Mora, D., Xu, W., Sourjik, V., Gavira, J.A., and Krell, T. (2022). Chemotaxis of the Human Pathogen Pseudomonas aeruginosa to the Neurotransmitter Acetylcholine. MBio *13*, e0345821. 10.1128/mbio.03458-21.

80. Schwarzer, C., Fischer, H., and Machen, T.E. (2016). Chemotaxis and binding of Pseudomonas aeruginosa to scratch-wounded human cystic fibrosis airway epithelial cells. PLoS One *11*, e0150109. 10.1371/journal.pone.0150109.

81. Sanguankiattichai, N., Buscaill, P., and Preston, G.M. (2022). How bacteria overcome flagellin pattern recognition in plants. Curr. Opin. Plant Biol. *67*, 102224. 10.1016/j.pbi.2022.102224.

82. Miao, E.A., Andersen-Nissen, E., Warren, S.E., and Aderem, A. (2007). TLR5 and Ipaf: dual sensors of bacterial flagellin in the innate immune system. Semin. Immunopathol. *29*, 275–288. 10.1007/s00281-007-0078-z.

83. Sutterwala, F.S., and Flavell, R.A. (2009). NLRC4/IPAF: a CARD carrying member of the NLR family. Clin. Immunol. *130*, 2–6. 10.1016/j.clim.2008.08.011.

84. Miao, E.A., Alpuche-Aranda, C.M., Dors, M., Clark, A.E., Bader, M.W., Miller, S.I., and Aderem, A. (2006). Cytoplasmic flagellin activates caspase-1 and secretion of interleukin 1β via Ipaf. Nat. Immunol. *7*, 569–575. 10.1038/ni1344.

85. Franchi, L., Amer, A., Body-Malapel, M., Kanneganti, T.-D., Özören, N., Jagirdar, R., Inohara, N., Vandenabeele, P., Bertin, J., Coyle, A., et al. (2006). Cytosolic flagellin requires Ipaf for activation of caspase-1 and interleukin 1β in salmonella-infected macrophages. Nat. Immunol. *7*, 576–582. 10.1038/ni1346.

86. Yang, X., Yang, F., Wang, W., Lin, G., Hu, Z., Han, Z., Qi, Y., Zhang, L., Wang, J., Sui, S.-F., et al. (2018). Structural basis for specific flagellin recognition by the NLR protein NAIP5. Cell Res. *28*, 35–47. 10.1038/cr.2017.148.

87. Halff, E.F., Diebolder, C.A., Versteeg, M., Schouten, A., Brondijk, T.H.C., and Huizinga, E.G. (2012). Formation and Structure of a NAIP5-NLRC4 Inflammasome Induced by Direct Interactions with Conserved N- and C-terminal Regions of Flagellin*. J. Biol. Chem. *287*, 38460–38472. 10.1074/jbc.M112.393512.

88. Lightfield, K.L., Persson, J., Brubaker, S.W., Witte, C.E., von Moltke, J., Dunipace, E.A., Henry, T., Sun, Y.-H., Cado, D., Dietrich, W.F., et al. (2008). Critical function for Naip5 in inflammasome activation by a conserved carboxy-terminal domain of flagellin. Nat. Immunol. *9*, 1171–1178. 10.1038/ni.1646.

89. Cullender, T.C., Chassaing, B., Janzon, A., Kumar, K., Muller, C.E., Werner, J.J., Angenent, L.T., Bell, M.E., Hay, A.G., Peterson, D.A., et al. (2013). Innate and adaptive immunity interact to quench microbiome flagellar motility in the gut. Cell Host Microbe *14*, 571–581. 10.1016/j.chom.2013.10.009.

90. Carvalho, F.A., Koren, O., Goodrich, J.K., Johansson, M.E.V., Nalbantoglu, I., Aitken, J.D., Su, Y., Chassaing, B., Walters, W.A., González, A., et al. (2012). Transient inability to manage

proteobacteria promotes chronic gut inflammation in TLR5-deficient mice. Cell Host Microbe *12*, 139–152. 10.1016/j.chom.2012.07.004.

91. Hawn, T.R., Verbon, A., Lettinga, K.D., Zhao, L.P., Li, S.S., Laws, R.J., Skerrett, S.J., Beutler, B., Schroeder, L., Nachman, A., et al. (2003). A common dominant TLR5 stop codon polymorphism abolishes flagellin signaling and is associated with susceptibility to legionnaires' disease. J. Exp. Med. *198*, 1563–1572. 10.1084/jem.20031220.

92. Merx, S., Zimmer, W., Neumaier, M., and Ahmad-Nejad, P. (2006). Characterization and functional investigation of single nucleotide polymorphisms (SNPs) in the human TLR5 gene. Hum. Mutat. *27*, 293. 10.1002/humu.9409.

93. Hacquard, S., Spaepen, S., Garrido-Oter, R., and Schulze-Lefert, P. (2017). Interplay Between Innate Immunity and the Plant Microbiota. Annu. Rev. Phytopathol. *55*, 565–589. 10.1146/annurev-phyto-080516-035623.

94. Albert, M., Jehle, A.K., Lipschis, M., Mueller, K., Zeng, Y., and Felix, G. (2010). Regulation of cell behaviour by plant receptor kinases: Pattern recognition receptors as prototypical models. Eur. J. Cell Biol. *89*, 200–207. 10.1016/j.ejcb.2009.11.015.

95. DeFalco, T.A., and Zipfel, C. (2021). Molecular mechanisms of early plant pattern-triggered immune signaling. Mol. Cell *81*, 4346. 10.1016/j.molcel.2021.09.028.

96. Roberts, R., Liu, A.E., Wan, L., Geiger, A.M., Hind, S.R., Rosli, H.G., and Martin, G.B. (2020). Molecular Characterization of Differences between the Tomato Immune Receptors Flagellin Sensing 3 and Flagellin Sensing 2. Plant Physiol. *183*, 1825–1837. 10.1104/pp.20.00184.

97. Katsuragi, Y., Takai, R., Furukawa, T., Hirai, H., Morimoto, T., Katayama, T., Murakami, T., and Che, F.-S. (2015). CD2-1, the C-Terminal Region of Flagellin, Modulates the Induction of Immune Responses in Rice. Mol. Plant. Microbe. Interact. *28*, 648–658. 10.1094/MPMI-11-14-0372-R.

98. Moroz, N., and Tanaka, K. (2020). FlgII-28 Is a Major Flagellin-Derived Defense Elicitor in Potato. Mol. Plant. Microbe. Interact. *33*, 247–255. 10.1094/MPMI-06-19-0164-R.

99. Fliegmann, J., and Felix, G. (2016). Immunity: Flagellin seen from all sides. Nat Plants *2*, 16136. 10.1038/nplants.2016.136.

100. Wei, H.-L., Chakravarthy, S., Worley, J.N., and Collmer, A. (2013). Consequences of flagellin export through the type III secretion system of Pseudomonas syringae reveal a major difference in the innate immune systems of mammals and the model plant Nicotiana benthamiana. Cell. Microbiol. *15*, 601–618. 10.1111/cmi.12059.

101. Buscaill, P., and van der Hoorn, R.A.L. (2021). Defeated by the nines: nine extracellular strategies to avoid microbe-associated molecular patterns recognition in plants. Plant Cell *33*, 2116–2130. 10.1093/plcell/koab109.

102. Sun, W., Dunning, F.M., Pfund, C., Weingarten, R., and Bent, A.F. (2006). Within-Species Flagellin Polymorphism in Xanthomonas campestris pv campestris and Its Impact on Elicitation of Arabidopsis FLAGELLIN SENSING2–Dependent Defenses. Plant Cell *18*, 764–779. 10.1105/tpc.105.037648.

103. Cheng, J.H.T., Bredow, M., Monaghan, J., and diCenzo, G.C. (2021). Proteobacteria Contain Diverse flg22 Epitopes That Elicit Varying Immune Responses in Arabidopsis thaliana. Mol. Plant. Microbe. Interact. *34*, 504–510. 10.1094/MPMI-11-20-0314-SC.

104. Colaianni, N.R., Parys, K., Lee, H.-S., Conway, J.M., Kim, N.H., Edelbacher, N., Mucyn, T.S., Madalinski, M., Law, T.F., Jones, C.D., et al. (2021). A complex immune response to flagellin epitope variation in commensal communities. Cell Host Microbe *29*, 635-649.e9. 10.1016/j.chom.2021.02.006.

105. Galkin, V.E., Yu, X., Bielnicki, J., Heuser, J., Ewing, C.P., Guerry, P., and Egelman, E.H. (2008). Divergence of quaternary structures among bacterial flagellar filaments. Science *320*, 382–385. 10.1126/science.1155307.

106. Yang, J., Zhang, E., Liu, F., Zhang, Y., Zhong, M., Li, Y., Zhou, D., Chen, Y., Cao, Y., Xiao, Y., et al. (2014). Flagellins of Salmonella Typhi and nonpathogenic Escherichia coli are differentially recognized through the NLRC4 pathway in macrophages. J. Innate Immun. *6*, 47–57. 10.1159/000351476.

107. Bardoel, B.W., van der Ent, S., Pel, M.J.C., Tommassen, J., Pieterse, C.M.J., van Kessel, K.P.M., and van Strijp, J.A.G. (2011). Pseudomonas evades immune recognition of flagellin in both mammals and plants. PLoS Pathog. *7*, e1002206. 10.1371/journal.ppat.1002206.

108. Deng, Y., Chen, H., Li, C., Xu, J., Qi, Q., Xu, Y., Zhu, Y., Zheng, J., Peng, D., Ruan, L., et al.

(2019). Endophyte Bacillus subtilis evade plant defense by producing lantibiotic subtilomycin to mask self-produced flagellin. Commun Biol *2*, 368. 10.1038/s42003-019-0614-0.

109. Sanders, C.J., Yu, Y., Moore, D.A., 3rd, Williams, I.R., and Gewirtz, A.T. (2006). Humoral immune response to flagellin requires T cells and activation of innate immunity. J. Immunol. *177*, 2810–2818. 10.4049/jimmunol.177.5.2810.

110. Gewirtz, A.T., Vijay-Kumar, M., Brant, S.R., Duerr, R.H., Nicolae, D.L., and Cho, J.H. (2006). Dominant-negative TLR5 polymorphism reduces adaptive immune response to flagellin and negatively associates with Crohn's disease. Am. J. Physiol. Gastrointest. Liver Physiol. *290*, G1157-63. 10.1152/ajpgi.00544.2005.

111. Parish, C.R. (1996). Immune deviation: a historical perspective. Immunol. Cell Biol. *74*, 449–456. 10.1038/icb.1996.75.

112. Pabst, O. (2012). New concepts in the generation and functions of IgA. Nat. Rev. Immunol. *12*, 821–832. 10.1038/nri3322.

113. Liu, G., Zhang, H., Zhao, C., and Zhang, H. (2020). Evolutionary History of the Toll-Like Receptor Gene Family across Vertebrates. Genome Biol. Evol. *12*, 3615–3634. 10.1093/gbe/evz266.

114. Karasov, T.L., Horton, M.W., and Bergelson, J. (2014). Genomic variability as a driver of plant-pathogen coevolution? Curr. Opin. Plant Biol. *18*, 24–30. 10.1016/j.pbi.2013.12.003.

115. Ma, W., Dong, F.F.T., Stavrinides, J., and Guttman, D.S. (2006). Type III effector diversification via both pathoadaptation and horizontal transfer in response to a coevolutionary arms race. PLoS Genet. *2*, e209. 10.1371/journal.pgen.0020209.

116. Barreiro, L.B., Ben-Ali, M., Quach, H., Laval, G., Patin, E., Pickrell, J.K., Bouchier, C., Tichit, M., Neyrolles, O., Gicquel, B., et al. (2009). Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. PLoS Genet. *5*, e1000562. 10.1371/journal.pgen.1000562.

117. Clarke, C.R., Chinchilla, D., Hind, S.R., Taguchi, F., Miki, R., Ichinose, Y., Martin, G.B., Leman, S., Felix, G., and Vinatzer, B.A. (2013). Allelic variation in two distinct Pseudomonas syringae flagellin epitopes modulates the strength of plant immune responses but not bacterial motility. New Phytol. *200*, 847–860. 10.1111/nph.12408.

118. Felix, G., Duran, J.D., Volko, S., and Boller, T. (1999). Plants have a sensitive perception system for the most conserved domain of bacterial flagellin. Plant J. *18*, 265–276. 10.1046/j.1365-313x.1999.00265.x.

119. Vetter, M.M., Kronholm, I., He, F., Häweker, H., Reymond, M., Bergelson, J., Robatzek, S., and de Meaux, J. (2012). Flagellin perception varies quantitatively in Arabidopsis thaliana and its relatives. Mol. Biol. Evol. *29*, 1655–1667. 10.1093/molbev/mss011.

120. McCann, H.C., Nahal, H., Thakur, S., and Guttman, D.S. (2012). Identification of innate immunity elicitors using molecular signatures of natural selection. Proc. Natl. Acad. Sci. U. S. A. *109*, 4215–4220. 10.1073/pnas.1113893109.

121. Cai, R., Lewis, J., Yan, S., Liu, H., Clarke, C.R., Campanile, F., Almeida, N.F., Studholme, D.J., Lindeberg, M., Schneider, D., et al. (2011). The plant pathogen Pseudomonas syringae pv. tomato is genetically monomorphic and under strong selection to evade tomato immunity. PLoS Pathog. *7*, e1002130. 10.1371/journal.ppat.1002130.

122. Werling, D., Jann, O.C., Offord, V., Glass, E.J., and Coffey, T.J. (2009). Variation matters: TLR structure and species-specific pathogen recognition. Trends Immunol. *30*, 124–130. 10.1016/j.it.2008.12.001.

123. Grueber, C.E., Wallis, G.P., and Jamieson, I.G. (2014). Episodic positive selection in the evolution of avian toll-like receptor innate immunity genes. PLoS One *9*, e89632. 10.1371/journal.pone.0089632.

124. Wlasiuk, G., Khan, S., Switzer, W.M., and Nachman, M.W. (2009). A history of recurrent positive selection at the toll-like receptor 5 in primates. Mol. Biol. Evol. *26*, 937–949. 10.1093/molbev/msp018.

125. Areal, H., Abrantes, J., and Esteves, P.J. (2011). Signatures of positive selection in Toll-like receptor (TLR) genes in mammals. BMC Evol. Biol. *11*, 368. 10.1186/1471-2148-11-368.

126. Wlasiuk, G., and Nachman, M.W. (2010). Adaptation and constraint at Toll-like receptors in primates. Mol. Biol. Evol. *27*, 2172–2186. 10.1093/molbev/msq104.

127. Su, Q., Chen, Y., and He, H. (2023). Molecular evolution of Toll-like receptors in rodents. Integr. Zool. 10.1111/1749-4877.12746.

128. Andersen-Nissen, E., Smith, K.D., Bonneau, R., Strong, R.K., and Aderem, A. (2007). A conserved surface on Toll-like receptor 5 recognizes bacterial flagellin. J. Exp. Med. *204*, 393–403. 10.1084/jem.20061400.

129. Sharma, V., Hecker, N., Walther, F., Stuckas, H., and Hiller, M. (2020). Convergent Losses of TLR5 Suggest Altered Extracellular Flagellin Detection in Four Mammalian Lineages. Mol. Biol. Evol. *37*, 1847–1854. 10.1093/molbev/msaa058.

130. Lewis, W.H., Tahon, G., Geesink, P., Sousa, D.Z., and Ettema, T.J.G. (2021). Innovations to culturing the uncultured microbial majority. Nat. Rev. Microbiol. *19*, 225–240. 10.1038/s41579-020-00458-8.

131. Cowan, D., Meyer, Q., Stafford, W., Muyanga, S., Cameron, R., and Wittwer, P. (2005). Metagenomic gene discovery: past, present and future. Trends Biotechnol. *23*, 321–329. 10.1016/j.tibtech.2005.04.001.

132. Taş, N., de Jong, A.E., Li, Y., Trubl, G., Xue, Y., and Dove, N.C. (2021). Metagenomic tools in microbial ecology research. Curr. Opin. Biotechnol. *67*, 184–191. 10.1016/j.copbio.2021.01.019.

133. Thomas, T., Gilbert, J., and Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. Microb. Inform. Exp. *2*, 3. 10.1186/2042-5783-2-3.

134. Charalampous, T., Kay, G.L., Richardson, H., Aydin, A., Baldan, R., Jeanes, C., Rae, D., Grundy, S., Turner, D.J., Wain, J., et al. (2019). Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. Nat. Biotechnol. *37*, 783–792. 10.1038/s41587-019-0156-5.

135. Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. Nat. Biotechnol. *35*, 833–844. 10.1038/nbt.3935.

136. Rodriguez-Brito, B., Rohwer, F., and Edwards, R. a. (2006). An application of statistics to comparative metagenomics. BMC Bioinformatics *7*, 162. 10.1186/1471-2105-7-162.

137. Li, H. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. Annu. Rev. Stat. Appl. *2*, 73–94. 10.1146/annurev-statistics-010814-020351.

138. Li, H., and Li, H. (2021). Introduction to special issue on statistics in microbiome and

metagenomics. Stat. Biosci. *13*, 197–199. 10.1007/s12561-021-09307-5.

139. Calle, M.L. (2019). Statistical analysis of metagenomics data. Genomics Inform. *17*, e6. 10.5808/GI.2019.17.1.e6.

140. Mathieu, A., Leclercq, M., Sanabria, M., Perin, O., and Droit, A. (2022). Machine learning and deep learning applications in metagenomic taxonomy and functional annotation. Front. Microbiol. *13*, 811495. 10.3389/fmicb.2022.811495.

141. Van Camp, P.-J., Prasath, V.B.S., Haslam, D.B., and Porollo, A. (2023). MGS2AMR: a gene-centric mining of metagenomic sequencing data for pathogens and their antimicrobial resistance profile. Microbiome *11*, 223. 10.1186/s40168-023-01674-z.

142. Casimiro-Soriguer, C.S., Loucera, C., Peña-Chilet, M., and Dopazo, J. (2022). Towards a metagenomics machine learning interpretable model for understanding the transition from adenoma to colorectal cancer. Sci. Rep. *12*, 450. 10.1038/s41598-021-04182-y.

143. Langmead, B., and Nellore, A. (2018). Cloud computing for genomic data analysis and collaboration. Nat. Rev. Genet. *19*, 325. 10.1038/nrg.2018.8.

144. Wang, L., Ding, R., He, S., Wang, Q., and Zhou, Y. (2023). A pipeline for constructing reference genomes for large cohort-specific metagenome compression. Microorganisms *11*, 2560. 10.3390/microorganisms11102560.

145. Abubucker, S., Segata, N., Goll, J., Schubert, A.M., Izard, J., Cantarel, B.L., Rodriguez-Mueller, B., Zucker, J., Thiagarajan, M., Henrissat, B., et al. (2012). Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput. Biol. *8*, e1002358. 10.1371/journal.pcbi.1002358.

146. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. *44*, D457-62. 10.1093/nar/gkv1070.

147. UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. *47*, D506–D515. 10.1093/nar/gky1049.

148. Liu, B., and Pop, M. (2008). ARDB—Antibiotic Resistance Genes Database. Nucleic Acids Res. *37*, D443–D447. 10.1093/nar/gkn656.

149. Zhu, L., Lian, Y., Lin, D., Huang, D., Yao, Y., Ju, F., and Wang, M. (2022). Insights into microbial contamination in multi-type manure-amended soils: The profile of human bacterial pathogens, virulence factor genes and antibiotic resistance genes. J. Hazard. Mater. *437*, 129356. 10.1016/j.jhazmat.2022.129356.

150. Liu, B., Zheng, D., Zhou, S., Chen, L., and Yang, J. (2022). VFDB 2022: a general classification scheme for bacterial virulence factors. Nucleic Acids Res. *50*, D912–D917. 10.1093/nar/gkab1107.

151. Röst, H.L., Sachsenberg, T., Aiche, S., Bielow, C., Weisser, H., Aicheler, F., Andreotti, S., Ehrlich, H.-C., Gutenbrunner, P., Kenar, E., et al. (2016). OpenMS: a flexible open-source software platform for mass spectrometry data analysis. Nat. Methods *13*, 741–748. 10.1038/nmeth.3959.

152. Keegan, K.P., Glass, E.M., and Meyer, F. (2016). MG-RAST, a metagenomics service for analysis of microbial community structure and function. Methods Mol. Biol. *1399*, 207–233. 10.1007/978-1-4939-3369-3_13.

153. Gonzalez, A., Navas-Molina, J.A., Kosciolek, T., McDonald, D., Vázquez-Baeza, Y., Ackermann, G., DeReus, J., Janssen, S., Swafford, A.D., Orchanian, S.B., et al. (2018). Qiita: rapid, web-enabled microbiome meta-analysis. Nat. Methods *15*, 796–798. 10.1038/s41592-018-0141-9.

154. Richardson, L., Allen, B., Baldi, G., Beracochea, M., Bileschi, M.L., Burdett, T., Burgin, J., Caballero-Pérez, J., Cochrane, G., Colwell, L.J., et al. (2023). MGnify: the microbiome sequence data analysis resource in 2023. Nucleic Acids Res. *51*, D753–D759. 10.1093/nar/gkac1080.

155. Alter, G.C., and Vardigan, M. (2015). Addressing Global Data Sharing Challenges. J. Empir. Res. Hum. Res. Ethics *10*, 317–323. 10.1177/1556264615591561.

156. Ten Hoopen, P., Finn, R.D., Bongo, L.A., Corre, E., Fosso, B., Meyer, F., Mitchell, A., Pelletier, E., Pesole, G., Santamaria, M., et al. (2017). The metagenomic data life-cycle: standards and best practices. Gigascience *6*, 1–11. 10.1093/gigascience/gix047.

157. Naaman, K., Grant, S., Kianersi, S., Supplee, L., Henschel, B., and Mayo-Wilson, E. (2023). Exploring enablers and barriers to implementing the Transparency and Openness Promotion Guidelines: a theory-based survey of journal editors. R Soc Open Sci *10*, 221093.

10.1098/rsos.221093.

158. Eckert, E.M., Di Cesare, A., Fontaneto, D., Berendonk, T.U., Bürgmann, H., Cytryn, E., Fatta-Kassinos, D., Franzetti, A., Larsson, D.G.J., Manaia, C.M., et al. (2020). Every fifth published metagenome is not available to science. PLoS Biol. *18*, e3000698. 10.1371/journal.pbio.3000698.

159. Ruaud, A., Pfister, N., Ley, R.E., and Youngblut, N.D. (2022). Interpreting tree ensemble machine learning models with endoR. PLoS Comput. Biol. *18*, e1010714. 10.1371/journal.pcbi.1010714.

160. Boulund, F., Berglund, F., Flach, C.-F., Bengtsson-Palme, J., Marathe, N.P., Larsson, D.G.J., and Kristiansson, E. (2017). Computational discovery and functional validation of novel fluoroquinolone resistance genes in public metagenomic data sets. BMC Genomics *18*, 682. 10.1186/s12864-017-4064-0.

161. Kaminski, J., Gibson, M.K., Franzosa, E.A., Segata, N., Dantas, G., and Huttenhower, C. (2015). High-Specificity Targeted Functional Profiling in Microbial Communities with ShortBRED. PLoS Comput. Biol. *11*, e1004557. 10.1371/journal.pcbi.1004557.

162. Petersen, T.N., Lukjancenko, O., Thomsen, M.C.F., Maddalena Sperotto, M., Lund, O., Møller Aarestrup, F., and Sicheritz-Pontén, T. (2017). Correction: MGmapper: Reference based mapping and taxonomy annotation of metagenomics sequence reads. PLoS One *12*, e0179778. 10.1371/journal.pone.0179778.

163. Segata, N., Boernigen, D., Tickle, T.L., Morgan, X.C., Garrett, W.S., and Huttenhower, C. (2013). Computational meta'omics for microbial community studies. Mol. Syst. Biol. *9*, 666. 10.1038/msb.2013.22.

164. Wang, Q., Fish, J.A., Gilman, M., Sun, Y., Brown, C.T., Tiedje, J.M., and Cole, J.R. (2015). Xander: employing a novel method for efficient gene-targeted metagenomic assembly. Microbiome *3*, 32. 10.1186/s40168-015-0093-6.

165. Li, D., Huang, Y., Leung, C.-M., Luo, R., Ting, H.-F., and Lam, T.-W. (2017). MegaGTA: a sensitive and accurate metagenomic gene-targeted assembler using iterative de Bruijn graphs. BMC Bioinformatics *18*, 408. 10.1186/s12859-017-1825-3.

166. Mende, D.R., Waller, A.S., Sunagawa, S., Järvelin, A.I., Chan, M.M., Arumugam, M., Raes, J., and Bork, P. (2012). Assessment of metagenomic assembly using simulated next generation

sequencing data. PLoS One *7*, e31386. 10.1371/journal.pone.0031386.

167. Kassen, R., and Rainey, P.B. (2004). The ecology and genetics of microbial diversity. Annu. Rev. Microbiol. *58*, 207–231. 10.1146/annurev.micro.58.030603.123654.

168. Beatson, S.A., Minamino, T., and Pallen, M.J. (2006). Variation in bacterial flagellins: from sequence to structure. Trends Microbiol. *14*, 151–155. 10.1016/j.tim.2006.02.008.

169. Paul, C.J., Twine, S.M., Tam, K.J., Mullen, J.A., Kelly, J.F., Austin, J.W., and Logan, S.M. (2007). Flagellin diversity in Clostridium botulinum groups I and II: a new strategy for strain identification. Appl. Environ. Microbiol. *73*, 2963–2975. 10.1128/AEM.02623-06.

170. Karpiński, P., Wultańska, D., Piotrowski, M., Brajerova, M., Mikucka, A., Pituch, H., and Krutova, M. (2022). Motility and the genotype diversity of the flagellin genes fliC and fliD among Clostridioides difficile ribotypes. Anaerobe *73*, 102476. 10.1016/j.anaerobe.2021.102476.

171. Wood, D.E., and Salzberg, S.L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. *15*, R46. 10.1186/gb-2014-15-3-r46.

172. Lu, J., Breitwieser, F.P., Thielen, P., and Salzberg, S.L. (2017). Bracken: estimating species abundance in metagenomics data. PeerJ Comput. Sci. *3*, e104. 10.7717/peerj-cs.104.

173. Murrell, B., Wertheim, J.O., Moola, S., Weighill, T., Scheffler, K., and Kosakovsky Pond, S.L. (2012). Detecting individual sites subject to episodic diversifying selection. PLoS Genet. *8*, e1002764. 10.1371/journal.pgen.1002764.

174. Hage, H., Couillaud, J., Salamov, A., Loussouarn-Yvon, M., Durbesson, F., Ormeño, E., Grisel, S., Duquesne, K., Vincentelli, R., Grigoriev, I., et al. (2023). An HMM approach expands the landscape of sesquiterpene cyclases across the kingdom Fungi. Microb Genom *9*. 10.1099/mgen.0.000990.

175. Hacquard, S., Garrido-Oter, R., González, A., Spaepen, S., Ackermann, G., Lebeis, S., McHardy, A.C., Dangl, J.L., Knight, R., Ley, R., et al. (2015). Microbiota and Host Nutrition across Plant and Animal Kingdoms. Cell Host Microbe *17*, 603–616. 10.1016/j.chom.2015.04.009.

176. Colombo, B.M., Scalvenzi, T., Benlamara, S., and Pollet, N. (2015). Microbiota and mucosal immunity in amphibians. Front. Immunol. *6*, 111. 10.3389/fimmu.2015.00111.

177. Kohl, K.D., Cary, T.L., Karasov, W.H., and Dearing, M.D. (2013). Restructuring of the amphibian gut microbiota through metamorphosis. Environ. Microbiol. Rep. *5*, 899–903. 10.1111/1758-2229.12092.

178. Fedewa, L.A. (2006). Fluctuating Gram-Negative Microflora in Developing Anurans. hpet *40*, 131–135. 10.1670/104-04N.1.

179. Rojas, C.A., Ramírez-Barahona, S., Holekamp, K.E., and Theis, K.R. (2021). Host phylogeny and host ecology structure the mammalian gut microbiota at different taxonomic scales. Anim Microbiome *3*, 33. 10.1186/s42523-021-00094-4.

180. Hale, V.L., Tan, C.L., Niu, K., Yang, Y., Knight, R., Zhang, Q., Cui, D., and Amato, K.R. (2017). Diet Versus Phylogeny: a Comparison of Gut Microbiota in Captive Colobine Monkey Species. Microb. Ecol., 1–13. 10.1007/s00248-017-1041-8.

181. Amato, K.R., G Sanders, J., Song, S.J., Nute, M., Metcalf, J.L., Thompson, L.R., Morton, J.T., Amir, A., J McKenzie, V., Humphrey, G., et al. (2019). Evolutionary trends in host physiology outweigh dietary niche in structuring primate gut microbiomes. ISME J. *13*, 576–587. 10.1038/s41396-018-0175-0.

182. Quach, H., Wilson, D., Laval, G., Patin, E., Manry, J., Guibert, J., Barreiro, L.B., Nerrienet, E., Verschoor, E., Gessain, A., et al. (2013). Different selective pressures shape the evolution of toll-like receptors in human and African great ape populations. Hum. Mol. Genet. *22*, 4829–4840. 10.1093/hmg/ddt335.

183. Moeller, A.H., Caro-Quintero, A., Mjungu, D., Georgiev, A.V., Lonsdorf, E.V., Muller, M.N., Pusey, A.E., Peeters, M., Hahn, B.H., and Ochman, H. (2016). Cospeciation of gut microbiota with hominids. Science *353*, 380–382. 10.1126/science.aaf3951.

184. Muehlbauer, A.L., Richards, A.L., Alazizi, A., Burns, M.B., Gomez, A., Clayton, J.B., Petrzelkova, K., Cascardo, C., Resztak, J., Wen, X., et al. (2021). Interspecies variation in hominid gut microbiota controls host gene regulation. Cell Rep. *37*, 110057. 10.1016/j.celrep.2021.110057.

185. Qin, M., Jiang, L., Qiao, G., and Chen, J. (2023). Phylosymbiosis: The Eco-Evolutionary Pattern of Insect-Symbiont Interactions. Int. J. Mol. Sci. *24*. 10.3390/ijms242115836.

186. Meehan, C.J., and Beiko, R.G. (2014). A phylogenomic view of ecological specialization in the

lachnospiraceae, a family of digestive tract-associated bacteria. Genome Biol. Evol. *6*, 703–713. 10.1093/gbe/evu050.

187. Biddle, A., Stewart, L., Blanchard, J., and Leschine, S. (2013). Untangling the Genetic Basis of Fibrolytic Specialization by Lachnospiraceae and Ruminococcaceae in Diverse Gut Communities. Diversity *5*, 627–640. 10.3390/d5030627.

188. Ríos-Covián, D., Ruas-Madiedo, P., Margolles, A., Gueimonde, M., de Los Reyes-Gavilán, C.G., and Salazar, N. (2016). Intestinal Short Chain Fatty Acids and their Link with Diet and Human Health. Front. Microbiol. *7*, 185. 10.3389/fmicb.2016.00185.

189. Koh, A., De Vadder, F., Kovatcheva-Datchary, P., and Bäckhed, F. (2016). From dietary fiber to host physiology: Short-chain fatty acids as key bacterial metabolites. Cell *165*, 1332–1345. 10.1016/j.cell.2016.05.041.

190. Ngou, B.P.M., Wyler, M., Schmid, M.W., Kadota, Y., and Shirasu, K. (2024). Evolutionary trajectory of pattern recognition receptors in plants. Nat. Commun. *15*, 308. 10.1038/s41467-023-44408-3.

191. Riera Romo, M., Pérez-Martínez, D., and Castillo Ferrer, C. (2016). Innate immunity in vertebrates: an overview. Immunology *148*, 125–139. 10.1111/imm.12597.

192. Boehm, T. (2012). Evolution of vertebrate immunity. Curr. Biol. *22*, R722-32. 10.1016/j.cub.2012.07.003.

193. Avila Santos, A.P., Kabiru Nata'ala, M., Kasmanas, J.C., Bartholomäus, A., Keller-Costa, T., Jurburg, S.D., Tal, T., Camarinha-Silva, A., Saraiva, J.P., Ponce de Leon Ferreira de Carvalho, A.C., et al. (2023). The AnimalAssociatedMetagenomeDB reveals a bias towards livestock and developed countries and blind spots in functional-potential studies of animal-associated microbiomes. Anim Microbiome *5*, 48. 10.1186/s42523-023-00267-3.

194. Corrêa, F.B., Saraiva, J.P., Stadler, P.F., and da Rocha, U.N. (2020). TerrestrialMetagenomeDB: a public repository of curated and standardized metadata for terrestrial metagenomes. Nucleic Acids Res. *48*, D626–D632. 10.1093/nar/gkz994.

195. Nata'ala, M.K., Avila Santos, A.P., Coelho Kasmanas, J., Bartholomäus, A., Saraiva, J.P., Godinho Silva, S., Keller-Costa, T., Costa, R., Gomes, N.C.M., Ponce de Leon Ferreira de Carvalho, A.C., et al. (2022). MarineMetagenomeDB: a public repository for curated and

standardized metadata for marine metagenomes. Environ Microbiome *17*, 57. 10.1186/s40793-022-00449-7.

196. Schmidt, T.S.B., Fullam, A., Ferretti, P., Orakov, A., Maistrenko, O.M., Ruscheweyh, H.-J., Letunic, I., Duan, Y., Van Rossum, T., Sunagawa, S., et al. (2024). SPIRE: a Searchable, Planetary-scale mIcrobiome REsource. Nucleic Acids Res. *52*, D777–D783. 10.1093/nar/gkad943.

197. Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. Curr. Protoc. Hum. Genet. *Chapter 7*, Unit7.20. 10.1002/0471142905.hg0720s76.

198. Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics *26*, 680–682. 10.1093/bioinformatics/btq003.

199. Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30*, 1236–1240. 10.1093/bioinformatics/btu031.

200. Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L., et al. (2004). The Pfam protein families database. Nucleic Acids Res. *32*, D138-41. 10.1093/nar/gkh121.

201. Sherrill-Mix, S. taxonomizr: Functions to Work with NCBI Accessions and Taxonomy. https://CRAN. R-project. org/package= taxonomizr.

202. Parks, D.H., Chuvochina, M., Rinke, C., Mussig, A.J., Chaumeil, P.-A., and Hugenholtz, P. (2022). GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res. *50*, D785–D794. 10.1093/nar/gkab776.

203. de la Cuesta-Zuluaga, J., Spector, T.D., Youngblut, N.D., and Ley, R.E. (2020). Genomic insights into adaptations of TMA-utilizing methanogens to diverse habitats including the human gut. bioRxiv, 2020.09.17.302828. 10.1101/2020.09.17.302828.

204. Youngblut, N.D., Reischer, G.H., Dauser, S., Maisch, S., Walzer, C., Stalder, G., Farnleitner, A.H., and Ley, R.E. (2021). Vertebrate host phylogeny influences gut archaeal diversity. Nat

Microbiol *6*, 1443–1454. 10.1038/s41564-021-00980-2.

205. Kodama, Y., Shumway, M., Leinonen, R., and International Nucleotide Sequence Database Collaboration (2012). The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res. *40*, D54-6. 10.1093/nar/gkr854.

206. Buchfink, B., Xie, C., and Huson, D.H. (2014). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60. 10.1038/nmeth.3176.

207. Katoh, K., Misawa, K., Kuma, K.-I., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. *30*, 3059–3066. 10.1093/nar/gkf436.

208. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973. 10.1093/bioinformatics/btp348.

209. Price, M.N., Dehal, P.S., and Arkin, A.P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. *26*, 1641–1650. 10.1093/molbev/msp077.

210. Letunic, I., and Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. *49*, W293–W296. 10.1093/nar/gkab301.

211. Mailund, T. (2019). Manipulating Data Frames: dplyr. In R Data Science Quick Reference: A Pocket Guide to APIs, Libraries, and Packages, T. Mailund, ed. (Apress), pp. 109–160. 10.1007/978-1-4842-4894-2_7.

212. McMurdie, P.J., and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One *8*, e61217. 10.1371/journal.pone.0061217.

213. Boehmke, B.C. (2016). Reshaping Your Data with tidyr. In Data Wrangling with R, Boehmke and B. C., eds. (Springer International Publishing), pp. 211–218. 10.1007/978-3-319-45599-0_21.

214. Wickham, H. (2010). Stringr: Modern, consistent string processing. R J. *2*, 38. 10.32614/rj-2010-012.

215. Barnett, D., Arts, I., and Penders, J. (2021). microViz: an R package for microbiome data visualization and statistics. J. Open Source Softw. *6*, 3201. 10.21105/joss.03201.

216. Castañeda, L.E., and Barbosa, O. (2017). Metagenomic analysis exploring taxonomic and functional diversity of soil microbial communities in Chilean vineyards and surrounding native forests. PeerJ *5*. 10.7717/peerj.3098.

217. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140. 10.1093/bioinformatics/btp616.

218. Wickham, H. (2011). Ggplot2. Wiley Interdiscip. Rev. Comput. Stat. *3*, 180–185. 10.1002/wics.147.

219. Ganten, D., and Ruckpaul, K. (2006). Batch Entrez. Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine; Springer: Berlin, Germany, 131.

220. Thompson, J.D., Gibson, T.J., and Higgins, D.G. (2002). Multiple sequence alignment using ClustalW and ClustalX. Curr. Protoc. Bioinformatics *Chapter 2*, Unit 2.3. 10.1002/0471250953.bi0203s00.

221. Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. *34*, W609-12. 10.1093/nar/gkl315.

222. Kosakovsky Pond, S.L., Posada, D., Gravenor, M.B., Woelk, C.H., and Frost, S.D.W. (2006). GARD: a genetic algorithm for recombination detection. Bioinformatics *22*, 3096–3098. 10.1093/bioinformatics/btl474.

223. Weaver, S., Shank, S.D., Spielman, S.J., Li, M., Muse, S.V., and Kosakovsky Pond, S.L. (2018). Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes. Mol. Biol. Evol. *35*, 773–777. 10.1093/molbev/msx335.

224. Kosakovsky Pond, S.L., and Frost, S.D.W. (2005). Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol. Biol. Evol. *22*, 1208–1222. 10.1093/molbev/msi105.

225. Thaiss, C.A., Levy, M., Suez, J., and Elinav, E. (2014). The interplay between the innate immune

system and the microbiota. Curr. Opin. Immunol. *26*, 41–48. 10.1016/j.coi.2013.10.016.

226. Rumbo, M., Nempont, C., Kraehenbuhl, J.-P., and Sirard, J.-C. (2006). Mucosal interplay among commensal and pathogenic bacteria: lessons from flagellin and Toll-like receptor 5. FEBS Lett. *580*, 2976–2984. 10.1016/j.febslet.2006.04.036.

227. Chu, H., and Mazmanian, S.K. (2013). Innate immune recognition of the microbiota promotes host-microbial symbiosis. Nat. Immunol. *14*, 668–675. 10.1038/ni.2635.

228. Kubinak, J.L., and Round, J.L. (2012). Toll-like receptors promote mutually beneficial commensal-host interactions. PLoS Pathog. *8*, e1002785. 10.1371/journal.ppat.1002785.

229. Zhao, Q., and Elson, C.O. (2018). Adaptive immune education by gut microbiota antigens. Immunology *154*, 28–37. 10.1111/imm.12896.

230. Alexander, K.L., Zhao, Q., Reif, M., Rosenberg, A.F., Mannon, P.J., Duck, L.W., and Elson, C.O. (2021). Human Microbiota Flagellins Drive Adaptive Immune Responses in Crohn's Disease. Gastroenterology *161*, 522-535.e6. 10.1053/j.gastro.2021.03.064.

231. Vijay-Kumar, M., Sanders, C.J., Taylor, R.T., Kumar, A., Aitken, J.D., Sitaraman, S.V., Neish, A.S., Uematsu, S., Akira, S., Williams, I.R., et al. (2007). Deletion of TLR5 results in spontaneous colitis in mice. J. Clin. Invest. *117*, 3909–3921. 10.1172/JCI33084.

232. Neville, B.A., Sheridan, P.O., Harris, H.M.B., Coughlan, S., Flint, H.J., Duncan, S.H., Jeffery, I.B., Claesson, M.J., Ross, R.P., Scott, K.P., et al. (2013). Pro-Inflammatory Flagellin Proteins of Prevalent Motile Commensal Bacteria Are Variably Abundant in the Intestinal Microbiome of Elderly Humans. PLoS One *8*. 10.1371/journal.pone.0068919.

233. Lloyd-Price, J., Arze, C., Ananthakrishnan, A.N., Schirmer, M., Avila-Pacheco, J., Poon, T.W., Andrews, E., Ajami, N.J., Bonham, K.S., Brislawn, C.J., et al. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature *569*, 655–662. 10.1038/s41586-019-1237-9.

234. Franzosa, E.A., Sirota-Madi, A., Avila-Pacheco, J., Fornelos, N., Haiser, H.J., Reinker, S., Vatanen, T., Hall, A.B., Mallick, H., McIver, L.J., et al. (2019). Gut microbiome structure and metabolic activity in inflammatory bowel disease 10.1038/s41564-018-0306-4.

235. Borodovich, T., Shkoporov, A.N., Ross, R.P., and Hill, C. (2022). Phage-mediated horizontal

gene transfer and its implications for the human gut microbiome. Gastroenterol. Rep. *10*, goac012. 10.1093/gastro/goac012.

236. Voogdt, C.G.P., Bouwman, L.I., Kik, M.J.L., Wagenaar, J.A., and van Putten, J.P.M. (2016). Reptile Toll-like receptor 5 unveils adaptive evolution of bacterial flagellin recognition. Sci. Rep. *6*, 19046. 10.1038/srep19046.

237. Campbell, E.A., Walden, H., Walter, J.C., Shukla, A.K., Beck, M., Passmore, L.A., and Xu, H.E. (2024). AlphaFold: Research accelerator and hypothesis generator. Mol. Cell *84*, 404–408. 10.1016/j.molcel.2023.12.035.

238. Varadi, M., and Velankar, S. (2023). The impact of AlphaFold Protein Structure Database on the fields of life sciences. Proteomics *23*, e2200128. 10.1002/pmic.202200128.

239. Sanderson, T., Bileschi, M.L., Belanger, D., and Colwell, L.J. (2023). ProteInfer, deep neural networks for protein functional inference. Elife *12*. 10.7554/eLife.80942.

240. Kulmanov, M., and Hoehndorf, R. (2021). DeepGOPlus: improved protein function prediction from sequence. Bioinformatics *37*, 1187. 10.1093/bioinformatics/btaa763.

241. Kans, J. (2024). Entrez Direct: E-utilities on the Unix Command Line (National Center for Biotechnology Information (US)).

242. NCBI Resource Coordinators (2018). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. *46*, D8–D13. 10.1093/nar/gkx1095.

243. Keck, F., Rimet, F., Bouchez, A., and Franc, A. (2016). phylosignal: an R package to measure, test, and explore the phylogenetic signal. Ecol. Evol. *6*, 2774–2780. 10.1002/ece3.2051.

244. Kembel, S. An introduction to the picante package. http://cran.uib.no/web/packages/picante/vignettes/picante-intro.pdf.

245. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152. 10.1093/bioinformatics/bts565.

246. Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659. 10.1093/bioinformatics/btl158.

247. Asnicar, F., Thomas, A.M., Beghini, F., Mengoni, C., Manara, S., Manghi, P., Zhu, Q., Bolzan,

M., Cumbo, F., May, U., et al. (2020). Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. Nat. Commun. *11*, 2500. 10.1038/s41467-020-16366-7.

248. Revell, L.J. (2024). phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). PeerJ *12*, e16505. 10.7717/peerj.16505.

Appendices

Appendix 1

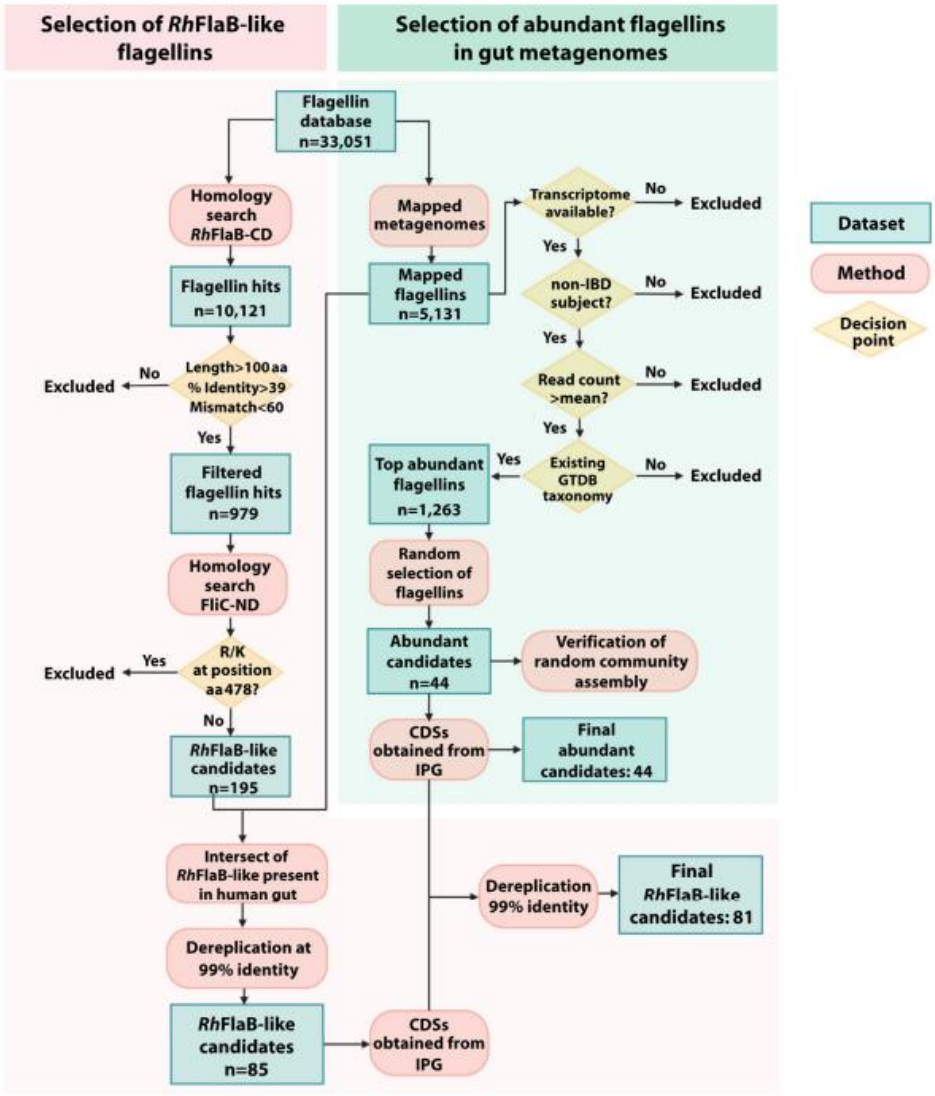

Figure S1. Selection of flagellin candidates. The flowchart illustrates the strategy used to identify abundant flagellins (n = 44) and *Rh*FlaB-like flagellins occurring in the human gut (n = 81). **Left:** *Rh*FlaB-like flagellins were identified by homology search of truncated C-terminal domain (CD) and N-terminal domain (ND) of FlaB of *Roseburia hominis* and FliC of *Salmonella Typhimurium*, respectively. Only *Rh*FlaB-like flagellins occurring in human gut metagenomes were kept. **Right:** Abundant flagellins from gut metagenomes of non-IBD subjects were randomly selected among flagellins with a read count over the median and an existing taxonomy on GTDB v95. The coding sequences (CDSs) of *Rh*FlaB-like candidates were de-replicated with the CDSs of previously identified abundant flagellins to avoid redundancy.

Appendix 2:

Supplementary Table 1:

https://docs.google.com/spreadsheets/d/19-l5GRwr_o4e5t_ervbdqyCB2MHiFdkF/edit#gid=1320004726

Appendix 3: Silent recognition of flagellins from human gut commensal bacteria by Toll-like receptor 5

INNATE IMMUNITY

Silent recognition of flagellins from human gut commensal bacteria by Toll-like receptor 5

Sara J. Clasen[1], Michael E. W. Bell[1], Andrea Borbón[1], Du-Hwa Lee[2], Zachariah M. Henseler[1], Jacobo de la Cuesta-Zuluaga[1], Katarzyna Parys[2], Jun Zou[3], Yanling Wang[3], Veronika Altmannova[4], Nicholas D. Youngblut[1], John R. Weir[4], Andrew T. Gewirtz[3], Youssef Belkhadir[2], Ruth E. Ley[1,5]*

Flagellin, the protein subunit of the bacterial flagellum, stimulates the innate immune receptor Toll-like receptor 5 (TLR5) after pattern recognition or evades TLR5 through lack of recognition. This binary response fails to explain the weak agonism of flagellins from commensal bacteria, raising the question of how TLR5 response is tuned. Here, we screened abundant flagellins present in metagenomes from human gut for both TLR5 recognition and activation and uncovered a class of flagellin-TLR5 interaction termed silent recognition. Silent flagellins were weak TLR5 agonists despite pattern recognition. Receptor activity was tuned by a TLR5-flagellin interaction distal to the site of pattern recognition that was present in *Salmonella* flagellin but absent in silent flagellins. This interaction enabled flagellin binding to preformed TLR5 dimers and increased TLR5 signaling by several orders of magnitude. Silent recognition by TLR5 occurred in human organoids and mice, and silent flagellin proteins were present in human stool. These flagellins were produced primarily by the abundant gut bacteria Lachnospiraceae and were enriched in nonindustrialized populations. Our findings provide a mechanism for the innate immune system to tolerate commensal-derived flagellins while remaining vigilant to the presence of flagellins produced by pathogens.

INTRODUCTION

Innate immune responses are initiated by pattern recognition receptors (PRRs) that evolved to detect conserved microbe-associated molecular patterns (MAMPs) (*1*). The Toll-like receptors (TLRs) are membrane-bound PRRs, widely expressed in many cell types, that activate proinflammatory pathways after MAMP binding to their horseshoe-shaped ectodomains (*2*). Because MAMPs are not unique to pathogens, a question that has persisted for decades is whether TLRs respond differently to ligands derived from beneficial or commensal microbiota relative to those produced by potentially pathogenic microbes (*3*). This question is especially relevant for TLRs that interface with the intestinal microbiota, such as TLR5, which is highly expressed by epithelial cells that line mucosal surfaces (*4*).

TLR5 is plasma membrane bound and binds extracellular flagellin, the protein subunit of the bacterial flagellum (*5*). Phylogenetically diverse bacteria produce structurally similar flagellins that consist of conserved N- and C-terminal D0-D1 domains separated by a hypervariable region (*6*). The MAMP recognized by TLR5 is located in the N-terminal D1 (nD1) and referred to as the TLR5 epitope (*7*, *8*). Studies on the FliC flagellin derived from the human pathogen *Salmonella enterica* serovar Typhimurium show that mutating key residues in this region (FliC PIM) reduces ligand potency and abolishes bacterial motility (*7*); crystal structures of FliC in complex with *Danio rerio* TLR5 confirm a direct interaction between these residues and leucine-rich repeat 9 (LRR9) in the N-terminal region of the receptor ectodomain (*9*, *10*). Furthermore, flagellins that do not stimulate TLR5, like FlaA from the human pathogen *Helicobacter pylori* ("*Hp*FlaA") have different amino acids in their TLR5 epitope site (*8*, *11*). TLR5's inability to respond to *Hp*FlaA is characterized as "evasion" and is presumed to occur through loss of TLR5 binding. Together, these studies demonstrate that robust TLR5 signaling requires the receptor ectodomain to bind the flagellin TLR5 epitope.

Commensal bacteria produce flagellins with TLR5 epitopes identical to those of FliC; however, these flagellins induce a range of TLR5 activity (*12*–*14*), raising the question of how the TLR5 response is tuned. Here, we addressed this question by investigating how TLR5 interacts with flagellins from commensal bacteria. We identified members of the Lachnospiraceae family as the major producers of flagellin in the human gut and measured TLR5 recognition and response to 40 of the most prevalent flagellins in the human gut. This approach led us to identify a type of flagellin that we termed "silent" due to the observation that these flagellins retained binding to TLR5 yet poorly activated TLR5 signaling. Last, we described how silent flagellins circumvented TLR5 stimulation, thereby providing additional insights into the mechanism of TLR5 activation.

RESULTS

Commensal bacteria produced flagellins that decoupled TLR5 binding from activation

To examine the TLR5 response to commensal-derived flagellins, we first searched for flagellins commonly encoded by the human gut microbiome. Flagellin diversity was vast: Of the 10 million proteins encoded by the human gut microbiome, more than 5000 different proteins were classified as flagellins (*15*). Most flagellin in the

[1]Department of Microbiome Science, Max Planck Institute for Biology, Tübingen 72076, Germany. [2]Gregor Mendel Institute (GMI), Austrian Academy of Sciences, Vienna BioCenter (VBC), Dr. Bohr-Gasse 3, Vienna, Austria. [3]Center for Inflammation, Immunity and Infection, Institute for Biomedical Sciences, Georgia State University, Atlanta, GA, USA. [4]Friedrich Miescher Laboratory of the Max Planck Society, Max-Planck-Ring 9, Tübingen 72076, Germany. [5]Cluster of Excellence EXC 2124 Controlling Microbes to Fight Infections, University of Tübingen, Tübingen, Germany.
*Corresponding author. Email: ruth.ley@tuebingen.mpg.de

human gut is produced by Lachnospiraceae (*16*), a prevalent and abundant family of Firmicutes that includes beneficial bacteria such as the butyrate producers of the *Roseburia* and *Eubacterium* genera (*17*). Selecting from the most abundant flagellins observed in 270 samples from individuals without inflammatory bowel disease (IBD) (*18*) (fig. S1), we expressed an initial 40 recombinant flagellins (34 belonging to Lachnospiraceae species) and screened these for both TLR5 signaling and recognition (figs. S2 and S3).

To quantify TLR5 recognition of flagellin, we measured the relative binding strength between the receptor and the TLR5 epitope (Fig. 1A, magenta region) using a truncated form of the human ectodomain, TLR5^{N14} [similar to the one used in the crystal structure complex in (*9*)] (Fig. 1B). This construct contained the first 14 of the 22 LRRs that comprise the ectodomain, including the flagellin binding site identified in the crystal structure (LRR9), followed by a C-terminal adaptor sequence tagged to immunoglobulin G (IgG)–Fc. Binding was quantified by incubating TLR5^{N14} with flagellins expressing C-terminal alkaline phosphatase (AP) and measuring AP activity. Consistent with its TLR5 epitope directly interacting with TLR5^{N14} (*9*), FliC bound strongly (Fig. 1C). In contrast, FliC PIM, which lacks three conserved residues in the epitope, showed a ninefold reduction in binding compared with FliC. Similarly, *Hp*FlaA, with its altered TLR5 epitope, failed to bind TLR5^{N14}. The D0 domain of FliC is unnecessary for binding to TLR5 (*9*, *19*), and we observed strong binding of FliC ΔD0 to TLR5^{N14} (Fig. 1, A and C). Together, these results showed that the flagellin-TLR5^{N14} interaction required the TLR5 epitope and thus reflected pattern recognition by TLR5 (*11*).

In our screen, most of the 40 flagellins from commensals have TLR5 epitopes whose key residues either are identical to those of FliC (21 of 40) or differ at only one position (16 of 40) (fig. S4A). In addition to flagellins from commensals, we included three flagellins from pathogens (FliC, *Vibrio cholerae* FlaB, and *Listeria monocytogenes* FlaA) and two negative controls (*Hp*FlaA and FliC PIM). We generated AP-tagged flagellins to assay TLR5^{N14} binding (fig. S2) and separately expressed N-terminal Myc-tagged flagellins to

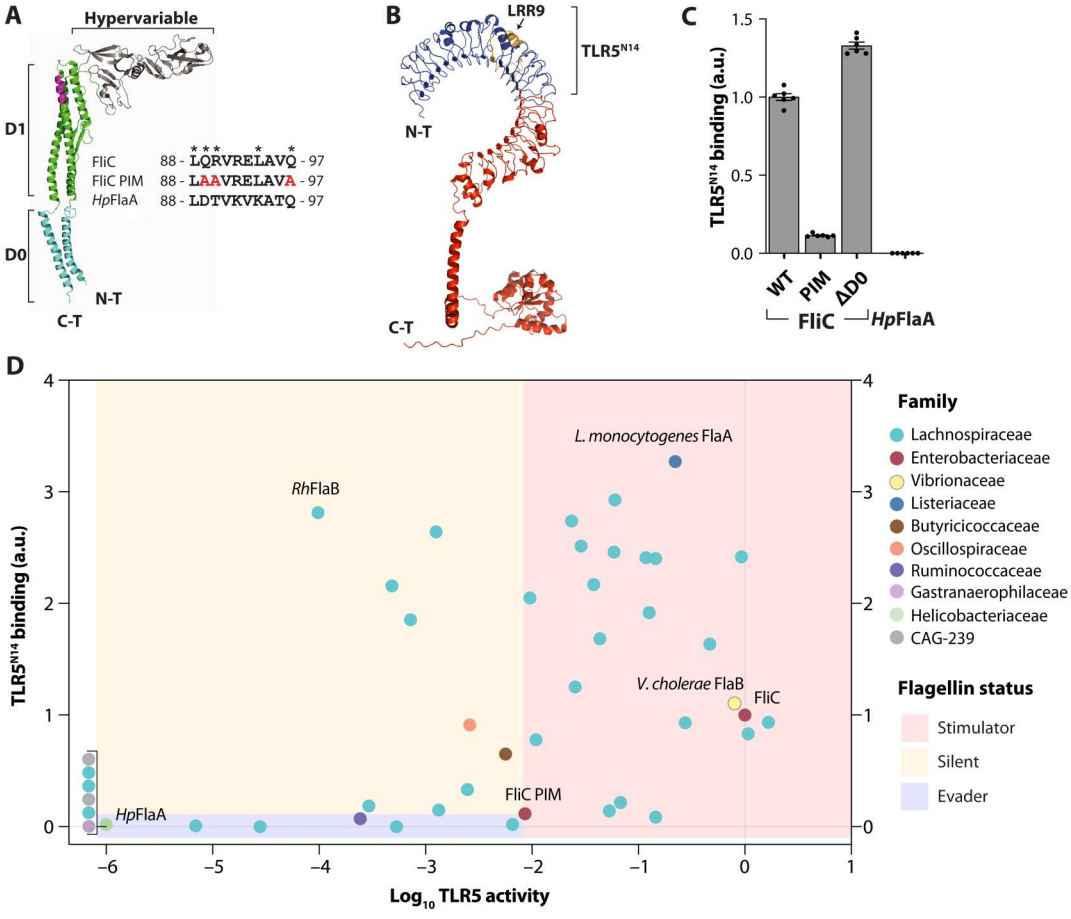

Fig. 1. Flagellins from human gut commensals were silently recognized by TLR5. (**A**) FliC crystal structure [PDB 3A5X from (*6*)] and multiple sequence alignment of nD1 TLR5 epitope (magenta) from *Salmonella* and *H. pylori* flagellins. Asterisks denote residues in FliC required for TLR5 recognition; red: residues mutated in FliC PIM. (**B**) Predicted structure of human TLR5 (*10*). Blue: Region present in TLR5^{N14}. Yellow: TLR5 epitope binding site, LRR9. (**C**) Flagellin binding to TLR5 ectodomain: absorbance units (a.u.) are relative to FliC. Error bars are SEM for $n = 6$. (**D**) TLR5 activity versus TLR5^{N14} binding for 40 flagellins abundant in gut microbiome (see fig. S1) and from pathogens. Circles represent individual flagellins; colors indicate family-level taxonomy (GTDB). TLR5^{N14} binding described in (B); data represent mean for $n \geq 2$. TLR5 activity represents negative EC_{50} normalized to flagellin expression in lysates. Data represent mean from at least three independent experiments; values normalized to FliC. Flagellin status is defined relative to FliC PIM: Stimulators are more active (EC_{50} < 10 nM), silent flagellins are less active (EC_{50} > 10 nM) with increased binding to TLR5^{N14}, and evaders are less active (EC_{50} > 10 nM) with reduced binding to TLR5^{N14}.

quantify TLR5 activation (fig. S3). Flagellins were incubated with nuclear factor κB (NF-κB) reporter human embryonic kidney (HEK) cells engineered to express TLR5, and NF-κB–dependent AP activity was measured as a readout for TLR5 activation.

Consistent with the notion that binding TLR5 leads to its activation, we generally observed a positive correlation between $TLR5^{N14}$ binding and TLR5 activity (Fig. 1D). Flagellins that induced a greater response than that of FliC PIM [corresponding roughly to a median effective concentration (EC_{50}) of <10 nM] were categorized as "stimulators" (red region in Fig. 1D) regardless of their ability to bind $TLR5^{N14}$; this described nearly half the flagellins in our screen. "Evaders," in contrast, bound and stimulated more weakly than FliC PIM (blue region). This group included *Hp*FlaA and 11 commensal-derived flagellins. A TLR5 activity score of ≤−3 corresponded broadly to an EC_{50} of >100 nM; these weak agonists showed high variability in our activity assay (fig. S4B). The remaining flagellins (9 of 40) resembled evaders with respect to TLR5 activation (stimulated worse than FliC PIM) but acted like stimulators with regard to $TLR5^{N14}$ binding (stronger interaction compared with FliC PIM; yellow region). We termed these unexpected ligands silent flagellins in reference to their inability to induce signaling despite intact TLR5 recognition.

Silent flagellins lacked an allosteric activator in D0 domain

We further investigated how silent flagellins decoupled TLR5 ectodomain binding from agonism. We selected FlaB from *Roseburia hominis* (*Rh*FlaB) as our representative silent flagellin because it bound $TLR5^{N14}$ the strongest among the silent flagellins from our initial screen (Fig. 1D). *R. hominis* is of wide interest because it is a common gut commensal species belonging to the Lachnospiraceae and generally thought to be anti-inflammatory and thus beneficial to host health (*20*). *R. hominis* is motile and expresses *Rh*FlaB in vivo (*20*, *21*). We purified recombinant *Rh*FlaB and observed that, in addition to binding $TLR5^{N14}$, it also bound full-length human TLR5 tagged with hemagglutinin (HA) (Fig. 2A, lane 5). FliC bound more full-length TLR5 compared with *Rh*FlaB, whereas the evader *Hp*FlaA showed an equal lack of binding to both truncated and full-length TLR5 (Fig. 2A, lane 4). We also validated that *Rh*FlaB was a weaker TLR5 agonist than FliC PIM, despite its intact TLR5 epitope (Fig. 2B and table S1).

Next, we tested whether *Rh*FlaB bound TLR5 through its TLR5 epitope. We constructed the flagellin *Rh*FlaB PIM, which carries the same mutations as FliC PIM that resulted in loss of binding to $TLR5^{N14}$. *Rh*FlaB PIM failed to bind the full-length receptor, consistent with TLR5 binding requiring the TLR5 epitope (Fig. 2C, lanes 5 and 6). However, unlike *Rh*FlaB PIM, FliC PIM showed no reduction in binding to full-length TLR5 (Fig. 2C, lanes 3 and 4). This result was unexpected, because FliC PIM did not bind $TLR5^{N14}$ (Fig. 1C) (*9*).

We hypothesized that FliC PIM bound the C-terminal LRRs of the TLR5 ectodomain at a location allosteric to the site of pattern recognition. Although the structure of this region of TLR5 is unsolved, the C-terminal LRRs are predicted to interact with the conserved D0 domain of flagellin (Fig. 1, A and B) (*22*). The D0 domain was not required for binding $TLR5^{N14}$ (Fig. 1C) and is also absent in the FliC-TLR5 crystal structure (*9*). Several studies previously reported the necessity of the FliC D0 for TLR5 activation (*9*, *19*). However, the mechanism is unclear and the authors unequivocally conclude that the D0 domain does not directly bind the receptor.

We tested for a TLR5 binding site in the FliC D0 by assessing the binding of FliC PIM ΔD0 to full-length TLR5. Because FliC PIM did not bind $TLR5^{N14}$, if the D0 bound TLR5 LRRs 15 to 22, then FliC PIM ΔD0 should have been unable to bind full-length TLR5. Consistent with an additional binding site in the D0 of FliC, FliC PIM ΔD0 showed a substantial loss of binding to TLR5 compared with FliC PIM and FliC ΔD0 (Fig. 2C, lane 8 versus lanes 4 and 7). Given our observation that the TLR5 epitope was necessary for *Rh*FlaB to bind the receptor, such that *Rh*FlaB PIM could not bind full-length TLR5, we predicted that the FliC D0 would restore TLR5 binding to *Rh*FlaB PIM. As expected, swapping FliC D0 for the native *Rh*FlaB D0 rescued *Rh*FlaB PIM binding (Fig. 2C, lanes 6 and 9). Together, these results suggested that FliC D0 allosterically bound TLR5, in direct contradiction to previous findings (*9*, *19*). The additional binding site also explained why FliC pulled down more full-length TLR5 than *Rh*FlaB (Fig. 2A).

The discovery of an allosteric TLR5 binding site in FliC prompted us to test its impact on TLR5 activation. We purified recombinant *Rh*FlaB chimera expressing the FliC D0 (*Rh*FlaB-FliC D0) and assayed TLR5 signaling. As expected from its greater ability to bind full-length TLR5, the chimeric flagellin was 100-fold more stimulatory than *Rh*FlaB with its native D0 (Fig. 2D). We confirmed the NF-κB reporter response by directly measuring phosphorylation of the NF-κB subunit p65 and inhibitor IkBα (Fig. 2E). After incubation with FliC or *Rh*FlaB-FliC D0, cells showed increased phosphorylation of p65 and IkBα, accompanied by reduced total IkBα levels, relative to untreated cells (lanes 3 and 5 versus lane 1). This indicates activation of the NF-κB pathway (*23*). Although cells treated with the positive control tumor necrosis factor–α (TNF-α) (lane 2) showed a modest increase in IkBα phosphorylation and degradation, incubation with *Rh*FlaB did not induce phosphorylation or degradation of IkBα (lane 4). Combined, these results showed that the FliC D0 activated *Rh*FlaB-dependent signaling of TLR5.

D0 binding site targeted flagellin to preformed TLR5 dimers

We hypothesized that the additional TLR5 binding site in the FliC D0 increased activity in part by enabling *Rh*FlaB-FliC D0 to interact with more TLR5 receptors than *Rh*FlaB. TLR5 activation requires the formation of a symmetric 2:2 flagellin:TLR5 complex (*9*, *24*). How this complex is assembled remains unclear, although it is widely stated that flagellin binding induces TLR5 dimerization (*19*, *25*–*27*). Early cryo–electron microscopy work revealed, however, that human TLR5 forms asymmetric homodimers in the absence of flagellin, a conformation likely associated with multiple ligand binding sites and thus a possible target of the FliC D0 domain (*28*). We investigated whether TLR5 formed unliganded dimers by briefly treating TLR5-HA HEK cells with the membrane-impermeable crosslinker BS^3 (Fig. 2F). In addition to monomeric TLR5, we detected a higher molecular weight species consistent with the size of a TLR5 dimer. This result suggested that preformed TLR5 dimers were present on the cell surface in the absence of ligand-induced dimerization that is commonly invoked for TLR5.

We tested the ability of FliC and *Rh*FlaB to interact with TLR5 dimers. We observed that FliC bound both monomer and dimer, whereas *Rh*FlaB only interacted with the monomer (Fig. 2F, lanes 3 and 4). The switching out of its native D0 for FliC D0 endowed

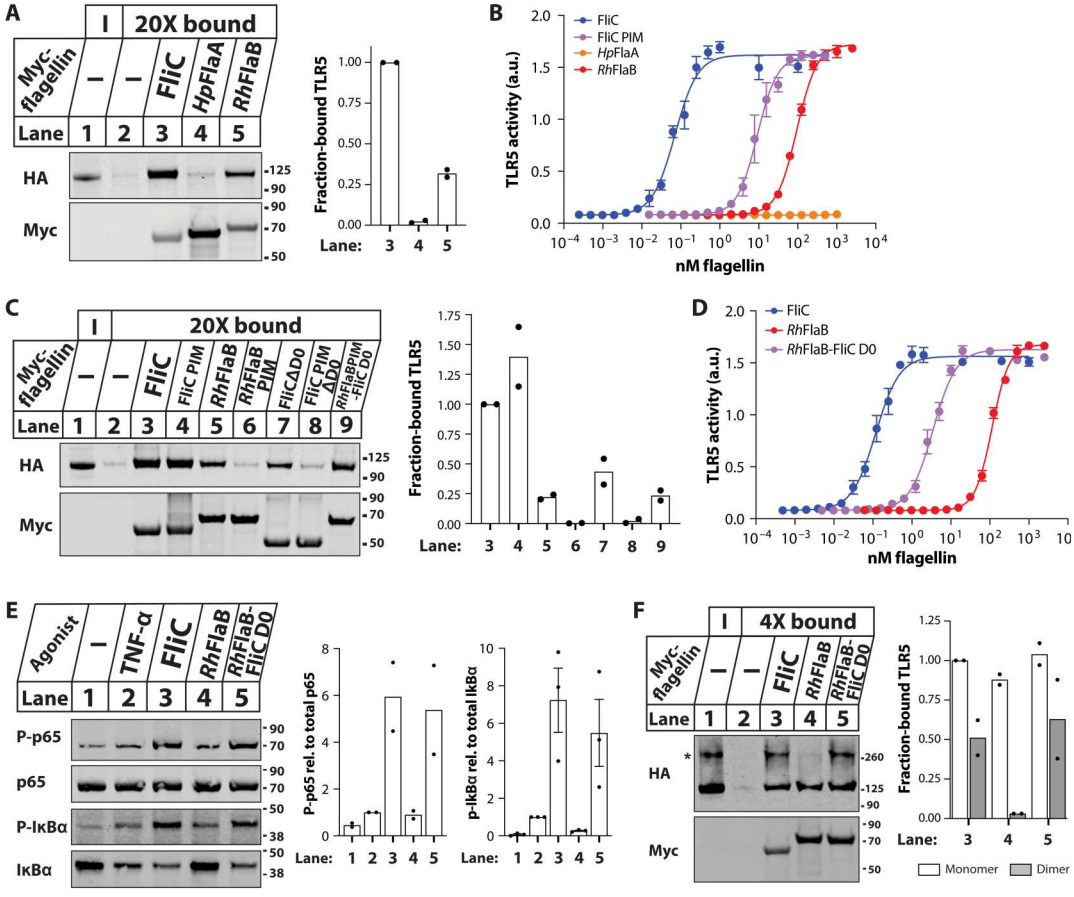

Fig. 2. Silent flagellin *Rh*FlaB lacks TLR5 binding site in D0. (**A**) Flagellin binding to full-length TLR5: TLR5-HA HEK lysates were incubated with 6xHis-Myc–tagged flagellins and purified on poly-HIS binding beads. Input ("I") and bound fractions ("20X bound") were analyzed by immunoblot. Left: representative blots from one of two independent experiments; right: densitometry quantification of HA signal in bound lanes relative to Myc signal, normalized to lane 3. (**B**) *Rh*FlaB-dependent TLR5 activity: TLR5 HEK-Blue cells were incubated with purified recombinant flagellins, and NF-κB–dependent AP levels were quantified. Error bars are SEM for $n = 3$. (**C**) Mapping TLR5 binding sites in flagellin, as in (A). (**D**) *Rh*FlaB chimera-dependent activation of TLR5, as in (B). (**E**) Activation of NF-κB pathway: TLR5 HEK-Blue cells were incubated with flagellins or TNF-α in the presence of MG132. Phosphorylation of p65 and IkBα in cell lysates was analyzed by immunoblot. Left: representative blots; right: quantification of phosphorylated signal relative to total protein signal, normalized to lane 2. (**F**) Flagellin binding to preformed TLR5 complexes: BS^3-crosslinked cells processed as in (A). Asterisk indicates the TLR5 dimer band. Quantification normalized to lane 3 monomer.

*Rh*FlaB the ability to bind the dimer (Fig. 2F, lane 5). This result suggested that the FliC D0 directly bound the ectodomain and activated TLR5 signaling in part by mediating binding to preformed TLR5 dimers. This observation supported our hypothesis that an allosteric binding site in its D0 enabled FliC to interact with more receptors than *Rh*FlaB.

The allosteric binding site was located exclusively in the conserved C-terminal D0 domain

Given the high conservation of the flagellin D0 domains, we next asked which domain (nD0 or cD0) harbored the allosteric binding site. We recombinantly expressed and purified *Rh*FlaB chimeras containing either the FliC nD0 or cD0 and tested their ability to stimulate TLR5. The FliC nD0 did not affect TLR5 signaling by *Rh*FlaB, whereas *Rh*FlaB-FliC cD0 phenocopied *Rh*FlaB-FliC D0 (Fig. 3A). The FliC cD0, therefore, mediated activation of TLR5. We further confirmed that the FliC cD0, but not the nD0, rescued *Rh*FlaB PIM binding to TLR5 (Fig. 3B, lane 6).

The *Rh*FlaB and FliC cD0s were 46% identical across 39 amino acids (Fig. 3C). Residues in the FliC cD0 are important for TLR5 activation (*19*), several of which were not conserved in *Rh*FlaB. We hypothesized that these amino acids mediate binding to the dimer and tested whether *Rh*FlaB mutants that express the equivalent residues in FliC activated TLR5 (Fig. 3D). However, of the three *Rh*FlaB mutants that we tested, only *Rh*FlaB H478R showed increased stimulation of TLR5. We concluded that the allosteric binding site likely encompasses multiple residues distributed across the cD0 domain.

Silent flagellins were prevalent in the human gut and enriched in nonindustrialized populations

Next, we assessed how widespread silent flagellins were in the human microbiome. We searched for peptide sequences of silent flagellins using a published database comprising more than 33,000 flagellins (Materials and Methods and fig. S1) (*29*). Candidate silent flagellins were selected on the basis of their presence in human gut metagenomes and by similarity to the C-terminal region

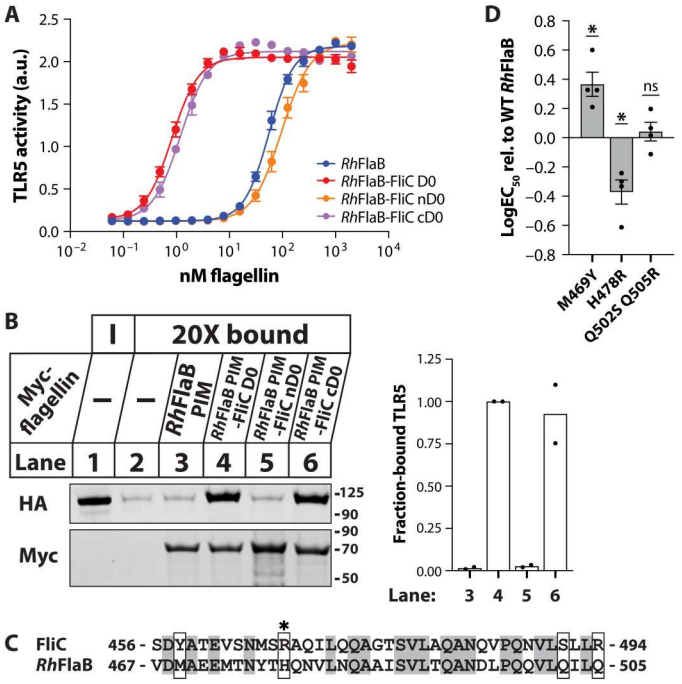

Fig. 3. FliC cD0 domain mediated binding to and activation of TLR5. (**A**) Effect of FliC nD0 and cD0 domains on *Rh*FlaB-dependent TLR5 activity, as in Fig. 2B. Data are means ± SEM for $n = 3$. (**B**) Effect of FliC nD0 and cD0 domains on *Rh*FlaB PIM binding to TLR5, as in Fig. 2A. Left: representative blots from one of two independent experiments; right: quantification of HA signal relative to Myc signal, normalized to lane 4. (**C**) Alignment of FliC and *Rh*FlaB cD0 domain amino acid sequences. Identical residues are shaded gray, and boxes indicate mutants tested in (D). Asterisk denotes residue that increased *Rh*FlaB-dependent TLR5 activation. (**D**) TLR5 activation by *Rh*FlaB cD0 mutants, as in (A). Data shown are differences in $\log EC_{50}$ of mutants relative to WT *Rh*FlaB, with EC_{50} determined by weighted, nonlinear regression analysis with Hill slope constrained to 1. Significance in mean differences between *Rh*FlaB WT and mutants was calculated by one-sample *t* test against theoretical mean equal to 0 (*$P < 0.05$; ns, not significant).

of *Rh*FlaB (Fig. 3 and fig. S1). The list was further curated to exclude flagellins containing a basic residue (R/K) at position *Rh*FlaB amino acid 478 based on our observation that *Rh*FlaB H478R showed a slight, but significant, increase in TLR5 stimulation (Fig. 3D). The final candidate silent flagellins were mostly, but not exclusively, from species belonging to the Lachnospiraceae family (74 of 76) (fig. S5).

To verify whether these 76 candidate silent flagellins were silent, we expressed them recombinantly to screen for both TLR5 signaling and $TLR5^{N14}$ binding (Fig. 4A). Compared with our initial screen (Fig. 1D), we were successful in enriching for silent flagellins: More than half (44 of 76) were weaker TLR5 agonists and stronger $TLR5^{N14}$ binders relative to FliC PIM (Fig. 4A, yellow region). Given its importance in Crohn's disease, we additionally tested the flagellin CBir1, a known weak TLR5 agonist (*13*, *30*). CBir1 bound $TLR5^{N14}$, categorizing it as a silent flagellin. The remaining 34 candidates were equally distributed among stimulators (red region) and evaders (blue region).

To assess whether the mechanism was the same for *Rh*FlaB as for the above set of silent flagellins, we examined the impact of swapping in the FliC D0 domain on a subset representing a broad range of $TLR5^{N14}$ binding strengths. Although the magnitude differed among candidates, the FliC D0 universally increased TLR5 signaling for all silent flagellins tested, including CBir1 (Fig. 4B). Moreover, these silent flagellins belonged to common taxa of the human gut microbiome, including multiple species of *Roseburia* (*17*). The FliC D0 did not affect TLR5 evasion by *Hp*FlaA, consistent with previous observations (*19*).

Given that flagellin is facultatively expressed and that expression in the gut can vary depending on external factors (*12*), we assessed the presence of silent flagellins directly from healthy human stool. Endogenous flagellins were isolated using TLR5 as bait and identified by mass spectrometry (MS). Peptides were searched against a custom flagellin database built from metagenome sequences generated from the same stool sample. Of the 12 flagellins identified, 10 were ascribed to Lachnospiraceae (table S2 and data file S1). This was consistent with the taxonomic affiliation of the abundantly expressed flagellins in humans without IBD (fig. S6, A and B) (*18*). We recombinantly expressed and purified the top two candidates to assay TLR5 signaling. Both flagellins weakly activated TLR5 with EC_{50} values greater than 100 nM (Fig. 4C and table S1). However, similar to *Rh*FlaB, swapping in the FliC D0 for the native D0 profoundly increased their ability to stimulate TLR5.

We further verified that silent recognition occurred when TLR5 was endogenously expressed in human organoids and in mice. In three-dimensionally cultured human colon organoids, FliC stimulated the secretion of interleukin-8 (IL-8) (Fig. 4D), a proinflammatory cytokine produced downstream from TLR5 activation (*31*). IL-8 levels in *Rh*FlaB-treated organoids were similar to those of *Hp*FlaA- and buffer-treated controls, whereas organoids incubated with *Rh*FlaB-FliC D0 had significantly higher IL-8 levels (Fig. 4D). We then characterized the response to silent flagellin *Rh*FlaB by wild-type (WT) and $Tlr5^{-/-}Nlrc4^{-/-}$ C57BL/6 mice. NLRC4 is an intracellular sensor of flagellin (*32*, *33*), and mice lacking both Tlr5 and Nlrc4 failed to respond to flagellin (*34*). WT mice injected with *Rh*FlaB had lower levels of the epithelial proinflammatory cytokine Cxcl1 compared with animals injected with FliC and *Rh*FlaB-FliC D0 (Fig. 4E). We saw a similar trend for IL-6, an inflammatory cytokine produced by myeloid cells (Fig. 4F). Consistent with these cytokine responses being flagellin-dependent, Cxcl1 and IL-6 levels were not elevated in $Tlr5^{-/-}Nlrc4^{-/-}$ double-knockout mice after injection. IL-18 levels were similar between WT animals injected with FliC and *Rh*FlaB (fig. S7). Because IL-18 production requires NLRC4, but not TLR5 (*34*), this suggested that silent flagellins induced an inflammatory response when detected intracellularly. Together, these results indicated that TLR5's silent recognition of flagellin was not species specific and that the FliC D0 activated both human and mouse endogenously expressed TLR5.

Our identification of an allosteric activator of TLR5 in FliC suggested a mechanism by which this receptor can respond to minute levels of stimulatory flagellin. Commensal members of the gut microbiome, such as members of Lachnospiraceae, can produce an array of flagellins that are silent, stimulatory, or evasive. *R. hominis* itself encodes several flagellins that fell into all three categories based on our binding and activation criteria (Fig. 5A and tables S1 and S3) (*20*). This within-species flagellin diversity reflected the flagellin diversity encoded broadly in human gut metagenomes, where all three types were detected (Fig. 5B and fig. S8). We observed that metagenomes from nonindustrialized populations encoded a greater proportion of all three flagellin types

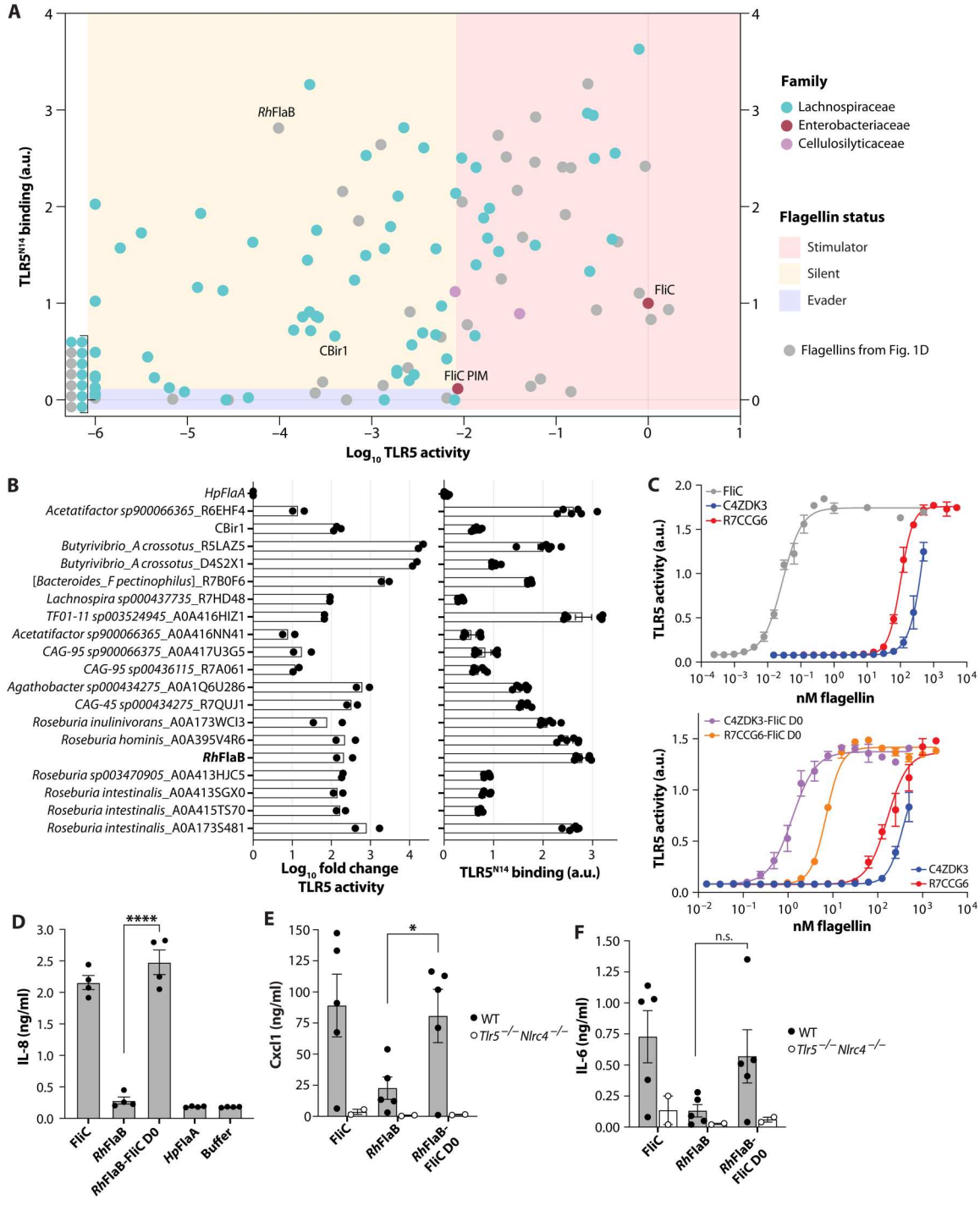

Fig. 4. Silent flagellins were widespread among Lachnospiraceae that colonize the human gut. (**A**) TLR5 activity versus TLR5^{N14} binding for *Rh*FlaB-like flagellins, as in Fig. 1D. Data are mean for $n \geq 2$; values relative to FliC. See fig. S5. (**B**) Native D0s swapped for FliC D0 in subset of Lachnospiraceae silent flagellins: TLR5 activity measured as in (A). Bar graph shows mean difference between WT and chimeric flagellins from at least two independent experiments. TLR5^{N14} binding on right. (**C**) Top: TLR5 activation by flagellins from human stool as in Fig. 2B. Bottom: Effect of FliC D0 on TLR5 activity. Data are means ± SEM for $n = 3$. See tables S1 and S2 and data file S1. (**D**) Flagellin-dependent responses in colonoids: IL-8 levels quantified by ELISA. Data represent means ± SEM. Significance between *Rh*FlaB and *Rh*FlaB-FliC D0 means was determined by unpaired, two-tailed *t* test (****$P < 0.0001$). (**E** and **F**) TLR5-dependent responses in mice: Serum Cxcl1 (E) and IL-6 (F) levels were measured by ELISA after intraperitoneal injection of flagellins. Significance between *Rh*FlaB and *Rh*FlaB-FliC D0 means was determined by unpaired, two-tailed *t* test (*$P < 0.05$).

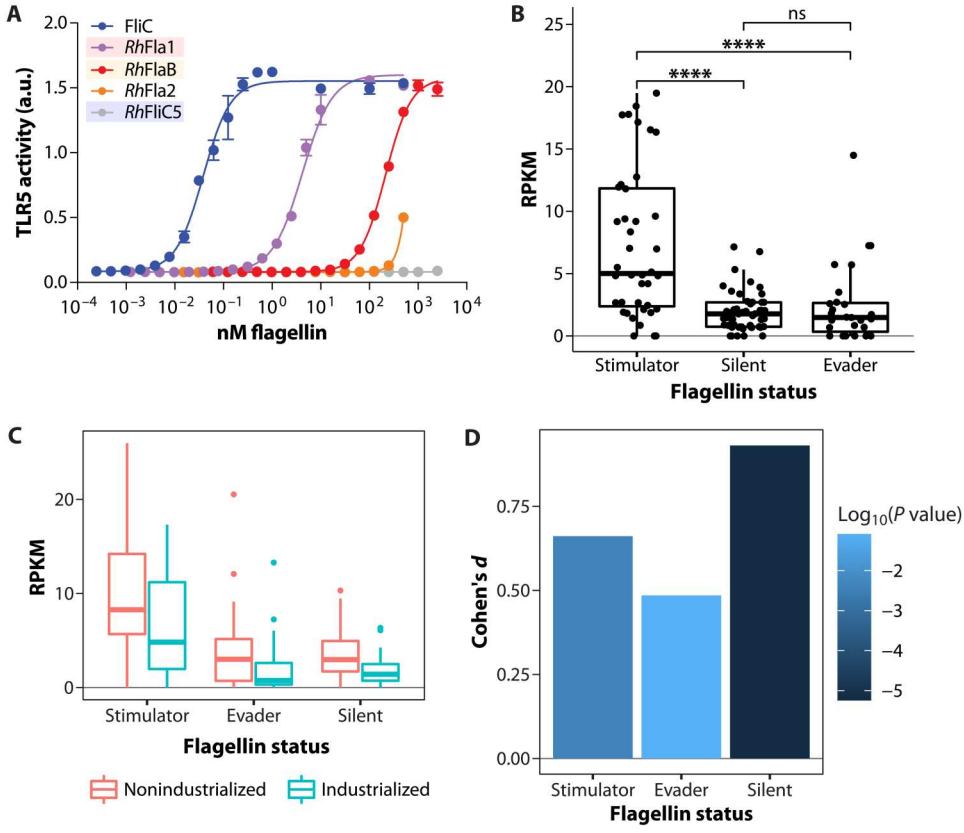

Fig. 5. Silent flagellins were enriched in nonindustrialized populations. (**A**) Within-species flagellin diversity: TLR5 activity of flagellins encoded by *R. hominis* (table S3) as in Fig. 2B. Data are means ± SEM for $n = 3$. Flagellins are shaded to reflect status (see fig. S4 for *Rh*FliC5 binding and table S1 for EC_{50} values). (**B**) Abundance of stimulator, silent, and evader flagellins: Boxplots show median RPKMs of flagellins in human metagenomes ($n = 1783$). Significance was determined by post hoc pairwise *t* tests [analysis of variance (ANOVA), type II, $F = 25.3$, $P = 6.8 \times 10^{-10}$]. ****$P < 0.0001$. See fig. S8. (**C**) Median RPKM of flagellins by industrialization status and flagellin status: See fig. S8 and data file S2. (**D**) Effect sizes (Cohen's *d*) and *t* test *P* values corresponding to industrialization versus nonindustrialization comparisons within each flagellin status.

compared with metagenomes from industrialized populations, despite lower relative abundance of Lachnospiraceae (Fig. 5C and fig. S6, B and C). The decrease in flagellin abundance with industrialization was most pronounced for the silent flagellins (Fig. 5D). Reduced flagellin diversity may reflect shifts in host-microbiome interactions associated with industrialization.

DISCUSSION

The understanding of how TLR5 interacts with its primary ligand, flagellin, comes mostly from the study of flagellins encoded by Pseudomonadota (formerly Proteobacteria), in particular, the pathogens *Salmonella* and *H. pylori*, and others such as *Escherichia coli*. These studies led to the discovery of the TLR5 epitope, a conserved region in the flagellin nD1, whose binding is considered to be required for TLR5 recognition and subsequent stimulation. We showed here that, in addition to the TLR5 epitope, the D0 of FliC allosterically bound TLR5, akin to a homotropic ligand. Our work indicated that, in addition to these modes of interaction (i.e., recognition followed by activation versus nonrecognition), a third mode, very common in commensal bacteria prevalent in the gut, allows bacteria to express flagellins that retain the TLR5 epitope without inducing a robust TLR5 response. Although commensal bacteria also produced stimulator and evader flagellins (all three types can be encoded in a single genome), our analysis of metagenomes indicated that silent flagellins are very common in the human gut and therefore represent a substantial, previously unappreciated, yet physiologically relevant, population of TLR5 ligands.

Our current model proposes that TLR5 adopts different conformations, as evidenced by the presence of both monomeric and dimeric TLR5, and that flagellins have different binding preferences for these receptor states. Although the FliC D0 conferred binding to dimeric TLR5, silent flagellins failed to bind this receptor state and were weak agonists relative to FliC as a result. However, receptor binding to the TLR5 epitope enabled silent flagellins to activate TLR5 at high concentrations, in contrast to *Hp*FlaA. Our data further suggest that FliC binding to TLR5 complexes induced a conformational change, rather than receptor dimerization, similarly to what has been described for other TLRs (*35*). This mechanism is compatible with previous studies that identify an additional TLR5 binding interface in the cD1 domain of flagellin (*7*, *9*). Although our work showed that FliC PIM ΔD0 failed to bind the receptor, this does not exclude a role for the cD1 domain in TLR5 activation: TLR5 conformational changes are likely accompanied by alterations in flagellin-TLR5 binding. Mutating the FliC cD1 is known to affect

TLR5 activation, consistent with a direct interaction between this region and the receptor in the crystal structure of the complex.

Together, our work highlights how pattern recognition by TLR5 can occur without downstream signaling. By probing into the weak agonism of flagellins produced by commensal gut bacteria, we discovered a third class of flagellins that contain the epitope recognized by TLR5 yet poorly activate the receptor. Allosteric activation of TLR5 allows the host to tolerate silent flagellins from commensal bacteria while remaining responsive to faint levels of stimulatory flagellin.

MATERIALS AND METHODS

Study design

The objective of this study was to identify flagellins encoded by common human gut bacteria and characterize their interaction with the host immune receptor TLR5. Publicly available gut metagenomes from non-IBD individuals were used to identify candidate flagellins. We recombinantly expressed these proteins and used biochemical assays to test their ability to bind and activate TLR5. We identified a class of flagellin, termed "silent flagellin," characterized by strong binding and weak activity. Mutational analysis revealed that this "silent recognition" by TLR5 resulted from the inability of silent flagellins to bind TLR5 dimers through a previously unidentified cD0 binding site. Last, we accessed the prevalence of silent flagellins in human gut metagenomes globally.

Reagents

Chemicals were obtained from Sigma-Aldrich [2-mercaptoethanol, bromophenol blue, bovine serum albumin (BSA), LB with agar, EDTA, imidazole, guanidine HCl, Igepal CA-630, N-acetylcysteine, nicotinamide, SB202190, SDS, and Tween 20], Carl Roth (bis-tris, bicine, Hepes, NaH_2PO_4, Na_2HPO_4, urea, tris, NaCl, KCl, methanol, milk powder, MES, and $MgCl_2$), or Thermo Fisher Scientific (glycerol) unless stated otherwise. Primary antibodies included mouse anti-Myc monoclonal 4A6 (Merck, 05-724), rat anti-HA monoclonal 3F10 (Roche, 11867423001), rabbit anti–phospho–NF-κB p65 (Ser^{536}) monoclonal 93H1 [Cell Signaling Technology (CST), 3033], rabbit anti–phospho-IkBα (Ser^{32}) monoclonal 14D4 (CST, 2859), mouse anti–NF-κB p65 monoclonal L8F6 (CST 6956), and mouse anti-IkBα monoclonal L35A5 (CST, 4814). Secondary antibodies were from LI-COR (IRDye 800CW goat anti-rat IgG and IRDye 680RD/800CW goat anti-mouse IgG).

$TLR5^{N14}$ binding to AP-tagged flagellins

Flagellin (prey) and $TLR5^{N14}$ (bait) constructs were cloned into modified pLIB vector containing N-terminal BiP or GP64 signal peptide, respectively, and C-terminal Strep-II affinity tag as described previously (*36*, *37*). Flagellin constructs were further modified to include C-terminal AP tag. $TLR5^{N14}$ bait construct was generated by inserting first 14 LRRs of human TLR5 and C-terminal variable leucine repeat adaptor sequence upstream of human IgG-Fc and 6XHis tag. Constructs were codon-optimized for *Drosophila melanogaster* expression and assembled using Gibson assembly [New England Biolabs (NEB), E2611]. Plasmids were transformed into NEB 5-alpha *E. coli* (NEB, C2987), cultured in 5 ml of LB supplemented with carbenicillin (100 μg/ml; Invitrogen), and purified using ZymoPURE plasmid miniprep kit (Zymo Research, D4211).

Sequence-verified constructs were transformed into DH10EMBaCY cells to generate recombinant baculovirus genomes (bacmids) as described previously (*36*, *38*). Bacmids were transfected into SF9 insect cells using FuGENE 6 transfection reagent (Promega, E2311) and expanded to generate initial (V0) viral stock. Two successive rounds of viral amplification were performed through 1:100 (v/v) inoculations of 25 ml of SF9 cell cultures with baculovirus-containing supernatant, each lasting 72 hours. The supernatant from the second culture (V2) was used to induce protein production. Protein production was optimized for each construct through expression trials in Hi5 cells using inoculation ratios from 1:10 to 1:1000. Proteins were harvested after 48 hours except for FliC PIM and *Hp*FlaA, which were harvested after 72 hours.

Secreted prey proteins in supernatant were collected after centrifugation (800 relative centrifugal force, 5 min) and filter-purified. Bait protein was obtained from Hi5 cell pellets by ultrasonication in lysis buffer [50 mM Hepes, 300 mM NaCl, 5% glycerol, 0.1% Tween 20, 5 mM 2-mercaptoethanol, 1× protease inhibitor (Serva, 39107)] and then filter-purified. Protein expression was confirmed by immunoblotting with Strep-Tactin conjugated to horseradish peroxidase (HRP) (IBA, 2-1502-001).

Flagellin candidates from human commensals were polymerase chain reaction (PCR)–amplified from pETM11 vectors using Phusion high-fidelity DNA polymerase (Thermo Fisher Scientific, F530S). Inserts were cloned into pECIA14 vector using Gibson assembly (NEBuilder HiFi DNA Assembly Master Mix E2621). Insect cell protein expression was performed as previously described (*39*, *40*) with minor modifications. AP-tagged flagellins were expressed via transient transfection in *D. melanogaster* Schneider 2 (S2) cells using Expres^2 TR (Expres^{2}ion Biotechnologies). During transfection, cells were shifted from 27° to 25°C. Protein expression was induced with 1 mM $CuSO_4$ 24 hours after transfection. Supernatant media were collected 72 hours after induction. Protease inhibitors (cOmplete, EDTA-free, Roche, 5056489001) and 0.02% NaN_3 were added to the media. Expressed proteins were confirmed by immunoblotting using anti–Flag-HRP (Sigma-Aldrich, A8592) antibody. Nine candidates failed to express, resulting in 40 abundant and 76 *Rh*FlaB-like candidates.

Each assay contained *Rh*FlaB, FliC, and *Hp*FlaA as controls. AP-tagged flagellins were quantified by incubating 25 μl of each prey sample in 100 μl of BluePhos microwell substrate (SeraCare, 5120-0059) for 30 min at room temperature. Prey concentrations were normalized to *Rh*FlaB based on their relative response to the BluePhos substrate by dilution in PBS-T [phosphate-buffered saline + 0.1% (v/v) Tween 20]. $TLR5^{N14}$-flagellin binding was assayed following a previously described method (*39*), with some modifications. Flagellin proteins and $TLR5^{N14}$ were prediluted 1:10 in PBS-T and then mixed by rocking for 2 hours at 4°C. Ninety-six–well Pierce protein A–coated plates (Thermo Fisher Scientific, 15132) were activated by two washes with 200 μl of PBS-T. $TLR5^{N14}$-flagellin mixtures (200 μl) were added to each well and incubated overnight at 4°C. Each sample had three replicates per plate, alongside corresponding flagellin-only wells as controls. AP activity was quantified using BluePhos microwell substrate, with the absorbance measured at 650 nm after 3 hours. Relative $TLR5^{N14}$-flagellin binding strength was quantified by subtracting the background absorbance from flagellin-only wells from mixed wells and then normalized to FliC. Negative values were set to zero.

Myc-tagged flagellins

Flagellin cDNAs were commercially synthesized (Integrated DNA Technologies) with N-terminal Myc tags and cloned into pETM11 vector downstream from 6x-histidine tag (NEBuilder HiFi DNA assembly) before transformation in TOP10 cells (Invitrogen). Mutated flagellins and flagellins containing the *Salmonella* FliC (UniProt ID: P06179) D0 (amino acids 1 to 41 and amino acids 456 to 494) were generated using Gibson assembly. Plasmid inserts were Sanger-sequenced and transformed into ClearColi BL21(DE3) cells (Lucigen).

Unless stated otherwise, ClearColi-transformed cells were grown at 37°C in LB supplemented with kanamycin (50 μg/ml; Gibco) to OD_{600} (optical density at 600 nm) of 0.4 to 0.6/ml. Protein expression was induced with 0.1 M isopropyl-β-D-thiogalactopyranoside (MP Biomedicals), and cells were incubated for 2.5 hours at 37°C. Cells were harvested by centrifugation (3400*g* for 15 min), and pellets were resuspended in 1:100 volume lysis buffer [10 mM tris-HCl (pH 8), 8 M urea, and 100 mM NaH_2PO_4]. Cells expressing *H. pylori* FlaA (or FlaA-FliC D0) were grown at 25°C for 4 hours after induction and then lysed in 10 mM tris-HCl (pH 8), 6 M guanidine HCl, and 100 mM Na_2HPO_4. Cells were lysed for 2 hours at room temperature, and the supernatant was collected after centrifugation (16,000*g* for 20 min). These lysates were used to screen for TLR5 activity in Figs. 1D and 4 (A and B).

For flagellin purification, cleared lysates were incubated with Ni-NTA (nitrilotriacetic acid) agarose (1:4 volume; QIAGEN) for 2 hours at room temperature. Agarose was then washed once with two column volumes (CVs) of lysis buffer and then washed twice with two CVs of lysis buffer (pH 6.3) and twice with two CVs of lysis buffer (pH 5.9). Ni-NTA–bound proteins were eluted with 0.5 CV of lysis buffer (pH 4.5). Eluates were dialyzed at 4°C in 20 mM tris-HCl (pH 8), 300 mM NaCl, and 5 mM $MgCl_2$ with the exception of *H. pylori* FlaA, CAZDK3, CAZDK3-FliC D0, and R7CCG6-FliC D0, which were dialyzed in 50 mM Hepes (pH 7.4), 114 mM NaCl, and 1.5 mM Na_2HPO_4. Purified proteins were quantified by bicinchoninic acid assay (BCA; Pierce), and aliquots were stored at −80°C. For mouse experiments, non–Myc-tagged flagellins were purified as described above and passed through polymyxin B–agarose (Sigma-Aldrich) before BCA quantification.

TLR5 HEK-Blue activity assay

HEK-Blue hTLR5 cells (InvivoGen) were grown in 5% CO_2 at 37°C in medium [GlutaMAX Dulbecco's modified Eagle's medium (DMEM) and 10% fetal bovine serum (FBS; Gibco)] containing selection antibiotics [Zeocin (100 μg/ml; InvivoGen) and blasticidin (30 μg/ml; InvivoGen)] to 90% confluence. Cells were detached using prewarmed PBS (pH 7.4; Gibco), resuspended in medium (1×10^5 per ml), and distributed in 96-well plates (180 μl per well). Purified flagellins were serially diluted in PBS, and 20 μl were added in triplicate to HEK cells for a final volume of 200 μl per well. Plates were incubated for 18 hours (5% CO_2, 37°C), and then 20 μl of medium from each well were added to 180 μl of QUANTI-Blue solution (InvivoGen). Absorbance at 635 nm was measured after 30 min. EC_{50} values for purified flagellins were calculated by plotting absorbance values against flagellin concentrations in Prism 9 (GraphPad) and performing weighted, nonlinear regression analysis. Unless stated otherwise, no constraints were set. LogEC_{50} values were averaged.

Flagellin candidates were screened for TLR5 activity (Figs. 1D and 4, A and B) as described above with the following modifications. HEK cells were plated at 190 μl per well, and bacterial lysates expressing flagellins were serially diluted in PBS such that final volume in the assay ranged from 1 to 1×10^{-6} μl. Lysates were added in duplicate to HEK cells to a final volume of 200 μl per well. The relative amount of flagellin expression was determined by quantifying Myc signal in 0.1 μl of lysate (diluted 10-fold in PBS) by immunoblotting. TLR5 activity was determined by plotting absorbance values against bacterial lysate volume in Prism 9 and calculating EC_{50} values from weighted, nonlinear regression analysis with the following constraints: Hill slope was set to 1 and top less than or equal to shared value. Activity was normalized to protein expression by calculating EC_{50} multiplied by Myc signal relative to FliC. Values were negative $\log_{10}$-transformed. Candidates with activity less than empty vector control were assigned $\log_{10}$ TLR5 activity values of −6.

Full-length TLR5-flagellin pull-down assays

hTLR5-HA HEK cells (InvivoGen) were incubated in 5% CO_2 at 37°C in medium (GlutaMAX DMEM and 10% FBS; Gibco) supplemented with selection antibiotic blasticidin (10 μg/ml). Cells were grown to 90% confluency in T75 flasks (Grenier), detached with prewarmed PBS (pH 7.4), and harvested by centrifugation (3400*g* for 3 min). Pellets were resuspended in 1 ml of prechilled lysis buffer [25 mM Tris-HCl (pH 7.5), 1% (v/v) Igepal CA-630, 100 mM NaCl, 5 mM imidazole, and 10% v/v glycerol] supplemented with 1× Halt protease inhibitors (Thermo Fisher Scientific).

Cells were incubated on ice for 15 min, followed by centrifugation at 4°C (855*g* for 20 min). Supernatant was collected, and protein levels were quantified by BCA. For each reaction, 500 μg of cell lysate were diluted in 300 μl of lysis buffer and incubated with 150 pmol of Myc-tagged flagellin or buffer control for 2 hours at 4°C. TALON resin (Takara) was washed twice with PBS (pH 7) and twice with lysis buffer before resuspension in PBS; 20 μl of washed resin were added to each reaction. After 2 hours of incubation at 4°C, resin was pelleted at 4°C (855*g* for 1 min) and washed 3× with 750 μl of lysis buffer supplemented with protease inhibitors (cOmplete, mini, EDTA-free). The bound fraction was eluted with 20 μl of lysis buffer containing 200 mM imidazole and transferred to fresh tubes. Samples (25 μg of input and eluted fractions) were analyzed by gel electrophoresis followed by immunoblotting.

For pull-downs with cross-linked lysates, hTLR5-HA HEK cells were plated in Nunc EasYdishes (Thermo Fisher Scientific, 150460) to 70% confluence. Immediately before BS^3 treatment, cells were washed with 1 ml of PBS and then incubated with 2.33 mM BS^3 cross-linker (Pierce, 21580) in 1.5 ml of PBS for 30 s on ice. Reaction was quenched with 1 M Tris-HCl (pH 8) to a final concentration of 50 mM. Crosslinker solution was removed, and cells were scraped into prechilled lysis buffer (750 μl) and then processed as described above with the following modification: The input sample was incubated with anti-HA resin (Roche, 11815016001) for 2 hours at 4°C, and bound proteins were eluted with 20 μl of HA peptide [1 mg/ml in Tris-buffered saline (TBS); Thermo Fisher Scientific, 26184] for 15 min at 37°C.

Flagellin diversity in human microbiome

Flagellins in genomes of isolates and metagenome-assembled genomes from microbial taxa present in the human gut microbiome were assessed using predicted proteome from Unified Human Gastrointestinal Genome catalog v.2.0 (available at www.ebi.ac.uk/metagenomics/genome-catalogues/human-gut-v2-0). We considered a predicted protein a flagellin if eggNOG ID was "COG1344" and annotated as flagellin (as opposed to other accessory proteins). On the basis of this, we identified 5404 (0.053%) flagellins spanning eight phyla.

Flagellin selection

The selection process of candidate flagellins is summarized in fig. S1. Using the flagellin database from (*29*), we annotated database sequences with InterProScan (*41*) and only retained sequences having both Pfam domains PF00669 (flagellin N-terminal domain, "ND") and PF00700 (flagellin C-terminal domain, "CD"). Our filtered database comprised 33,051 flagellin protein sequences. To identify flagellins abundant in the human gut, we filtered the hits from metagenomes of healthy individuals from the IBD multi-omics database of Lloyd-Price *et al.* (*18*) using the median of read counts as a cutoff. The taxonomy of those accessions was assigned using the taxonomizr R package (*42*) to obtain their taxids and a further search of their tax IDs within Genome Taxonomy Database (GTDB) release 95 (https://data.ace.uq.edu.au/public/gtdb/data/releases/latest/). We searched the remaining accessions using Entrez Direct (*43*) in the National Center for Biotechnology Information (NCBI) Identical Protein Groups database (*44*) to obtain their assembly accessions, which were further searched within the GTDB taxonomy. We ranked the resulting accessions by their read counts and randomly selected across the list to get a broad taxonomic representation of flagellins from the human gut, reducing the dataset to 44 candidates. We tested the random community assembly of the final candidates using different metrics: (i) Cmean and Pagel's Lambda, implemented in the phylosignal R package (*45*), and the standardized effect size (SES) of (ii) mean pairwise distance (MPD) and mean nearest taxon distance (MNTD), implemented in the picante R package (*46*). We retrieved the coding sequence (CDS) of each protein accession from the NCBI Identical Protein Groups database (*44*).

To assess the expression profile of flagellins in the human gut, we retrieved publicly available gut metatranscriptome data (available at https://ibdmdb.org/tunnel/public/HMP2/MTX/1750/products) (*18*). We restricted our assessment to samples from the 26 healthy controls with associated metatranscriptomes in (*18*), selecting one sample (corresponding to their earliest visit) per individual. We filtered the downloaded HUMAnN2 tables (*47*) by retaining features annotated as flagellin and whose abundance could be attributed to a specific taxon by HUMAnN2's tiered search. We further validated that the retained flagellins contained the N-terminal and C-terminal domains. HUMAnN2's tiered search provides NCBI taxonomy annotation; we reannotated the contributing species by matching the NCBI taxonomy annotation to that of GTDB release 95 (*48*).

*Rh*FlaB-like candidates were identified using the flagellin database previously described. A truncated CD protein sequence of *Rh*FlaB (accession: WP_014081191.1) was mapped against the database using DIAMOND blastp (*49*) with parameters "--max-target-seqs 0 --evalue 10^{-3} --very-sensitive." The resulting 10,121 hits were filtered by (i) the median of the length of the alignment (100 amino acids), (ii) the sequence identity (39%), (iii) and the number of mismatches (n = 60), which resulted in 979 protein sequences. We mapped the ND flagellin from *Salmonella* Typhimurium (accession: AHA06007.1) against the selected hits with DIAMOND blastp using parameters "--max-target-seqs 0 --evalue 10-3 --very-sensitive," which reduced the list to 919 protein sequences. The resulting list was manually curated by selecting those sequences that did not have either amino acids arginine or lysine in position 478 of the alignment, leaving a total of 195 sequences.

To reduce the *Rh*FlaB-like flagellins to only those occurring in the human gut, we mapped gut metagenomes from the IBD multi-omics database (*18*) and the Franzosa *et al.* (*50*) dataset to our flagellin database using DIAMOND blastx with parameters "--max-target-seqs 1 --evalue 1e-3 --very-sensitive." The resulting 5131 protein accessions were intersected with the 195 flagellins from the former step, resulting in 145 flagellins that meet the sequence composition criteria and are also present in the human gut. From these, we dereplicated at 99% sequence identity with CD-HIT (*51*, *52*) using the parameters "-c 0.99 -M 16000 -n 5," which reduced the dataset to 85 protein accessions. We retrieved their CDSs from the NCBI Identical Protein Groups database (*44*). To avoid nucleotide sequence redundancy with the previously identified abundant flagellins, we merged the lists of CDSs, and the final nucleotide sequence list was dereplicated using CD-HIT with parameters "-c 0.99 -M 16000 -n 5."

Estimating abundance of silent, stimulator, and evader flagellins in human gut metagenomes

We used metagenomes from the curatedMetagenomicData dataset (*53*). Samples were selected on the basis of the following criteria: (i) shotgun metagenomes sequenced using the Illumina HiSeq platform with a median read length of >95 base pairs; (ii) available Sequence Read Archive accessions; (iii) labeled as adults or seniors or with a reported age of ≥18 years; (iv) without report of antibiotic consumption; (v) without report of pregnancy; (vi) nonlactating women; and (vii) labeled as "healthy" [based on random forest algorithm described in (*53*)]. In cases of multiple samples per individual, one sample was randomly selected. The final dataset consisted of 1783 samples from 21 studies (see data file S2). Only the forward reads were used, and all metagenomes were subsampled to 1 million reads.

We mapped the metagenome reads to all stimulator, silent, and evader flagellins (n = 126) via DIAMOND blastx with the following parameters: --sensitive --iterate --evalue 1e-5 --query-cover 90. The results were used to calculate reads per kilobase of target per million reads (RPKM) for each flagellin in each sample. Cohen's d was calculated with the effsize R package (*54*).

NF-κB activation assay

HEK-Blue hTLR5 cells were plated in six-well plates and grown overnight to 90% confluency in medium (GlutaMAX DMEM and 10% FBS; Gibco). Cells were washed with 1 ml of Hanks' balanced salt solution (HBSS; Gibco, catalog no. 14025092) and then incubated with flagellin (10 nM), TNF-α (10 ng/μl; CST, catalog no. 16769), or buffer diluted in 1 ml of HBSS containing 50 μM (R)-MG132 (Sigma-Aldrich, catalog no. M8699) for 1 hour at 37°C. Cells were harvested and resuspended in 200 μl of lysis buffer [25 mM Tris-HCl (pH 7.5), 1% (v/v) Igepal CA-630, 100 mM NaCl, and 10%

(v/v) glycerol] supplemented with 50 μM MG132 and protease inhibitors. After incubating on ice for 15 min, cells were centrifuged at 4°C (855*g*) for 20 min. Supernatants were collected, protein levels were quantified by BCA, and 40 μg of lysates were diluted in Laemmli loading buffer for analysis by immunoblotting.

Immunoblotting

Protein samples were diluted in 1× Laemmli loading buffer [50 mM Tris-HCl (pH 6.8), 2% (w/v) SDS, 2% (v/v) glycerol, 0.05% (w/v) bromophenol blue, and 2.5% (v/v) 2-mercaptoethanol] and incubated at 95°C for 5 min. Samples and Chameleon Duo Prestained protein ladder (LI-COR) were loaded in 4 to 12% NuPAGE Bis-Tris protein gels (Invitrogen) and run in Mini Gel Tank (Invitrogen) with MES buffer [50 mM MES, 50 mM Tris (pH 7.3), 0.1% (w/v) SDS, and 1 mM EDTA] at 125 V. After electrophoresis, gels were transferred to nitrocellulose membranes (Pierce) at 10 V for 70 min at 4°C in transfer buffer [25 mM bicine, 25 mM Bis-Tris (pH 7.2), 1 mM EDTA, and 10% (v/v) methanol] using the Mini Blot Module (Invitrogen). Membranes were blocked for 30 min with 5% (w/v) milk in TBS-T [19 mM tris (pH 7.6), 137 mM NaCl, 2.7 mM KCl, and 0.1% (v/v) Tween 20] and then incubated in primary antibody for 1 hour at room temperature or overnight at 4°C. Membranes were washed 3× for 5 min with TBS-T and then incubated with secondary antibody for 30 min at room temperature. Membranes were washed 3× before imaging on the Odyssey CLx (LI-COR). Unless stated otherwise, primary and secondary antibodies were diluted in TBS-T with 5% (w/v) milk to 0.1 μg/ml (anti-HA), 0.5 μg/ml (anti-Myc), and 0.05 μg/ml (IRDyes). For NF-κB activation immunoblots, primary antibodies were diluted 1:1000 in TBS-T with 5% (w/v) BSA.

Band intensities from immunoblots were quantified in Image Studio (LI-COR). For pull-downs, the fraction of TLR5 bound to Myc-tagged bait was calculated by first subtracting the background HA signal and then dividing by the Myc signal.

Stool sample preparation for proteomics

A stool sample was previously obtained from an adult female without known disease (Cornell University Institutional Review Board, protocol number 1106002281). Stool (166 mg) was resuspended in PBS (pH 7.4) supplemented with protease inhibitors to 0.03% (w/v) and vortexed at maximum speed for 20 min. Lysates were cleared after centrifugation at 13,700*g* for 10 min at 4°C. Protein levels in the supernatant were quantified by BCA, aliquoted, and stored at −80°C. TLR5-HA HEK lysates were prepared from ~6.5×10^6 cells as described above in lysis buffer lacking imidazole and incubated with 90 μl of anti-HA resin for 2 hours at 4°C. Beads were washed 3× with lysis buffer before incubation with 1.8 mg of stool proteins for 2 hours at 4°C. Before incubation, stool proteins were boiled for 5 min at 98°C and diluted in lysis buffer to a final volume of 300 μl. After incubation, beads were washed 3× with lysis buffer and bound proteins were analyzed by MS.

Flagellin database for proteomics

We extracted total DNA from the above stool sample using the QIAGEN PowerSoil kit following the manufacturer's instructions. Library preparation and sequencing were performed as described above. After quality control, the sequencing depth of the sample was 210,366,585 paired reads. We mapped the sequencing reads to our curated flagellin database using DIAMOND v.2.0.9 (*55*) using parameters "--evalue 1e-3 --ultra-sensitive." The abundance of mapped reads was transformed to reads per kilobase (RPK). We assigned taxonomy to the flagellin hits by matching the NCBI protein ID and its associated assembly to GTDB r95. The final database contained 1028 flagellin sequences.

NanoLC-MS/MS analysis and MS data processing

Proteins were eluted from the washed beads and purified with a 12% NuPAGE Novex bis-tris gel (Invitrogen). Tryptic in-gel digestion of proteins was performed as described previously (*56*), and extracted peptides were desalted using C18 StageTips (*57*). Eluted peptides were subjected to liquid chromatography–tandem mass spectrometry (LC-MS/MS) analysis.

Peptide analysis was performed on an Easy-nLC 1200 system coupled to an Exploris 480 mass spectrometer (Thermo Fisher Scientific) as described elsewhere (*58*) with slight modifications: Peptides were injected onto the column in high-performance liquid chromatography (HPLC) solvent A (0.1% formic acid) at a flow rate of 500 nl/min and subsequently eluted with a 227-min segmented gradient of 10-33-50-90% HPLC solvent B (80% acetonitrile in 0.1% formic acid). During peptide elution, the flow rate was kept constant at 200 nl/min.

In each scan cycle, the 20 most intense precursor ions were sequentially fragmented using higher-energy collisional dissociation (HCD) fragmentation. For both precursors and fragment ions, the maximal injection time mode was set to auto and the AGC target was set to standard. Precursor masses with charge states between 2 and 6 were selected with a resolution of 60,000 at a minimum intensity of 400,000 at a scan range of 300 to 1750 mass/charge ratio (*m*/*z*). They were excluded from further selection for 30 s. HCD collision energy for peptide fragmentation was set to 28%, and resolution for MS2 scans was set to 30,000.

MS data were processed with MaxQuant software suite version 1.6.7.0 (*59*), and database search was performed using the Andromeda search engine (*60*), a module of the MaxQuant. MS/MS spectra were searched against a *Homo sapiens* database obtained from UniProt (released 10 July 2020, 97,795 entries), a database containing 1028 flagellin sequences from various organisms (European Nucleotide Archive, study accession number: PRJEB47632), and a database consisting of 246 commonly observed contaminants. In database search, full tryptic specificity was required and up to two missed cleavages were allowed. Carbamidomethylation of cysteine was set as fixed modification, whereas oxidation of methionine and acetylation of protein N terminus were set as variable modifications. Mass tolerance for precursor ions was set to 4.5 parts per million (ppm) and to 20 ppm for fragment ions. Peptide, protein, and modification site identifications were reported at a false discovery rate of 0.01, estimated by the target/decoy approach (*61*). For protein group quantitation, a minimum of two quantified peptides were required. All search parameters were kept to default values except for the following: Minimal peptide length of 5 amino acids was required, and the iBAQ algorithm was used to estimate quantitative values by dividing the sum of peptide intensities of all detected peptides by the number of theoretically observable peptides of the matched protein (*62*). See data file S1.

Organoid experiments

Organoids derived from human colon were cultured in basement membrane extract (Cultrex PathClear Reduced Growth Factor,

Type 2) and growth medium [Advanced DMEM/F12 supplemented with 50% (v/v) L-WRN conditioned medium described in (*63*), 1 mM Hepes (Gibco), 1× GlutaMAX (Gibco), 1× B27 (Gibco), 1 mM N-acetylcysteine, 10 nM gastrin (Sigma-Aldrich), epidermal growth factor (50 ng/ml; PeproTech), 10 mM nicotinamide, 500 nM A83-01 (Tocris), 10 μM SB202190,10 μM Y27632 (Tocris), and 250 nM CHIR99021 (Tocris)] for 4 days in 96-well plates. Organoids were incubated for 18 hours in the presence of flagellins (10 nM) or buffer control diluted in growth medium at 37°C. IL-8 cytokine levels were quantified from 20 μl of medium according to manufacturer′s instructions [R&D Systems, Human IL8 DuoSet enzyme-linked immunosorbent assay (ELISA), catalog no. DY208].

Mouse experiments

All animal studies were performed at Georgia State University under an approved animal protocol (Institutional Animal Care and Use Committee, no. A17047). Eight-week-old female or male C57BL/6 WT and *Tlr5*$^{-/-}$/*Nlrc4*$^{-/-}$ mice, bred at Georgia State University, were treated with 10 or 20 μg of FliC, *Rh*FlaB, or *Rh*FlaB-FliC D0 by intraperitoneal injection. Two hours later, blood was collected from these mice via retrobulbar intraorbital capillary plexus, and serum was isolated by centrifugation at 4°C using Mini-Collect serum separator tubes (Greiner Bio-One). Cxcl1, IL-6, and IL-18 in serum were quantitated using ELISA kits from R&D Systems (catalog nos. DY453, DY406, and 7625, respectively) according to the manufacturer′s instructions. Absorbance values lower than blank controls were set to zero.

General data analysis and visualization

Data were processed, analyzed, and visualized with snakemake (*64*), conda (*65*), and R (*66*) and the following R packages: dplyr (*67*), tidyr (*68*), ggplot2 (*69*), and ggpubr (*70*). Protein Data Bank (PDB) structure was color-coded using PyMOL (*71*).

Statistics

Statistical analyses on mouse, organoid, and *Rh*FlaB mutant data were performed in Prism 9 (GraphPad). Differences between groups were assessed by unpaired, two-tailed *t* test. Unless stated otherwise, all other analyses were performed in R. We defined $P < 0.05$ as significant.

Supplementary Materials

This PDF file includes:
Figs. S1 to S8
Tables S1 to S3

Other Supplementary Material for this manuscript includes the following:
Data files S1 and S2
Table S4

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. C. A. Janeway Jr., Approaching the asymptote? Evolution and revolution in immunology. *Cold Spring Harb. Symp. Quant. Biol.* **54**, 1–13 (1989).
2. I. Botos, D. M. Segal, D. R. Davies, The structural biology of Toll-like receptors. *Structure* **19**, 447–459 (2011).
3. R. Medzhitov, Toll-like receptors and innate immunity. *Nat. Rev. Immunol.* **1**, 135–145 (2001).
4. B. Chassaing, R. E. Ley, A. T. Gewirtz, Intestinal epithelial cell toll-like receptor 5 regulates the intestinal microbiota to prevent low-grade inflammation and metabolic syndrome in mice. *Gastroenterology* **147**, 1363–1377.e17 (2014).
5. F. Hayashi, K. D. Smith, A. Ozinsky, T. R. Hawn, E. C. Yi, D. R. Goodlett, J. K. Eng, S. Akira, D. M. Underhill, A. Aderem, The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. *Nature* **410**, 1099–1103 (2001).
6. S. Maki-Yonekura, K. Yonekura, K. Namba, Conformational change of flagellin for polymorphic supercoiling of the flagellar filament. *Nat. Struct. Mol. Biol.* **17**, 417–422 (2010).
7. K. D. Smith, E. Andersen-Nissen, F. Hayashi, K. Strobe, M. A. Bergman, S. L. R. Barrett, B. T. Cookson, A. Aderem, Toll-like receptor 5 recognizes a conserved site on flagellin required for protofilament formation and bacterial motility. *Nat. Immunol.* **4**, 1247–1253 (2003).
8. M. A. B. Kreutzberger, C. Ewing, F. Poly, F. Wang, E. H. Egelman, Atomic structure of the *Campylobacter jejuni* flagellar filament reveals how ε Proteobacteria escaped Toll-like receptor 5 surveillance. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 16985–16991 (2020).
9. S.-I. Yoon, O. Kurnasov, V. Natarajan, M. Hong, A. V. Gudkov, A. L. Osterman, I. A. Wilson, Structural basis of TLR5-flagellin recognition and signaling. *Science* **335**, 859–864 (2012).
10. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
11. E. Andersen-Nissen, K. D. Smith, K. L. Strobe, S. L. R. Barrett, B. T. Cookson, S. M. Logan, A. Aderem, Evasion of Toll-like receptor 5 by flagellated bacteria. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 9247–9252 (2005).
12. T. C. Cullender, B. Chassaing, A. Janzon, K. Kumar, C. E. Muller, J. J. Werner, L. T. Angenent, M. E. Bell, A. G. Hay, D. A. Peterson, J. Walter, M. Vijay-Kumar, A. T. Gewirtz, R. E. Ley, Innate and adaptive immunity interact to quench microbiome flagellar motility in the gut. *Cell Host Microbe* **14**, 571–581 (2013).
13. T. Steiner, S. Ivison, C. Wang, C. Elson, P-0163: The A4-Fla2 flagellin, a dominant antigen in Crohn?s disease, is a poor TLR5 agonist. *Inflamm. Bowel Dis.* **15**, S55 (2009).
14. B. A. Neville, P. O. Sheridan, H. M. B. Harris, S. Coughlan, H. J. Flint, S. H. Duncan, I. B. Jeffery, M. J. Claesson, R. P. Ross, K. P. Scott, P. W. O'Toole, Pro-inflammatory flagellin proteins of prevalent motile commensal bacteria are variably abundant in the intestinal microbiome of elderly humans. *PLOS ONE* **8**, e68919 (2013).
15. A. Almeida, S. Nayfach, M. Boland, F. Strozzi, M. Beracochea, Z. J. Shi, K. S. Pollard, E. Sakharova, D. H. Parks, P. Hugenholtz, N. Segata, N. C. Kyrpides, R. D. Finn, A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
16. K. L. Alexander, Q. Zhao, M. Reif, A. F. Rosenberg, P. J. Mannon, L. W. Duck, C. O. Elson, Human microbiota flagellins drive adaptive immune responses in Crohn's disease. *Gastroenterology* **161**, 522–535.e6 (2021).
17. P. Louis, H. J. Flint, Diversity, metabolism and microbial ecology of butyrate-producing bacteria from the human large intestine. *FEMS Microbiol. Lett.* **294**, 1–8 (2009).
18. J. Lloyd-Price, C. Arze, A. N. Ananthakrishnan, M. Schirmer, J. Avila-Pacheco, T. W. Poon, E. Andrews, N. J. Ajami, K. S. Bonham, C. J. Brislawn, D. Casero, H. Courtney, A. Gonzalez, T. G. Graeber, A. B. Hall, K. Lake, C. J. Landers, H. Mallick, D. R. Plichta, M. Prasad, G. Rahnavard, J. Sauk, D. Shungin, Y. Vázquez-Baeza, R. A. White III; IBDMDB Investigators, J. Braun, L. A. Denson, J. K. Jansson, R. Knight, S. Kugathasan, D. P. B. Mc Govern, J. F. Petrosino, T. S. Stappenbeck, H. S. Winter, C. B. Clish, E. A. Franzosa, H. Vlamakis, R. J. Xavier, C. Huttenhower, Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).
19. V. Forstnerič, K. Ivičak-Kocjan, T. Plaper, R. Jerala, M. Benčina, The role of the C-terminal D0 domain of flagellin in activation of Toll like receptor 5. *PLOS Pathog.* **13**, e1006574 (2017).
20. A. M. Patterson, I. E. Mulder, A. J. Travis, A. Lan, N. Cerf-Bensussan, V. Gaboriau-Routhiau, K. Garden, E. Logan, M. I. Delday, A. G. P. Coutts, E. Monnais, V. C. Ferraria, R. Inoue, G. Grant, R. I. Aminov, Human gut symbiont *Roseburia hominis* promotes and regulates innate immunity. *Front. Immunol.* **8**, 1166 (2017).
21. S. H. Duncan, R. I. Aminov, K. P. Scott, P. Louis, T. B. Stanton, H. J. Flint, Proposal of *Roseburia faecis* sp. nov., *Roseburia hominis* sp. nov. and *Roseburia inulinivorans* sp. nov., based on isolates from human faeces. *Int. J. Syst. Evol. Microbiol.* **56**, 2437–2441 (2006).
22. S. N. Klimosch, A. Försti, J. Eckert, J. Knezevic, M. Bevier, W. von Schönfels, N. Heits, J. Walter, S. Hinz, J. Lascorz, J. Hampe, D. Hartl, J.-S. Frick, K. Hemminki, C. Schafmayer, A. N. R. Weber, Functional TLR5 genetic variants affect human colorectal cancer survival. *Cancer Res.* **73**, 7232–7242 (2013).

23. A. T. Gewirtz, T. A. Navas, S. Lyons, P. J. Godowski, J. L. Madara, Cutting edge: Bacterial flagellin activates basolaterally expressed TLR5 to induce epithelial proinflammatory gene expression. *J. Immunol.* **167**, 1882–1885 (2001).
24. K. Ivičak-Kocjan, G. Panter, M. Benčina, R. Jerala, Determination of the physiological 2:2 TLR5:Flagellin activation stoichiometry revealed by the activity of a fusion receptor. *Biochem. Biophys. Res. Commun.* **435**, 40–45 (2013).
25. M.-H. Khani, M. Bagheri, A. Dehghanian, A. Zahmatkesh, S. Moradi Bidhendi, Z. Salehi Najafabadi, R. Banihashemi, Effect of C-terminus modification in *Salmonella typhimurium* FliC on protein purification efficacy and bioactivity. *Mol. Biotechnol.* **61**, 12–19 (2019).
26. I. A. Hajam, P. A. Dar, I. Shahnawaz, J. C. Jaume, J. H. Lee, Bacterial flagellin—A potent immunomodulatory agent. *Exp. Mol. Med.* **49**, e373 (2017).
27. A. H. López-Yglesias, X. Zhao, E. K. Quarles, M. A. Lai, T. VandenBos, R. K. Strong, K. D. Smith, Flagellin induces antibody responses through a TLR5- and inflammasome-independent pathway. *J. Immunol.* **192**, 1587–1596 (2014).
28. K. Zhou, R. Kanai, P. Lee, H.-W. Wang, Y. Modis, Toll-like receptor 5 forms asymmetric dimers in the absence of flagellin. *J. Struct. Biol.* **177**, 402–409 (2012).
29. D. Hu, P. R. Reeves, The remarkable dual-level diversity of prokaryotic flagellins. *mSystems* **5**, e00705–19 (2020).
30. M. J. Lodes, Y. Cong, C. O. Elson, R. Mohamath, C. J. Landers, S. R. Targan, M. Fort, R. M. Hershberg, Bacterial flagellin is a dominant antigen in Crohn disease. *J. Clin. Invest.* **113**, 1296–1306 (2004).
31. A. T. Gewirtz, P. O. Simon Jr., C. K. Schmitt, L. J. Taylor, C. H. Hagedorn, A. D. O'Brien, A. S. Neish, J. L. Madara, *Salmonella typhimurium* translocates flagellin across intestinal epithelia, inducing a proinflammatory response. *J. Clin. Invest.* **107**, 99–109 (2001).
32. L. Franchi, A. Amer, M. Body-Malapel, T.-D. Kanneganti, N. Özören, R. Jagirdar, N. Inohara, P. Vandenabeele, J. Bertin, A. Coyle, E. P. Grant, G. Núñez, Cytosolic flagellin requires Ipaf for activation of caspase-1 and interleukin 1β in salmonella-infected macrophages. *Nat. Immunol.* **7**, 576–582 (2006).
33. E. A. Miao, C. M. Alpuche-Aranda, M. Dors, A. E. Clark, M. W. Bader, S. I. Miller, A. Aderem, Cytoplasmic flagellin activates caspase-1 and secretion of interleukin 1β via Ipaf. *Nat. Immunol.* **7**, 569–575 (2006).
34. M. Vijay-Kumar, F. A. Carvalho, J. D. Aitken, N. H. Fifadara, A. T. Gewirtz, TLR5 or NLRC4 is necessary and sufficient for promotion of humoral immunity by flagellin. *Eur. J. Immunol.* **40**, 3528–3534 (2010).
35. E. Latz, A. Verma, A. Visintin, M. Gong, C. M. Sirois, D. C. G. Klein, B. G. Monks, C. J. McKnight, M. S. Lamphier, W. P. Duprex, T. Espevik, D. T. Golenbock, Ligand-induced conformational changes allosterically activate Toll-like receptor 9. *Nat. Immunol.* **8**, 772–779 (2007).
36. F. Weissmann, G. Petzold, R. V. Linden, P. J. H. In 't Veld, N. G. Brown, F. Lampert, S. Westermann, H. Stark, B. A. Schulman, J.-M. Peters, biGBac enables rapid gene assembly for the expression of large multisubunit protein complexes. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E2564–E2569 (2016).
37. V. Altmannova, A. Blaha, S. Astrinidis, H. Reichle, J. R. Weir, InteBac: An integrated bacterial and baculovirus expression vector suite. *Protein Sci.* **30**, 108–114 (2021).
38. C. Bieniossek, T. Imasaki, Y. Takagi, I. Berger, MultiBac: Expanding the research toolbox for multiprotein complexes. *Trends Biochem. Sci.* **37**, 49–57 (2012).
39. E. Smakowska-Luzan, G. A. Mott, K. Parys, M. Stegmann, T. C. Howton, M. Layeghifard, J. Neuhold, A. Lehner, J. Kong, K. Grünwald, N. Weinberger, S. B. Satbhai, D. Mayer, W. Busch, M. Madalinski, P. Stolt-Bergner, N. J. Provart, M. S. Mukhtar, C. Zipfel, D. Desveaux, D. S. Guttman, Y. Belkhadir, An extracellular network of *Arabidopsis* leucine-rich repeat receptor kinases. *Nature* **553**, 342–346 (2018).
40. K. Parys, N. R. Colaianni, H.-S. Lee, U. Hohmann, N. Edelbacher, A. Trgovcevic, Z. Blahovska, D. Lee, A. Mechtler, Z. Muhari-Portik, M. Madalinski, N. Schandry, I. Rodríguez-Arévalo, C. Becker, E. Sonnleitner, A. Korte, U. Bläsi, N. Geldner, M. Hothorn, C. D. Jones, J. L. Dangl, Y. Belkhadir, Signatures of antagonistic pleiotropy in a bacterial flagellin epitope. *Cell Host Microbe* **29**, 620–634.e9 (2021).
41. P. Jones, D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A. F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S.-Y. Yong, R. Lopez, S. Hunter, InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
42. S. Sherrill-Mix, Functions to work with NCBI accessions and taxonomy [R package taxonomizr version 0.8.0] (2021); https://CRAN.R-project.org/package=taxonomizr.
43. J. Kans, *Entrez Direct: E-utilities on the Unix Command Line* (National Center for Biotechnology Information, 2022).
44. D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, J. Ostell, K. D. Pruitt, E. W. Sayers, GenBank. *Nucleic Acids Res.* **46**, D41–D47 (2018).
45. F. Keck, F. Rimet, A. Bouchez, A. Franc, phylosignal: An R package to measure, test, and explore the phylogenetic signal. *Ecol. Evol.* **6**, 2774–2780 (2016).
46. S. W. Kembel, P. D. Cowan, M. R. Helmus, W. K. Cornwell, H. Morlon, D. D. Ackerly, S. P. Blomberg, C. O. Webb, Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26**, 1463–1464 (2010).
47. E. A. Franzosa, L. J. McIver, G. Rahnavard, L. R. Thompson, M. Schirmer, G. Weingart, K. S. Lipson, R. Knight, J. G. Caporaso, N. Segata, C. Huttenhower, Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).
48. D. H. Parks, M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P.-A. Chaumeil, P. Hugenholtz, A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
49. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
50. E. A. Franzosa, A. Sirota-Madi, J. Avila-Pacheco, N. Fornelos, H. J. Haiser, S. Reinker, T. Vatanen, A. B. Hall, H. Mallick, L. J. McIver, J. S. Sauk, R. G. Wilson, B. W. Stevens, J. M. Scott, K. Pierce, A. A. Deik, K. Bullock, F. Imhann, J. A. Porter, A. Zhernakova, J. Fu, R. K. Weersma, C. Wijmenga, C. B. Clish, H. Vlamakis, C. Huttenhower, R. J. Xavier, Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat. Microbiol.* **4**, 293–305 (2019).
51. W. Li, A. Godzik, Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
52. L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
53. E. Pasolli, L. Schiffer, P. Manghi, A. Renson, V. Obenchain, D. T. Truong, F. Beghini, F. Malik, M. Ramos, J. B. Dowd, C. Huttenhower, M. Morgan, N. Segata, L. Waldron, Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* **14**, 1023–1024 (2017).
54. M. Torchiano, Efficient effect size computation [R package effsize version 0.8.1] (2020); https://CRAN.R-project.org/package=effsize.
55. B. Buchfink, K. Reuter, H.-G. Drost, Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
56. N. Borchert, C. Dieterich, K. Krug, W. Schütz, S. Jung, A. Nordheim, R. J. Sommer, B. Macek, Proteogenomics of *Pristionchus pacificus* reveals distinct proteome structure of nematode models. *Genome Res.* **20**, 837–846 (2010).
57. J. Rappsilber, M. Mann, Y. Ishihama, Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).
58. M. Schmitt, T. Sinnberg, K. Bratl, K. Zittlau, C. Garbe, B. Macek, N. C. Nalpas, Proteogenomics reveals perturbed signaling networks in malignant melanoma cells resistant to BRAF inhibition. *Mol. Cell. Proteomics* **20**, 100163 (2021).
59. J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
60. J. Cox, N. Neuhauser, A. Michalski, R. A. Scheltema, J. V. Olsen, M. Mann, Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
61. J. E. Elias, S. P. Gygi, Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).
62. B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, M. Selbach, Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
63. H. Miyoshi, T. S. Stappenbeck, In vitro expansion and genetic modification of gastrointestinal stem cells in spheroid culture. *Nat. Protoc.* **8**, 2471–2482 (2013).
64. F. Mölder, K. P. Jablonski, B. Letcher, M. B. Hall, C. H. Tomkins-Tinch, V. Sochat, J. Forster, S. Lee, S. O. Twardziok, A. Kanitz, A. Wilm, M. Holtgrewe, S. Rahmann, S. Nahnsen, J. Köster, Sustainable data analysis with Snakemake. *F1000Res.* **10**, 33 (2021).
65. Anaconda Software, Distribution *Anaconda* documentation (2020); https://docs.anaconda.com/.
66. R Core Team, R: A language and environment for statistical computing (2018); www.R-project.org/.
67. H. Wickham, R. François, L. Henry, K. Müller, dplyr: A grammar of data manipulation (2021); https://CRAN.R-project.org/package=dplyr.
68. H. Wickham, tidyr: Tidy messy data (2021); https://CRAN.R-project.org/package=tidyr.
69. H. Wickham, ggplot2: Elegant graphics for data analysis (2016); https://ggplot2.tidyverse.org.
70. A. Kassambara, ggpubr: "ggplot2" based publication ready plots (2020); https://CRAN.R-project.org/package=ggpubr.
71. L. L. C. Schrödinger, W. DeLano, PyMOL (2020); www.pymol.org/pymol.
72. Y. Perez-Riverol, J. Bai, C. Bandla, D. García-Seisdedos, S. Hewapathirana, S. Kamatchinathan, D. J. Kundu, A. Prakash, A. Frericks-Zipper, M. Eisenacher, M. Walzer, S. Wang, A. Brazma, J. A. Vizcaíno, The PRIDE database resources in 2022: A hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.* **50**, D543–D552 (2022).

Acknowledgments: Human colon organoids from the Stanford Tissue Bank were provided by J. Y. Co from the laboratories of M. R. Amieva and D. Monack. We thank I. Droste-Borel and B. Macek from the Proteome Center Tübingen, C. Lanz and O. Weichenrieder from the Genome Center at the MPI for Biology, and I. Hang and J. Neuhold at the Vienna BioCenter Core Facilities (VBCF ProTech). **Funding:** This work was supported by the Max Planck Society and grants from the Austrian Academy of Sciences through the Gregor Mendel Institute and the Vienna Science and Technology Fund Project (LS17-047) (Y.B.). **Author contributions:** S.J.C. and R.E.L. conceptualized the study. S.J.C., M.E.W.B., D.-H.L., K.P., Z.M.H., and V.A. developed and performed the biochemical experiments. A.B., J.d.l.C.-Z., and N.D.Y. conducted the bioinformatic analyses. J.Z. and Y.W. performed the mouse work. S.J.C., A.B., and N.D.Y. visualized the data. Resources were provided by Z.M.H., J.R.W., Y.B., A.T.G., and R.E.L. N.D.Y. and R.E.L. supervised the study. S.J.C., M.E.W.B., J.d.l.C.-Z., A.B., N.D.Y., D.-H.L., A.T.G., Y.B., K.P., V.A., J.R.W., and R.E.L. wrote and revised the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Raw sequence data of samples used for proteomics are available from European Nucleotide Archive (study accession number PRJEB47632). MS proteomics data are deposited to the ProteomeXchange Consortium via the PRIDE partner repository with dataset identifier PXD033249 (*72*). All other data needed to evaluate the conclusions in the paper are available in the paper or the Supplementary Materials.

Submitted 25 April 2022
Accepted 4 November 2022
Published 6 January 2023
10.1126/sciimmunol.abq7001