Interspecies Interactions Between the Nematode *Pristionchus pacificus* and Bacterial Food Sources

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Marina Agapi Athanasouli
aus Amarousio, Griechenland

Tübingen
2024

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

Tag der mündlichen Qualifikation:	20.12.2024
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter/-in:	Prof. Dr. Ralf J. Sommer
2. Berichterstatter/-in:	Prof. Dr. Boris Macek

ACKNOWLEDGEMENTS

Sometimes it takes a village, albeit in this case a small one. The chief of it, to whom I would like to extend my biggest thanks, is my advisor Christian Rödelsperger. His guidance, support and patience were instrumental to the completion of this PhD. Under his supervision, I learned more things than I could imagine and grew as a scientist; for this I'm forever grateful.

This PhD would not have been possible without both the financial and academic support of Ralf Sommer, who accepted me into his lab and to whom I'm extremely thankful. I would also like to extend my thanks to my university supervisor Boris Macek, whose insight proved very useful throughout these five and a half years. Many thanks go out to my TAC members Nick Youngblut and Thorsten Langner for their direction during my PhD.

Special thanks go to Walli who had the tenacity and bravery to train a bioinformatician in library sequencing protocols while making sure that I left the lab standing. Christian W, I'm eternally grateful for your friendship and for always making sure that the S2 lab was running smoothly, especially when I complained because the incubators were misbehaving (and for the first Schnapps in Germany that doesn't taste like 100% ethanol). George, Athena, Sara, Radhika, my worm sensei Peng, Heike, Hanh and Matthias, thank you for the laughs, for the good times and for making my time here much more fun than it would have ever been without you.

I would not be where I am today without the constant support of my family and friends back home, my wall to lean on when things became overwhelming. I have no words to express my gratitude to my parents who made sure I had everything I needed to succeed in life. I wouldn't be the woman or the sports fan I'm today without my mom and my dad. And I wouldn't be nearly as brave if it wasn't for my sister, who was courageous enough to become a historian in a family of chemists and biologists.

The last and greatest acknowledgement goes to my partner in nerd, football and life in general. Tobi, thank you for listening to my presentations an endless number of times for prep, for hearing me read out this thesis an even higher number of times, for being there during the best and the worst and for making sure I completed this PhD half-sane.

This work is dedicated to my late father. Dad, I did it. I miss you.

I. SUMMARY

The past few decades biological research has turned its attention towards the interspecies interactions taking place between multicellular organisms and their associated microbial communities. The focus has been placed mainly on symbiotic prokaryotes and their ability to modulate their host's physiological traits and responses, with the most representative example belonging to the field of human gut microbiome research. While currently popular, these interactions consist only a fraction of the inter- and intra-species relationships within a given niche. Ecological communities are often represented as complex networks, in an attempt to illustrate the multifaceted interplay present as well as the communication between the interacting partners that may determine their evolutionary trajectories and influence their adaptation. The study of such systems depends on the ability to establish simple but representative proxies. Nematodes, such as *Pristionchus pacificus*, and their diverse environment possess an array of properties that deem them an excellent choice to study interspecies interactions. During the last years, we have gathered important information regarding the effect of bacterial diets on *P. pacificus*' behavior along with the worm's impact on the associated bacterial communities. My work aims to broaden the scope of the questions and hypotheses formed, from specific genes or traits to entire regulatory networks as well as outline the importance of the information harnessed from the bacteria itself. To this end, I initially reconstructed the phylogeny and identified the metabolic potential of a subset of bacteria isolated from *Pristionchus*-associated environments in previous studies. This strategy allowed me to uncover the variation of the predicted metabolic potential across the phylogeny, detect genus-specific metabolic signatures in the bacterial strains and associate them with nematode traits from previous works. Furthermore, I identified environmentally responsive regulatory networks in *P. pacificus* through the implementation of a gene co-expression network constructed with transcriptomic profiles of the nematode and almost half of the sequenced bacterial strains as a dietary source. Lastly, I combined the collected data both from the bacteria and the nematode in a bipartite network and revealed the bacterial metabolic pathways with the strongest effect on the worm's regulatory networks. Overall, my work explores the interspecies interactions between *P. pacificus* and bacterial food sources and offers novel insights on a larger scale.

II. DEUTSCHE ZUSAMMENFASSUNG

In den letzten Jahrzehnten lag ein wesentlicher Schwerpunkt der biologischen Forschung auf dem Einfluss des Darmmikrobioms auf den menschlichen Stoffwechsel und das Immunsystem. Allerdings machen diese Wechselwirkungen nur einen Bruchteil aller Interaktionen zwischen Arten innerhalb einer Nische aus. Die Interaktionen zwischen Partnern innerhalb einer ökologische Gemeinschaft bilden ein komplexes Netzwerk, das die Anpassungsfähigkeit und damit die Evolution der ganzen Gemeinschaft beeinflusst. Um solche Systeme besser zu verstehen, ist es unumgänglich, repräsentative Ersatzstrukturen zu schaffen. Nematoden wie Pristionchus pacificus und ihre vielfältige Umgebung besitzen eine Reihe von Eigenschaften, die sie zu einer ausgezeichneten Wahl für die Untersuchung von Interaktionen zwischen Arten machen. In den letzten Jahren haben wir wichtige Erkenntnisse über die Auswirkungen der bakteriellen Ernährung auf das Verhalten von P. pacificus und den Einfluss des Wurms auf die zugehörigen bakteriellen Gemeinschaften gewonnen. Meine Doktorarbeit zielt darauf ab, den Umfang der Fragen und Hypothesen zu erweitern, von spezifischen Genen oder Merkmalen bis hin zu ganzen regulatorischen Netzwerken, und den Einfluss verschiedener Bakterien besser zu beschreiben. Zu diesem Zweck habe ich 84 Genome von Bakterien sequenziert, die in einer früheren Studie aus Pristionchus-assoziierten Umgebungen isoliert wurden. Anschließend habe ich deren Phylogenie rekonstruiert und die Variation des vorhergesagten Stoffwechselpotenzials untersucht. Dadurch konnte ich gattungsspezifische Stoffwechselsignaturen erkennen und sie mit Nematodenmerkmalen in Verbindung zu bringen. Darüber hinaus identifizierte ich umweltabhängige regulatorische Netzwerke in P. pacificus durch die Analyse eines Gen-Koexpressionsnetzwerks, das aus Expressionsprofilen des Nematoden auf 38 bakteriellen Stämmen erstellt wurde. Schließlich habe ich die gesammelten Daten der Bakterien und des Fadenwurms in einem zweistufigen Netzwerk kombiniert und die bakteriellen Stoffwechselwege mit den stärksten Auswirkungen auf die regulatorischen Netzwerke des Wurms ermittelt. Insgesamt erforscht meine Arbeit die Wechselwirkungen zwischen P. pacificus und bakteriellen Nahrungsquellen und bietet neue Erkenntnisse in größerem Maßstab.

III. LIST OF PUBLICATIONS

1. Rödelsperger C, **Athanasouli M**, Lenuzzi M, Theska T, Sun S, Dardiry M, Wighard S, Hu W, Sharma DR, Han Z. 2019. Crowdsourcing and the feasibility of manual gene annotation: A pilot study in the nematode *Pristionchus pacificus*. *Scientific Reports* 9: 1-9.
2. **Athanasouli M**, Witte H, Weiler C, Loschko T, Eberhardt G, Sommer RJ, Rödelsperger C. 2020. Comparative genomics and community curation further improve gene annotations in the nematode *Pristionchus pacificus*. *BMC Genomics* 21, 708.
3. **Athanasouli M** & Rödelsperger C. 2022. Analysis of repeat elements in the Pristionchus pacificus genome reveals an ancient invasion by horizontally transferred transposons. *BMC Genomics* 23, 523.
4. Wighard SS, **Athanasouli M**, Witte H, Rödelsperger C, Sommer RJ. 2022. A New Hope: A Hermaphroditic Nematode Enables Analysis of a Recent Whole Genome Duplication Event. *Genome Biology and Evolution* 14: evac169.
5. **Athanasouli M**, Akduman N, Röseler W, Theam P, Rödelsperger C. 2023. Thousands of *Pristionchus pacificus* orphan genes were integrated into developmental networks that respond to diverse environmental microbiota. *PLoS Genetics* 19(7): e1010832.
6. Rödelsperger C, Röseler W, **Athanasouli M**, Wighard S, Herrmann M, Sommer RJ. 2024. The genome assembly of *Rhabditoides inermis* from a complex microbial community reveals further evidence for parallel gene family expansions. bioRxiv: 2024.08.02.60598. *Submitted to Genome Biology and Evolution*.
7. Piskobulu V, **Athanasouli M**, Witte H, Feldhaus C, Streit A, Sommer RJ. 2024. High nutritional conditions influence feeding plasticity in *Pristionchus pacificus* and render worms non-predatory. bioRxiv:2024.08.27.609904
8. **Athanasouli M**, Loschko T, Rödelsperger C. 2024. Interspecies systems biology links candidate bacterial pathways to nematode gene expression, behavior, and survival. bioRxiv:2024.10.02.616249 (Submitted to *ISME*).

1. INTRODUCTION

"*Whenever we look at life, we look at networks.*"

Capra and Luisi - The Systems View of Life: A Unifying Vision (2014)

1.1 Reductionism and holism in biology

Until the early 19th century, what we know today as "biology" was essentially natural philosophy. In fact, the first appearances of the term are traced to its systematic use by Jean Baptiste Lamarck (1744-1829) and Reinhold Gottfried Treviranus (1776-1837) in his book *Biologie oder Philosophie der lebenden Natur für Naturforscher und Aerzte*[1, 2]. Up to that point, the investigation of natural phenomena was part of natural philosophy, indicating that the majority of discourse would take place within the realm of philosophy itself. Regardless of the intellectual or scientific framework, interpretations of biological entities have been historically centered either around a holistic approach which suggests viewing them as wholes or a reductionist approach which mandates dismantling them in order to derive a general explanation from the individual parts[3]. While seemingly polar opposites, a combination of reductionism and holism is necessary in most cases[4]. Dissecting any system (for example, a cell) can reduce its complexity and aid both with the investigation and the comprehension of it; however, the interactions between its components and their behavior within the system should be taken into account as well[5]. At the same time, both approaches can present significant drawbacks. A strictly reductionist strategy implies that describing each part of a system can essentially describe the entirety, while an exclusively systemic method can prove daunting and result in incorrect models and assumptions[3].

In the following sections, I will give a brief historical summary of the back and forth between reductionism and holism within the context of natural sciences and in particular biology. I will then focus on the transition from holism to what we know today as Systems Biology and its use of networks as an essential tool of the field. Lastly, I will describe interspecies interactions both generally as well as within the scope of host-microbiome interactions.

1.1.1 A centuries-long debate

The discourse between Western reductionists and holists can be traced back to ancient Greece. The first rationalist formulations came from the Milesian school of thought (circa 5-6th century BC) where its most prominent figures Thales, Anaximander and Anaximenes attempted to explain natural phenomena pragmatically; devoid of any mythological influences[6-8]. Their work is noted as the first incidence of a rationalist approach to philosophy and science[5] and includes their individual notions about the source of natural phenomena[8]. According to Thales and Anaximenes, everything comes from water and air respectively while Anaximander credited an infinite, primitive element as the foundation of life[8]. Their contributions laid the groundwork for a materialistic view of the world - that is the explanation of nature solely with matter as its source[9-11]. The rationalist ideas of the Milesian school were carried on by Leucippus (circa 5th century BC) and his student Democritus (460-370 BC) who established atomism, the theory that everything is comprised from invisible, physically indivisible parts ("*atoma*"), the interactions of which shape nature[7, 8]. The atomists did not consider a final cause as the explanation for an event; rather they questioned the conditions that led to it, providing a mechanistic interpretation[7, 8].

On the opposite side of the atomists stood the teleologists Plato (circa 427-348 BC) and his student Aristotle (384-322 BC). Both were critical of the materialist school of thought and opposed a rationalist approach to the interpretation of nature[7, 10, 12]. The teleologists focused on the "*why*", defining the end goal ("*telos*") of a form as the cause for its existence[13]. While there were significant differences in their teleological perspective - namely Aristotle's attempt to introduce reason in Plato's belief system - they both maintained that order in a system cannot be sufficiently explained through a materialist lens[7, 8, 13-15]. This summarized discourse between mechanism and teleology is the earliest documented.

The Aristotelian view of nature was at the forefront until the Scientific Revolution in the 16th century[16]. In his book *History of the inductive sciences*, William Whewell describes the transition from philosophical explanations to "*dependence upon external observation*"[17], which according to Bertrand Russell relied heavily on the patience of the observer and their courage to form hypotheses[8]. The development of empirical science based on experimental, observation-based evidence found its strongest proponent in Francis Bacon (1561-1626) who rejected the teleologists and favored the

atomist school of thought[8, 18-20]. At the same time, in opposition to Bacon's inductive approach, rationalism was experiencing a revival with René Descartes (1596-1650). With an emphasis on reason, Descartes approached science and the world in general from a mechanistic, deductive point of view, where everything can be broken down in parts that interact with each other resembling the gears of a well-oiled machine[21, 22]. This theory is therefore known as *Descartes' clockwork* and is representative of a reductionist, mechanistic scientific process, that seeks to explain natural phenomena using solely physical and chemical terms while equating them to machinery components[5, 16, 23, 24]. Descartes' ambition to describe everything mechanistically is evident by his depiction of biological functions such as blood circulation and the nervous system in the form of hydraulic systems, paving the way for the rise of mechanistic biology, especially in the field of physiology[25]. His endeavors along with the contributions of Nicolaus Copernicus (1473-1543), Johannes Kepler (1571-1630) and Galileo Galilei (1564-1642) were instrumental in ushering science from the Middle Ages to its modern era and setting the path forward for Newton (1647-1727)[8]. Newton's work provided a mechanistic, deterministic explanation of nature and cemented his status as one of the most important figures of the Scientific Revolution as well as contemporary science[25].

The first reactions to Descartes' mechanistic view of biology came from supporters of two holistic approaches: the vitalists and later the organicists[5, 25]. While both claimed that investigating different parts of a system at a time cannot produce a reliable general explanation of it, they exhibited irreconcilable differences. Vitalism assumed an invisible force (*entelechy* or *vis essentialis*) distinguishing inanimate entities from living organisms, while organismic biologists believed that the key was in the organization of a system and refused to accept invisible forces as a prerequisite for understanding natural phenomena[5, 16, 24, 26]. Vitalism was supported by notable philosophers and biologists such as Hans Driesch[27], but was abandoned due to the reliance on a nonmaterial force to describe life[26, 28]. On the contrary, organismic biology was integral to the emergence of *systems thinking*[5]; initially through Lawrence J. Henderson (1878-1942), a biochemist and supporter of organicism, who depicted biological entities as systems, outlined their ability to self-regulate to facilitate their survival and explored their interactions with the environment[29]. An unexpected and yet significant contribution in favor of incorporating a holistic methodology in biology came

from the field of psychology in Germany and specifically the "*Gestaltpsychologie*" movement[5]. The founder Max Wertheimer (1880-1943), along with his collaborators Wolfgang Köhler (1887-1967) and Kurt Koffka (1887-1941), focused on their subjects' perception and argued that it is not the individual parts of a whole that determine its behavior; rather their interactions within it (in this case, the brain)[30, 31].

While outside the scope of this thesis, it is worth noting that the shift away from reductionist approaches was heavily aided by the groundbreaking discoveries taking place in the field of physics at that time. The first decades of the twentieth century were marked by the conceptualization of quantum mechanics and signaled the transition from the deterministic view of the universe in the classical Newtonian approach to the statistical laws calculating the probabilities of events within a system on an atomic and sub-atomic level[5, 22].

Before elaborating on the establishment of Systems Theory in the field of biology by Ludwig von Bertalanffy, it is imperative to mention (the arguably lesser known) Aleksandr Bogdanov and his theory of tektology[32]. Bogdanov, in his quest to decode social structures, declared that all phenomena are governed by organizational laws and promoted the study of a system both through the relationship of its parts and within the context of its environment[33]. Tektology preceded von Bertalanffy's *General System Theory* (GST) by 30 years, however the latter found recognition and was deemed necessary to achieve a greater understanding of living organisms[34]. Von Bertalanffy's intention as an organicist was to disconnect biology from both vitalism and mechanism[26, 32]. His published work on GST spanned more than two decades and focused on the need to adopt a holistic approach. A summary of GST is provided in his book *General System Theory* (p.37): "*General system theory is a general science of 'wholeness' [..] In elaborate form it would be a mathematical discipline, in itself purely formal but applicable to the various empirical sciences. For sciences concerned with 'organized wholes', it would be of similar significance to that which probability theory has for sciences concerned with 'chance events'*"[35]. Throughout his body of work, von Bertalanffy vehemently opposed the dominance of physics in the field of biology and maintained that biological systems should be researched separately[5, 36]. Failing to reconcile the second law of thermodynamics with the maintenance and stability of biological systems, he developed a theory known as *Fließgleichgewicht*[36] which stated that, contrary to physical

systems, biological systems are "*open*" due to the flow of matter and energy between them and their environment, allowing them to exist far from equilibrium[5, 36-39].

When examining the rise of a systemic way of thinking, it is necessary to mention the impact of Jan C. Smuts. In his book *Holism and Evolution* (1926), Smuts heavily criticized the separatist, reductionist approach dominating nineteenth century natural philosophy and insisted on the need to adopt a holistic strategy[4]. He argued that despite the establishment and acceptance of Darwinism, mechanistic approaches were still prevalent. It's in this work that the term *holism* appears for the first time[3] as a starting point for describing the universe (p. 88): "*This character of 'wholeness' meets us everywhere and points to something fundamental in the universe. Holism […] is the term here coined for this fundamental factor operative towards the creation of wholes in the universe*"[4]. Despite initially presenting holism as the stark opposite of mechanism (and thus reductionism), he concluded that both approaches carry merit and are useful, with holism being fundamental and able to incorporate mechanism. Considering the complexity of biological structures, in many cases it is necessary to break them into smaller pieces in order to explore and explain them. However, Smuts argued, we should not be tempted to characterize this strategy as mechanistic but rather evolutionary[4].

Parallel to the development of GST by von Bertalanffy, the field of cybernetics began to flourish. The term was established by Norbert Wiener in 1948 who defined it as the "*control and communication in the animal and the machine*"[40]. While an intersection with biology might have seemed counterintuitive at that point, cyberneticists introduced certain principles and theories that took center stage in the later development of systems biology. In particular, Wiener recognized the importance of the organizational aspects governing biological phenomena such as feedback and control that are necessary to advance the understanding of living organisms[5]. The proposed cybernetic model comprised of two entities connected to each other: the regulated system and its controller[40, 41]. The system produces an output which represents the variable in need for regulation and receives as input a reference that is necessary to be maintained in order to produce the wanted output. Taking into account that every living system is subject to perturbations, the existence of a control mechanism registering the differences between the reference input and the actual output is essential, in order to inform the controller about necessary changes to the system's

response[40, 41]. This control mechanism is known as a negative feedback loop and is considered a homeostat in cases where the reference input must remain the same[40]. On the contrary, a positive feedback loop seeks to amplify the change between the reference input and the system's output[40-42]. Less than two decades later, feedback regulation as described by the cyberneticists would be used by François Jacob and Jacques Monod to explain the molecular framework of gene regulation[43, 44] and ultimately would be considered key to the self-organization (and consequently the self-maintenance) of an organism[42].

Researching living organisms as wholes requires accounting for the relationships and interactions between their parts, indicating their high complexity which increases even further when the focus shifts to entire communities[45]. A shift towards a holistic examination of biological systems mandated a general definition of a system's properties, which entailed decoding its internal organization and the patterns emerging from it. The identified patterns should therefore be representative and descriptive of each examined system[45, 46].

1.1.2 The debate moves to the field of biology: from Systems Theory to Systems Biology

The general definition of a system depicts it as a set of its individual parts and their interactions[47]. Considering the fact that systems theory focuses heavily on organization and the resulting behavior, it is only reasonable to conclude that systems biology - the intertwining of systems theory and biology - seeks a systemic (or holistic) understanding of concepts such as the structure of proteins or metabolic pathways and the connections between them[43]. The inherent interdisciplinary nature of such an approach can expand our understanding of living systems significantly. In comparison to molecular biology which focuses more on the individual low-level components, systems biology accounts for a high number of interacting partners and their relationships[3, 48]. This is the main reason molecular biology has been regularly labelled as reductionist[22, 48]. On the other hand, a key caveat in any systemic approach is the inability to predict the properties of a system solely from its components, as well as the behavior of the individual components through different states, which might be controlled systemically[48].

It is important at this point to briefly discuss reductionism within the context of biology. Biological reductionism belongs in one of three categories: *ontological*, *methodological* or *epistemological*[28]. Ontological reductionism suggests that all living entities are made up by the same inorganic materials, limiting them essentially to an aggregation of molecules or atoms[28, 49]. Questions regarding the structure of organisms belong in this domain, whose most radical form encloses the argument between mechanism and vitalism. The practical aspect of reductionist research falls under methodological reductionism[48, 49]. In cases where a choice has to be made between the study of underlying causes (lower level) or a higher organizational level of living organisms, this approach favors the investigation of the lower levels[28]. Methodological reductionism can aid the investigation of a complex system greatly by simplifying it through fragmentation and examination of its core processes. Historically, methodological reductionism has been closely associated with Francis Bacon as well as Descartes[3]. The last type of biological reductionism (epistemological) suggests that biological laws are essentially subcategories of physical laws[22, 28]. In general, epistemological reductionism occurs when two scientific fields partially overlap, pointing towards a potentially universal set of laws or theories[3, 28]. While appearing similar to methodological reductionism, the two should not be confused.

The contribution of all three types to science, and particularly, biology throughout the centuries has been of utmost significance. For example, the application of already existing theories to the explanation of biological phenomena by epistemological reductionism has uncovered the existence of patterns in nature[28]. Molecular biology approaches are inherently reductionist and belong to the methodological domain, due to the dependence on *decomposition* (focus on the processes taking place in a system's components), which indicates the existence of levels within the system, and *localization* (connecting the identified processes to the internal structure)[48]. Considering the above, it is evident that reductionist approaches would not succeed in explaining complex nonlinear systems such as the human brain or ecological and social structures[22, 50]. Furthermore, the exponentially increasing and highly diverse -omics data exposes the complexity of biological systems[22]. These datasets as well as the computational tools currently available have made possible the application of *recomposition* in biological research. *Recomposition* involves the acquisition of all necessary information in order to identify the interactions between a

system's components and investigate whether its behavior can be modelled[48, 50]. According to Hiroki Kitano, there are four factors to consider when attempting to understand and describe a biological system[51]:

i. The structures present within a system and their modus operandi
ii. The behavior of a system in response to fluctuations
iii. The mechanisms stabilizing the system
iv. The elected strategy for the representation of the biological system

1.1.3 Properties of biological systems

The non-linearity of biological systems indicates that small changes in their components can result in disproportionately large effects in other parts of it[5, 22]. Stuart Kauffman characteristically states that complex systems "*exist on the edge of chaos*" to emphasize the occurrence of sudden changes in response to a small perturbation[52]. Therefore, the ability of a biological system to remain stable through environmental fluctuations is imperative for its survival.

Cyberneticists already emphasized the importance of feedback loops with regards to self-maintenance. A classic example of a necessary negative feedback in a biological system is temperature regulation; when the environmental temperature induces a change beyond the acceptable limits in homeotherms, control mechanisms such as shivering are activated to diminish the effect of the disruption and stabilize the system[41, 53]. Generally, homeostatic mechanisms rely on negative feedback loops. As mentioned in the previous section, positive feedback loops act in the exact opposite manner. For example, the levels of oxytocin released during childbirth in humans increase constantly in response to signals from both the central nervous system and the pressure the fetus places on the cervix in order to aid with the process, as well as post birth to help the uterus revert to its normal size[54]. Further instances of positive feedback loops in nature have been observed in the nesting behavior of the great blue heron and the bluegill sunfish; specifically the preference of both for constructing central colonies due to the benefits they receive for food resources in the case of the former and the higher vulnerability of peripheral colonies in the case of the latter[42, 55, 56]. It is evident that positive feedback loops can escalate, with catastrophic effects to a system, similar to a snowball increasing in size and gaining momentum. Therefore, negative feedback as well as external constraints (e.g., crowding) manage or

antagonize positive feedback when boundaries need to be introduced[42]. The necessary communication and flow of information required for this type of regulation is a characteristic feature of dynamic systems[42, 57].

The survival of a biological system is tied to its ability to self-organize[58]. The term *self-organization* refers to the formation of higher-level patterns through interactions taking place in the lower levels of the system using information specific to each (low) level[42]. The most important development regarding the self-organizational aspect of biological systems came from the work of Ilya Prigogine (1917-2003), who established the term *dissipative* to describe the appearance of structures in systems existing far from equilibrium, following non-linear kinetics[42, 59, 60]. The patterns generated through these non-linear interactions consist emergent systemic properties and are characteristic of complex systems[5, 42]. The term *emergent* is used to describe their absence from the individual components (lower levels) of a system; they *emerge* during the processes taking place and lead the way from simplicity to complexity[42]. The resulting complexity requires the existence of some internal hierarchical organization within the system. It also implies the existence of level-specific laws that are absent in lower levels as a consequence of emergence, making the study of the whole system based solely on general features impossible[5, 22]. This reaffirms Smuts' conclusion that neither reductionism nor holism are enough on their own - in most cases both approaches are required to adequately represent a complex system[4].

Patterns are not necessarily indicative of a self-organized system. To assume self-organization simply by the presence of a pattern would be a problematic conclusion. The identification of the internal mechanism responsible for the emergence of a pattern is crucial, as is disrupting the system or interfering with internal processes to prove that self-organization is responsible for its stability and not external contributors[42]. Small perturbations can result in a transition from the pattern present under standard conditions to a new one through bifurcation, a process observed regularly in self-organizing system with positive feedback[42]. Taking into account the manifestation of large behavioral changes as a response to small alterations in the parameters of biological systems, the impact of adaptable emergent properties is profound, especially from an evolutionary standpoint.

1.1.4 Networks in Systems Biology: a bridge between local and global approaches

Identifying the organization of a system and its properties is one of the main concerns in Systems Biology. Encoding a system as an extensive network of interactions is a powerful tool that can provide important insight into the inner workings of its subsystems[61]. Depending on a variety of factors such as the design, the choice of model, the quality of the data and more, the patterns emerging through a network representation can reflect systemic ones, thus enabling the formulation of hypotheses about the system's internal mechanisms. Additionally, the organization and structure of a system can easily be described through a network.

Although networks belong to graph theory in mathematics, they are implemented in a variety of research fields. In biology, networks have found applications in the representation of protein-protein interactions, metabolic pathways, gene co-expression and more[62, 63]. The general description of a network or graph is a set of elements (nodes or vertices) and their connections (edges). Going by its mathematical notation, a graph G is a pair of (V, E) where V denotes a set of vertices and E a set of their edges[63]. There are several types of graphs, however the most widely used are undirected, directed, weighted and bipartite graphs and will be briefly discussed below.

Undirected graph: The simplest form of a graph in which nodes are connected with a single edge (*neighbors*)[63]. A gene co-expression network is a representative example of this type of graph.

Directed graph: The connection (edge) between two vertices i and j is an arrow, demonstrating direction and, in a biological context, possibly regulation or inhibition as well[63]. From the pair of interconnected nodes, one is considered the "head" and the other the "tail", with the edge directing a one-way connection from the tail to the head vertex[64]. A directed graph can be denoted as an ordered $G = (V, E, f)$, where f represents a function associating each edge in E to an ordered pair of nodes in V[63]. Metabolic pathways as well as regulatory networks belong in this category[62].

Weighted graph: here the set of edges E between two vertices is associated with a weight function $w: E \rightarrow R$, where R represents all real numbers[63]. The weight w_{ij} of an edge connecting nodes i and j usually reflects a feature of the connection between the nodes - for example a larger value indicates a more reliable or a stronger connection.

Weighted graphs are currently the most popular networks in biology and bioinformatics and are used mainly for exploring similarities between sequences or structures[62, 63].

Bipartite graph: an undirected graph occurring when the vertices V are divided in two sets V' and V'' and connections between vertices of the same set are prohibited, so that each vertex from V' can only be connected to vertices in V'' and vice versa[63]. Bipartite graphs have been extensively used to describe ecological relationships[62, 63].

Given the fact that networks may represent complex biological systems, understanding their organization is crucial. Every graph possesses certain general properties that can shed light on the entity represented through it. One of the most important and defining graph features is the degree distribution $P(k)$[63]. A prerequisite for this measure is knowing the degree of a node which is equal to the number of edges associated with it. If k denotes the number of connections a vertex has, then the degree distribution is the probability of k being equal to the degree of a random node in the network[63]. In cases where a network follows a *power-law* distribution, it is considered *scale-free*, indicating that network nodes are either hubs with a high number of connections, the removal of which will alter the topology of the network significantly, or nodes with only a few edges whose removal will not have a great impact[63]. Many networks that model real systems, such as social or biological networks are scale-free[65]. Other highly informative network properties include graph density and the clustering coefficient which represent the ratio of present edges to the number of all possible edges and the tendency of network nodes to form clusters respectively[63]. The last important graph feature is the distance between two nodes, a measure that corresponds to the shortest possible path, i.e., the least number of edges needed to reach one node with the other as starting point[63].

The main purpose of a graph representation is to determine whether the observed properties are characteristic to the network and consequently the system described[61]. Several models have been proposed to describe the structure of a graph and the connections within it, with the most widely known ones being the "*Erdos-Rényi*", the "*Watts-Strogatz*" and the "*Barabási-Albert*" models[63]. The *Erdos-Rényi* model was introduced for the description of random graphs whose main characteristic is their homogeneity, deeming them inappropriate to represent biological systems[66]. The *Watts-Strogatz* model was developed for networks with a high clustering coefficient and on average short distances between nodes, such as metabolic pathways[67]. Lastly, the

Barabási-Albert model describes scale-free networks that are capable to evolve through time[68]. In this model, the network has the ability to expand by adding new nodes in a manner that agrees with the established degree distribution and without the random addition of edges. This model is highly representative of biological systems, with protein-protein interaction networks being a classic example of it[63, 65].

The progress of the -omics approaches in biology has increased the amount of data produced exponentially. The variety of sequencing data that can be easily acquired (genomics, transcriptomics, etc.) has the potential to provide valuable insights regarding the structure as well as the regulation of biological systems, with several existing cases attesting to that. For instance, Ideker *et al* integrated existing data to a network representing the utilization of galactose in the yeast, then proceeded to disrupt each component of the metabolic pathway, registered their effect through transcriptomics to adjust the network and investigated whether the protein levels reflected the mRNA response[69]. Arrieta-Ortiz *et al* integrated 151 publicly available transcriptomes of *Clostridioides difficile* to a gene regulatory network that was able to uncover the adaptation of the pathogen through the infection of the host and make predictions regarding its interactions with the established intestinal commensals[70]. Systemic approaches to cancer research have gained significant ground as well, as they allow the examination of large, diverse cohorts simultaneously. For example, Zheng et al combined data from breast and head-and-neck cancer screens as well as multi-omics data from other studies to construct a network based on the interactions between cancer-related proteins, uncovering a hierarchical system which resembled a map of smaller protein complexes within a larger group that ultimately represented cancer-specific mutations[71].

The list of studies that adopted systems biology approaches to research complex systems, especially in the last 20 years, is extensive. In most instances, they involve the integration of datasets that are both diverse and, in certain cases, multi-terabyte in size with the end goal of producing an easily interpretable result; a feat that may ultimately prove challenging. As described in this section, networks have the capacity to incorporate and process a wide range of data, given the correct choice of model. Their application in biological research has already provided a greater understanding of mechanisms characterized by high complexity such as protein-protein interactions, metabolic pathways, gene co-expression, phylogenetic networks

describing evolutionary relationships, ecological chains, species interactions and the list goes on.

1.2 Interspecies interactions

The majority of ecosystems consist of multispecies communities, as isolated organisms are an extremely rare phenomenon[72]. The interactions within these communities shape the evolution of their component species and drive their adaptation[73]. The complex networks formed through interspecies interactions suggest that fluctuations of one species in the system, regardless of their initial magnitude, may have significant ramifications on the rest. Therefore, traits such as the survival or reproduction of species are influenced both by their own genetic properties as well as the ones of their interacting partners[72].

Ecological interactions can be classified into the following six groups based on their effect (beneficial, neutral or harmful) on the interacting organisms[74]:

Mutualism: both interacting partners benefit from the relationship[75]. A key example of mutualism can be found in the symbiotic relationship between plants and species of the *Glomeromycota* fungus known as *Arbuscular mycorrhiza*[76]. Research spanning almost two centuries has shown that the fungi promote plant growth by colonizing the plant roots and releasing inside them the nutrients they have absorbed from the soil while taking advantage of a microenvironment rich in nutrients[76].

Neutralism: species interact without a significant impact, for instance by coexisting in the same community[77].

Commensalism: one interacting species receives a benefit while the other one remains unaffected[77]. Epiphytes are a prime representation of this type of interactions. They establish themselves and grow on host trees in order to receive a benefit such as support or access to increased sunlight without a positive or negative impact on their host[78, 79].

Amensalism: one species is disadvantaged through the interaction and the other remains unaffected[77]. An instance of amensalism is observed in cases where the development of a plant is inhibited due to toxic chemical compounds released in the soil by a neighboring plant ("*negative allelopathy*")[75].

Antagonism: a species gains an advantage at the expense of another[77]. Antagonistic interactions include predator-prey and host-parasite relationships[80-82].

Competition: a number of species compete for the same limited resources, possibly affecting their fitness negatively[77]. When entering the intestine of a human host, food-borne pathogens such as *Salmonella enterica* trigger an immune response that they can survive due to the fact that they must outcompete the resident microbial population for resources[75, 83].

While these definitions are crucial for navigating the complex network of interactions in ecological communities, it should be noted that reassignment of individual cases between different categories is a common phenomenon. Inter- as well as intra-specific relationships within a community are multifaceted and sometimes the lines between certain types of interactions can be blurred. Furthermore, the continuous acquisition of new data as well as the improvement and reanalysis of existing datasets can shed new light and offer novel perspectives.

The last decades have seen a stark increase in research regarding the ecological interactions between eukaryotes and microbes, particularly humans and their associated microbiota, which is defined as the community of colonizing microorganisms within a niche[84]. The symbiotic relationship between a host and their microbiota usually falls under the umbrella of mutualism or commensalism[85]; however, it has been shown that the established microbes can enable parasitism by a third party at the expense of the host, as in the cases of an infection by parasitic nematodes and their establishment in mice and *Drosophila neotestacea*[86]. The microbes and their genetic properties, commonly known as *microbiome*[84], are instrumental for a variety of physiological processes, ranging from digestion to immune responses, and the overall homeostasis[84, 87]. The host-microbe system is regularly referred to as the *holobiont*[88] and it encases a variety of interactions between the interacting partners such as the synthesis of amino acids and vitamins which mediate both host-microbe and microbe-microbe interactions within this system[89]. Representative examples include lantibiotics produced by the commensals *Streptococcus salivarius* (oral cavity), *Staphylococcus epidermidis* (skin), and the intestinals *Ruminococcus gnavus* and *Clostridium nexile* which prevent the colonization of pathogens closely related to them and as result, protect their host[89-91].

Currently, the most widely researched host-microbe system is the commensal microbial community in the human gastrointestinal tract[84]. Research has shown that the composition of the gut microbiome may vary among individuals as it is influenced

by a multitude of factors[92]. Its significance regarding human biological processes has been highlighted[92] and perturbations altering the community makeup have been linked to a variety of multifactorial metabolic disorders such as obesity, diabetes, fatty liver disease and more[87, 93]. Additionally, the metabolic potential and capacity of the gut microbiome allows direct interactions with drugs and influences their metabolism in many ways, for example by competing with them for molecules or by producing enzymes that degrade them[94].

The increasing interest towards the human microbiome and its biomedical potential has shifted the attention even more towards the general interactions of multicellular organisms and their microbial environment. Gaining novel insights and decoding interspecies interactions besides the human-microbe system can only serve as a valuable resource and enrich our understanding of the dynamics and evolutionary trajectories among interacting partners in a variety of ecological communities.

1.2.1 The nematode-microbiota system as a model to study interspecies interactions

Nematodes, commonly known as *roundworms*, comprise important components in the majority of biospheres on the planet due to their abundance[95]. Soil nematodes in particular are core elements in their respective ecosystems worldwide and are linked to complex and highly diverse microbial communities[96, 97]. In addition to their indispensable ecological role, they are also vital in varying fields of research, ranging from ecology to neurobiology due to certain properties they possess. Specifically, their short generation time, their mode of reproduction, the ability to maintain them in monoxenic cultures and the molecular toolkits available make them a convenient, malleable system and contributed significantly to the establishment of the nematode *Caenorhabditis elegans* as a model organism for human-related research[98].

Despite the abundance of studies, especially involving *C. elegans*, very little is known about nematode-associated microbes and their impact on the worm. The *C. elegans* N2 strain - used in the majority of experiments - has been consistently isolated from its natural microbiota and maintained in monoxenic cultures consisting mostly of *Escherichia coli* OP50, while simultaneously tolerating treatments that eliminate its associated microbiome[99, 100]. However, the ecological interactions identified through the study of wild isolates have shed light on the overlooked associations with its

environmental microbiota, leading to the identification and functional characterization of the associated bacterial communities[101-105]. Further investigation of the nematode's transcriptomic response to intestine-colonizing bacteria has revealed incredibly interesting species-specific signatures with regards to intestinal bacterial strains[99, 106].

In spite of the fame that *C. elegans* has enjoyed, there is a wide variety of characterized nematodes exhibiting traits that can open new avenues of research, particularly in the context of worm-microbiota interactions. *Pristionchus pacificus* is a free-living, soil-dwelling nematode, found to exploit scarab beetles as its host[107]. Initially developed for comparative research with *C. elegans* regarding the vulva formation, *P. pacificus* was ultimately established as a model for phenotypic plasticity due to a mouth-form dimorphism it presents[108]. In particular, it can develop either a narrow buccal cavity with a single tooth called *stenostomatous* (St) that constitutes the worm a bacterivore or a wider one with two teeth called *eurystomatous* (Eu) that allows it to predate on other nematodes[109]. *P. pacificus* is a representative case regarding the impact of ecological interactions, as it has been shown that the mouth polyphenism is influenced by a variety of environmental factors, such as temperature, diet, culture type and more[110-113]. Additionally, compared to *C. elegans*, *P. pacificus* has exhibited higher tolerance to toxin-producing bacteria which can potentially allow for a richer bacterial microbiome[114, 115]. The synergy within its associated microbial community, observed through the diminished negative effect of native pathogenic bacteria in mixtures with non-pathogens[114], hints at the level of complexity characterizing the interspecies interactions which include the nematode. Despite the diversity present in the worm's natural environment, our knowledge so far is nematode-centric, focusing mostly on the impact of a diet on life-history traits or their effect on the mouth-form[114, 116, 117].

It is clear that the interplay within the intricate insect-nematode-microbe system is not fully decoded, especially with regards to the microbiota of the worm. While we have gained important insight about the dietary effect in general, we have not specified bacterial triggers acting as modulators of entire regulatory networks in *P. pacificus*. With this dissertation, I will illustrate an effort towards this direction. Starting with the characterization of bacterial strains previously isolated from the worm's natural habitat[118] and the identification of their metabolic potential, I predicted bacterial metabolic signatures; that is the variation of the metabolic potential between the

different bacterial genera processed. I leveraged this information in order to associate the bacterial metabolic potential with nematode traits quantified in previous studies[118]. This allowed me to narrow down the bacterial metabolites with a potentially strong impact from thousands to a few tens. The associations with nematode survival indicated a strong negative effect of paerucumarin and rhabduscin, a finding that I confirmed by supplementing available coumarin derivatives to the worm. By acquiring dietary transcriptomic profiles of *P. pacificus* and using them to construct a gene co-expression network, I organized the almost 29000 genes present to 8150 clusters. Performing functional enrichment analysis of the network's modules led me to the identification of environmentally responsive regulatory pathways in the nematode. Lastly, I combined the bacterial metabolic potential and the modules in a variation of a bipartite network in order to estimate the interactions between the two datasets and identified the network modules potentially regulated by diet-derived metabolites.

2. THESIS AIMS

The main goal of this dissertation was to investigate the interspecies interactions between the nematode *Pristionchus pacificus* and bacterial food sources found in its natural environment. To achieve this, I formulated three questions that are crucial in order to understand the complex relationship between *P. pacificus* and the bacteria: (i) what information can be harvested from bacterial genomes regarding their metabolic potential, (ii) how can transcriptomic profiles of the worm from different diets reveal environmentally responsive regulatory networks and lastly (iii) how are they associated to the bacterial metabolic potential. To find the answers, I carried out the following tasks:

1) I sequenced, assembled and annotated the genomes of 84 bacterial strains isolated previously from *Pristionchus*-associated environments and used them to predict their metabolic potential as well as the phenotypes present.
2) I reconstructed the bacterial phylogeny and compared it to the metabolic potential which uncovered the variation regarding the metabolic pathways present and, based on that, a metabolic signature of different bacterial genera.
3) I associated the presence/absence of bacterial metabolic pathway groups (MPGs) with available data regarding nematode traits such as survival and chemoattraction to a food source. As a result, I was able to decrease the number of bacteria-derived metabolites with a significant effect on survival and chemoattraction from thousands to 40 and 24 respectively. I validated the predicted negative effect of paerucumarin on survival through supplementation of coumarin derivatives to the nematode.
4) I produced transcriptomic profiles of *P. pacificus* on 38 diets from the bacterial phylogeny and created a gene co-expression network. By performing enrichment analysis of the modules with a variety of available transcriptomic datasets and functional annotations, I assigned functional labels to the network modules and discovered regulatory networks in the nematode.
5) I combined all the generated data in a bipartite network and linked the responses of the coexpression modules to specific MPGs. Using this approach, I determined the environmentally responsive co-expression modules as well as the MPGs acting as potential triggers.

3. RESULTS

3.1 Crowdsourcing and the feasibility of manual gene annotation: A pilot study in the nematode Pristionchus pacificus

Rödelsperger C, **Athanasouli M**, Lenuzzi M, Theska T,S un S, Dardiry M, Wighard S, Hu W, Sharma DR, Han Z

Scientific Reports 9, 18789 (2019). DOI: 10.1038/s41598-019-55359-5

3.1.1 Synopsis

The ability to decode the genome of an organism, particularly a novel one, relies heavily on ortholog detection in already established species. This is necessary in order to gain knowledge about the biological processes taking place and develop a molecular toolkit. However, broadening our knowledge as well as creating functional annotations can be hampered by inaccurate gene annotations. This is an issue often encountered outside of the study of model organisms such as *C. elegans* where the application of prediction software is accompanied by a large number of experimental approaches that aid functional annotation.

To improve the latest *Pristionchus* pacificus' gene annotation, we devised an approach where evidence from transcriptomic and comparative genomic data was coupled with manual, community-based revision of gene models we deemed suspicious in the existing annotation. We observed that most cases in need of correction consisted of gene predictions that resulted either from the artificial fusion of two gene models or their complete absence. Our methodology raised the number of *P. pacificus* genes to 28036 from the previously predicted 25517 and improved the completeness of single copy orthologs significantly from 86% to 97% while simultaneously demonstrating the power of community curation.

3.1.2 Own contribution

I manually curated approximately one third of the candidate genes in the study and trained other curators. I estimate my contribution at 20%.

3.2 Comparative genomics and community curation further improve gene annotations in the nematode *Pristionchus pacificus*

Athanasouli M, Witte H, Weiler C, Loschko T, Eberhardt G, Sommer RJ, Rödelsperger C

BMC Genomics 21, 708 (2020). DOI: 10.1186/s12864-020-07100-0

3.2.1 Synopsis

Our previous work initiated the improvement of the *P. pacificus* gene annotation and revealed the power of community-based curation when combined with a comparative genomics approach. While our first analysis targeted genes with suspected fused or completely absent gene models, this study investigated one-to-one orthologs with *C. elegans* that were suspiciously larger or smaller in length. The gene models in need for inspection were identified by comparing protein lengths between the orthologs as well as genes exhibiting unusual combinations of PFAM domains that are absent in *C. elegans, Bursaphelenchelus xylophilus* and *Strongyloides ratti*. The analysis resulted in 4311 *P. pacificus* gene models in need of revision, of which 1367 were corrected and 1579 were removed. These findings were incorporated to a new version of the existing gene annotation, totaling 28896 genes and a BUSCO completeness of 97.6%, which was slightly higher than the previous version.

3.2.2 Own contribution

I performed part of the computational analysis and wrote part of the manuscript. I estimate my contribution at 30%.

3.3 Analysis of repeat elements in the *Pristionchus pacificus* genome reveals an ancient invasion by horizontally transferred transposons

Athanasouli M, Rödelsperger C

BMC Genomics 23, 523 (2022). DOI: 10.1186/s12864-022-08731-1

3.3.1 Synopsis

Eukaryotic genomes consist of large fractions of repetitive DNA sequences, which vary between species and often stem from transposable elements. Generally, repetitive elements prove problematic during genome assemblies and gene annotation efforts. Due to the fact that transposons were part of what was termed “junk DNA” until fairly recently, we lack knowledge about their role in eukaryotes and even more so regarding their impact on the evolution of nematodes.

While the quality of the current *P. pacificus* genome is high, we are lacking insight on many of its features. In this work, we wanted to explore the repetitive aspect of the nematode’s genome. During the first steps of our analysis, we evaluated 12 different software tools for the detection and/or annotation of the transposons detected as well as their agreement on the predictions. We found that RepeatModeller2 exhibited the most common patterns with the majority of the approaches and suited our needs best. Using the RepeatModeller2 annotations, we observed that retrotransposons were the most abundant class while available transcriptomic evidence suggested potential activity of DNA transposons and LINEs. Interestingly, RepeatModeller2 indicated the presence of 58 transposons characterized as *Zisuptons*. Phylogenetic analysis of the Zisupton sequences revealed the process of horizontal transfer as the means of acquisition. Overall, this was the first comprehensive analysis of transposable elements in the worm and enriched our knowledge on their presence as well as their emergence in the *P. pacificus*’ genome.

3.3.2 Own contribution

I conceptualized the study. I performed the majority of the computational analysis, curated and interpreted the data. I visualized the results and wrote the majority of the manuscript. I estimate my contribution at 80%.

3.4 A New Hope: A Hermaphroditic Nematode Enables Analysis of a Recent Whole Genome Duplication Event

Wighard S, **Athanasouli M**, Witte H, Rödelsperger C, Sommer RJ

Genome Biology and Evolution 14:evac169 (2022). DOI: 10.1093/gbe/evac169

3.4.1 Synopsis

Whole genome duplication (WGD) involves the duplication of the entire genome of an organism which can define the evolutionary trajectory of species. Duplicated genes (ohnologs), while initially redundant, may either assume new roles or be silenced. This study describes for the first time *Allodiplogaster sudhausi*, a nematode belonging in the same family as *P. pacificus* and exhibiting multiple mouth-forms as well, but with an impressively larger body size. Whole genome sequencing combined with karyotyping and comparative analysis including two closely related *Allodiplogaster* species showed that the nematode has undergone a WGD specific to *A. sudhausi*. A large number of ohnologs were found to be under evolutionary constraint and certain protein domains and metabolic pathways such as homeodomains and the mTOR pathway were favored in terms of retention post-WGD. Analysis of the transposon landscape of *A. sudhausi* and comparison with *P. pacificus* as well as *C. elegans* revealed the significantly higher number of repeat elements present in *A. sudhausi* and specifically the overrepresentation of the transposon family DNA/hAT-A. Subsequent gene editing experiments with single or double mutants for pair of ohnologs and assessment of the resulting phenotypes demonstrated a type of redundancy associated with WGD.

3.4.2 Own contribution

I performed the computational analysis regarding the transposable elements. I estimate my contribution at 5%.

3.5 Thousands of *Pristionchus pacificus* orphan genes were integrated into developmental networks that respond to diverse environmental microbiota

Athanasouli M, Akduman N, Röseler W, Theam P, Rödelsperger C

PLoS Genetics 19(7): e1010832 (2023). DOI: 10.1371/journal.pgen.1010832

3.5.1 Synopsis

Compared to *C. elegans*, *P. pacificus* contains a high amount of taxonomically-restricted (orphan) genes. The nematode's initial genome assembly indicated that approximately one third of its genes can be classified as orphans and subsequent analysis involving nine more diplogastrids as well three outgroup species was able to distribute all *P. pacificus* genes in 14 gene age classes. The lack of homologs in other species renders their functional annotation almost impossible without extensive experimental investigation. There are many theories regarding their emergence, with some attributing their appearance to the need for environmental adaptation.

In this work, we're providing a first attempt to determine whether orphan genes have integrated in regulatory networks as well as assign them a functional label. To achieve this, we created a gene co-expression network from transcriptomic profiles produced by growing the nematode on 24 diets and processed the network modules containing more than 50 genes. We observed that the co-expression modules exhibited both distinct expression patterns across the different diets and compared to other modules as well as distinct regulatory motifs. Through enrichment analyses with both transcriptomic data from previous studies as well as functional annotations from databases such as KEGG and PFAM, we were able to assign labels to 22 out of the 28 modules. However, the most interesting finding consisted of the 3727 orphan genes we were able to classify as environmentally responsive, indicating that their emergence can be linked to necessity.

3.5.2 Own contribution

I performed the computational analysis in this study, interpreted and visualized the results and wrote a significant portion of the manuscript. I estimate my contribution at 70%.

3.6 The genome assembly of *Rhabditoides inermis* from a complex microbial community reveals further evidence for parallel gene family expansions

Rödelsperger C, Röseler W, **Athanasouli M**, Wighard S, Herrmann M, Sommer RJ

bioRxiv (2024). DOI: 10.1101/2024.08.02.605984

3.6.1 Synopsis

Nematodes consist the most abundant group of animals worldwide. They are found in a variety of ecosystems and are used consistently in biological research. Studies have estimated that the majority of nematodes are not yet isolated and characterized, suggesting constant reevaluations of the existing phylogeny. A correct and consistent phylogeny is crucial in order to follow the evolution of genes. The *Diplogastridae* family, where *P. pacificus* belongs, has been under revision the last few years. Research has been particularly conflicted about the sister species to the family, with *Rhabditoides inermis* - a beetle-associated nematode - appearing to be crucial for its determination.

To gain an insight on the nematode phylogeny, we obtained and sequenced *R. inermis* from the beetle *Nicrophorus vespilloides*. Genome assembly and evaluation indicated that some of the constructed scaffolds belonged to microbes other than the standard diet of Escherichia coli OP50, which after further investigation, led to the assembly of a complete *Vanrija albida* genome. Genomic and karyotyping analysis resulted in the identification of five chromosomes with a high percentage of completeness. Using the transcriptomic and protein sequences of 40 species including *C. elegans* and *P. pacificus*, we reconstructed the nematode phylogeny and showed that the misplacement of *R. inermis* in previous studies as a sister group to *Diplogastridae* was due to the use of 18S sequences instead.

3.6.2 Own contribution

I performed the computational analysis relating to the microbial genomes. I estimate my contribution at 5%.

3.7 High nutritional conditions influence feeding plasticity in *Pristionchus pacificus* and render worms non-predatory

Piskobulu V, **Athanasouli M**, Witte H, Feldhaus C, Streit A, Sommer RJ.
bioRxiv (2024). DOI: 10.1101/2024.08.27.609904

3.7.1 Synopsis

Plasticity accommodates an organism's need to adapt to its environment and can manifest as morphological or physiological changes. Nutrition is a major contributor in the emergence of polyphenisms, regulating the organism's metabolism and responding to potential stressors. *P. pacificus* exhibits a mouth-form dimorphism that distinguishes it from other nematodes and modulates its behavior, rendering it either a bacterivore or predator towards other nematodes. Diverse environmental contributors such as temperature or culture type have been found to regulate this dimorphism. While the role of nutrition and particularly low-quality nutrients has been studied with regards to the impact on the nematode's metabolism, very little is known about their effect on the mouth-form.

This work focuses on the impact of high-nutrition conditions on fat storage and subsequently the mouth-form. Through the supplementation of various monosaccharides and fatty acids to the worm, we observed a delayed development as well as an alteration to the mouth-form. Focusing specifically on the effect of glucose and oleic acid, we noticed an increase to the fat storage by staining, which was subsequently confirmed through RNA sequencing and differential gene expression analysis. Further functional enrichment of the differentially expressed genes using the KEGG database associated them with lipolysis and lipogenesis, resulting in the identification of candidate genes. In particular, delta-9 fatty acid desaturases and peroxisomal beta-oxidation genes were assessed for their effect both on mouth-form plasticity and fat storage. The results indicated that fatty acid biosynthesis as well as the peroxisome are intrinsically connected with nutritional-associated changes in the worm's mouth-form.

3.7.2 Own contribution

I performed the bioinformatic analysis and wrote the corresponding part of the manuscript. I estimate my contribution at 10%.

3.8 Interspecies systems biology links candidate bacterial pathways to nematode gene expression, behavior, and survival

Athanasouli M, Loschko T, Rödelsperger C.

bioRxiv (2024). DOI: 10.1101/2024.10.02.616249

3.8.1 Synopsis

Interspecies interactions have emerged as one of the most important drivers of both evolution and adaptation. While host-microbiota research is usually centered around the human gut microbiome, nematodes such as *C. elegans* and *P. pacificus* serve as powerful systems for exploring interactions with the microbiota. In this study, we examined the interactions between *P. pacificus* and bacterial food sources. We analyzed the genomes from a subset of bacterial strains isolated from associated environments in previous works and determined their metabolic potential. The variation of the predicted bacterial metabolic pathways present indicated the existence of genus-specific metabolic signatures which could influence the transcriptomic response of the worm. In order to associate the observed variation with a transcriptomic response in the nematode, we generated transcriptomic profiles using 38 of the sequenced strains as a food source and incorporated them in a gene co-expression network. Enrichment analysis of the co-expression modules similar to our previous work enabled us to functionally label them and compare them to the previous co-expression network, a process which led to the discovery of environmentally responsive processes in the nematode. Furthermore, we incorporated the bacterial data and the transcriptomic profiles in a bipartite network and were able to determine the environmentally sensitive modules (and thus biological processes) as well as the bacterial metabolic pathways acting as potential triggers.

3.8.2 Own contribution

I performed the experiments, generated and analyzed the data, visualized the results and wrote the majority of the manuscript. I estimate my contribution at 85%.

4. DISCUSSION

Pristionchus pacificus is part of a complex ecological community, exhibiting strong associations both with the insects serving as its host and microbes[119-121]. The presence and abundance of nematodes on the insect carcasses as well as their dietary preferences for cofactor-producing bacteria can alter the composition of the *Pristionchus*-associated microbial community[117, 122], which in turn can have a significant impact on the worm's metabolism[114, 116, 117], exemplifying the bidirectional effect of their interactions. These findings have become a great starting point for demystifying the complex network of interactions, where *P. pacificus* is a dominant component; however, they generate even more questions about the interplay with the native microbiota.

The goal of my work was two-fold: to begin with, I aimed to identify the metabolic potential of the native bacterial community associated with *P. pacificus*, in an effort to uncover its variation across the bacterial phylogeny. Subsequently, I sought to associate the identified variation with the differential transcriptomic response of the nematode to these bacteria, in order to reduce the thousands of bacterial metabolic pathways to a small number of candidate metabolites with a strong effect on the worm. To achieve this, I initially divided the analysis between the nematode and the bacteria and at a later stage combined the data in a bipartite network to gain insight about both interacting partners.

One of the most important aspects of this study depended on the ability to recognize regulatory networks in the worm using diverse transcriptomic profiles as input. To test the robustness of the method, I first created a gene co-expression network with P. pacificus transcriptomes from 22 bacteria isolated from Pristionchus-associated environments as well as the Escherichia coli strains OP50 and HB101. I standardized the co-expression network by altering parameters and tested it via random subsampling and reconstruction with a different number of diets each time, a process which demonstrated that inclusion of a higher number of dietary transcriptomes leads to more conservative networks. This outcome ensured that the final network was an accurate representation, in spite of the high correlation between the acquired different transcriptomes. I observed that the resulting co-expression modules exhibited distinct expression profiles, a finding that, in combination with

available developmental time-course data[123], allows predictions about the developmental speed of the nematodes when grown on different bacteria.

Using the phylostratigraphic analysis from previous studies that distributed the *P. pacificus* genes in 14 age classes[124], I was able to classify the co-expression modules as either young or conserved and simultaneously uncovered the rapid integration of over 3000 diplogastrid-specific genes in regulatory networks. This discovery was particularly interesting because it has already been theorized that the emergence of novel genes may be associated to environmental adaptation[125]. Further enrichment analyses connected 22 of the 28 largest co-expression modules with biological processes that also included previously identified gene sets such as target genes of the mouth-form regulatory network expressed in gland cells[126]. Among other findings, I discovered that oogenesis-associated modules consisted mostly of conserved genes while the second largest co-expression module was young and linked to spermatogenesis, exhibiting a difference in the evolutionary trajectory of the two processes. Altogether, this work served a double purpose: it laid the foundation for the methodological approach regarding the transcriptomic data acquired through different bacterial diets and demonstrated the information this type of analysis can produce.

Returning to the investigation of the dietary sources, I obtained the genomes of 84 bacterial strains previously isolated from the worm's natural envinronment[118] and reconstructed their phylogeny. While determining the bacterial nomenclature, I noticed the low homology of three strains to already identified bacterial species in the non-redundant version of NCBI, as well as the inability to assign a specific genus to another two closely related strains. The attempt to reconstruct their individual phylogenetic trees with their best hits and available bacterial genomes of the same genus revealed the presence of previously uncharacterized bacterial species in the collection. As a next step, I used the MetaCyc database[127] to predict the metabolic pathways present in the bacterial strains based on their genomes, which constituted their metabolic potential. Grouping the bacterial metabolic pathways using common presence/absence patterns across the phylogeny allowed for dimensionality reduction, as their number decreased from 2902 individual metabolic pathways to 712 groups (MPGs). It is worth noting that, in many occasions, the MPGs reflected either similar metabolic processes or processes in succession. Mapping the MPGs to the bacterial phylogeny revealed the underlying variation, with strains from the same

genus sharing significantly more MPGs compared to bacteria belonging in different genera, an observation hinting at genus-specific signatures in the bacteria.

From the 84 sequenced strains, I supplemented 38 to *P. pacificus* as a monoxenic food source to acquire transcriptomic profiles of the worm, which I subsequently used as input to a gene co-expression network. By applying the previously published methodology, I was able to functionally annotate 50 of the 60 modules containing more than 20 genes. Furthermore, I observed similarities between the two co-expression networks, revealing groups of genes that cluster together regardless of the conditions, as was the case with the astacin/mouth-form/gland cell module. The combination of the co-expression modules and the bacterial MPGs in an adaptation of the bipartite networks commonly used in ecology enabled me to identify the environmentally responsive regulatory networks in the worm as well as their potential "triggers" (MPGs). The bipartite network also revealed that intestine-associated modules appeared to be the most sensitive to different bacterial MPGs, placing the intestine in the limelight and exhibiting its importance.

As mentioned previously, this work is intended to start a conversation and more importantly an exploration around the broader spectrum of complex interactions taking place in the *Pristionchus pacificus* niche. Naturally, a study like this one comes with certain caveats. Approximately one third of the dietary bacteria attempted had a strong negative effect on the nematode, making it impossible to retrieve sufficient worms/RNA which could have resulted in stronger separation between the different transcriptomic responses. Additionally, while the monoxenic cultures present a great advantage in this case, they fail to depict the dynamic relationships within a bacterial community, especially considering the effect of *P. pacificus*' preferences on it as Lo et al have shown[117].

Nevertheless, the work presented in this thesis shows the success of a strategy incorporating diverse bacterial and nematode data towards a common goal. More importantly, the ability to narrow down thousands of metabolic pathways and their products to less than a hundred and isolate the groups of responsive genes has the potential to detect previously undiscovered interactions with the microbiota. Such analysis offers great insight to bacterial metabolites or processes that would otherwise be overlooked but may be worth exploring, especially with regards to an organism exhibiting a distinct plastic trait.

Reflecting on both the methodology and the findings of this thesis, a question arises: where does it stand with regards to holism and reductionism? Is it looking at an entire system from above or trying to identify specific mechanisms within it? The answer lies somewhere in the middle, with an inclination towards holism. In my work, I attempt to reconcile the best of both worlds; striving to understand the story that the emerging patterns of the bacterial metabolic potential and the gene co-expression network are telling while trying to specify the mechanisms responding to a diet, a metabolic pathway or a supplemented metabolite. The bipartite network offered a holistic glimpse into the interactions present in the “bacterial MPG-nematode transcriptomic response” system; at the same time, it exhibited high levels of complexity that would be impossible to disentangle mechanistically within the scope of this thesis. Future studies can make use of this data to screen selected candidate metabolites for their effect on specific gene expression modules.

5. REFERENCES

1. Humphreys J. Lamarck and the general theory of evolution. *Journal of Biological Education* 1996; 30: 295-303. DOI: 10.1080/00219266.1996.9655520.
2. Ruppel S. Nature as a Huge Organism: Gottfried Reinhold Treviranus (1776–1837) and Early Ecology in German Romantic Science. *Histories* 2023; 3: 112-128. DOI: 10.3390/histories3020009.
3. Fang FC and Casadevall A. Reductionistic and holistic science. *Infect Immun* 2011; 79: 1401-1404. 2011/02/16. DOI: 10.1128/IAI.01343-10.
4. Smuts JC. *Holism and evolution*. Cape Town: N & S Press, 1926.
5. Capra F and Luisi PL. *The Systems View of Life: A Unifying Vision*. Cambridge University Press, 2014.
6. *The Cambridge Companion to Early Greek Philosophy*. Cambridge: Cambridge University Press, 1999.
7. Kalfas V. *The study of the Heavens [Undergraduate textbook]*. Kallipos, Open Academic Editions, 2022.
8. Russell B. *History of Western Philosophy*. Routledge, 2004.
9. Cherniss H. The Characteristics and Effects of Presocratic Philosophy. *Journal of the History of Ideas* 1951; 12. DOI: 10.2307/2707750.
10. Russell B. Non-Materialistic Naturalism. *The Kenyon Review* 1942; 4: 361–365.
11. Thomson A. Materialistic Theories of Mind and Brain. *Between Leibniz, Newton, and Kant*. 2023, pp.197-225.
12. Grosholz E. Plato and Leibniz against the Materialists. *Journal of the History of Ideas* 1996; 57. DOI: 10.1353/jhi.1996.0016.
13. Ariew A. Teleology. *The Cambridge Companion to the Philosophy of Biology*. 2007, pp.160-181.
14. Solmsen F. Plato and the Concept of the Soul (Psyche): Some Historical Perspectives. *Journal of the History of Ideas* 1983; 44. DOI: 10.2307/2709171.
15. Santas G. Knowledge and Belief in Plato's Republic. *Greek Studies in the Philosophy and History of Science*. 1990, pp.45-59.
16. Mayr E. *What Makes Biology Unique?* 2009.
17. Whewell W. *The Philosophy of the Inductive Sciences*. 2014.
18. Klein J and Giglioni G. Francis Bacon. In: N. ZE, (ed.). *The Stanford Encyclopedia of Philosophy*. Fall 2020 ed.: Metaphysics Research Lab, Stanford University, 2020.
19. Calvente S and Manzo S. Empiricism, Early Modern. *Encyclopedia of Early Modern Philosophy and the Sciences*. 2020, pp.1-13.
20. Popper KR. Back to the Pre-Socratics: The Presidential Address. *Proceedings of the Aristotelian Society* 1958: 1-24.
21. Percival RS. Descartes' Model of Mind. In: Lilienfeld SO and Cautin RL (eds) *The Encyclopedia of Clinical Psychology*. 2015, pp.1-8.
22. Mazzocchi F. Complexity in biology. Exceeding the limits of reductionism and determinism using complexity theory. *EMBO Rep* 2008; 9: 10-14. 2008/01/05. DOI: 10.1038/sj.embor.7401147.
23. Kellenberger E. The evolution of molecular biology. *EMBO Rep* 2004; 5: 546-549. 2004/06/02. DOI: 10.1038/sj.embor.7400180.
24. Allen GE. Mechanism, vitalism and organicism in late nineteenth and twentieth-century biology: the importance of historical context. *Stud Hist Philos Biol Biomed Sci* 2005; 36: 261-283. 2005/06/01. DOI: 10.1016/j.shpsc.2005.03.003.

25. Bechtel W. Addressing the Vitalist's Challenge to Mechanistic Science: Dynamic Mechanistic Explanation. *Vitalism and the Scientific Image in Post-Enlightenment Life Science, 1800-2010*. 2013, pp.345-370.
26. Gilbert SF and Sarkar S. Embracing complexity: organicism for the 21st century. *Dev Dyn* 2000; 219: 1-9. 2000/09/07. DOI: 10.1002/1097-0177(2000)9999:9999<::AID-DVDY1036>3.0.CO;2-A.
27. Hofer V. Philosophy of Biology around the Vienna Circle: Ludwig von Bertalanffy, Joseph Henry Woodger and Philipp Frank. *History of Philosophy of Science*. 2002, pp.325-333.
28. Ayala FJ. Biological Reductionism: The Problems and Some Answers. In: Yates FE (ed) *Self-organizing systems: The emergence of order*. New York & London: Plenum Press, 1987, pp.315-324.
29. Fry I. On the Biological Significance of the Properties of Matter: L. J. Henderson's Theory of the Fitness of the Environment. *Journal of the History of Biology* 1996; 29: 155-196.
30. Leahey T. Gestalt psychology. *The Cambridge History of Philosophy 1870–1945*. 2003, pp.649-657.
31. Wagemans J, Feldman J, Gepshtein S, et al. A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations. *Psychol Bull* 2012; 138: 1218-1252. 2012/08/01. DOI: 10.1037/a0029334.
32. Bello RE. The Systems Approach: A. Bogdanov and L. von Bertalanffy. *Studies in Soviet Thought* 1985; 30: 131–147.
33. Gorelik G. Bogdanov's Tektology: Its Nature, Development and Influence. *Studies in Soviet Thought* 1983; 26: 39-57.
34. Dudley P. Back to basics? Tektology and General System Theory (GST). *Systems Practice* 1996; 9: 273-284. DOI: 10.1007/bf02169018.
35. von Bertalanffy L. *General System Theory: Foundations, Development*. New York: George Braziller, 1968.
36. Hammond D. The legacy of Ludwig von Bertalanffy and its relevance for our time. *Systems Research and Behavioral Science* 2019; 36: 301-307. DOI: 10.1002/sres.2598.
37. Marín VH. General System Theory and the Ecosystem Concept. *Bulletin of the Ecological Society of America* 1997; 78: 102-104.
38. von Bertalanffy L. The History and Status of General Systems Theory. *The Academy of Management Journal* 1972; 15: 407-426.
39. Terreberry S. The Evolution of Organizational Environments. *Administrative Science Quarterly* 1968; 12: 590-613.
40. Wiener N. *Cybernetics or Control and Communication in the Animal and the Machine*. 2019.
41. Waterman TH. Systems Theory and Biology—View of a Biologist. *Systems Theory and Biology*. 1968, pp.1-37.
42. Sneyd J, Theraula G, Bonabeau E, et al. *Self-Organization in Biological Systems*. 2001.
43. Wolkenhauer O. Systems biology: the reincarnation of systems theory applied in biology? *Brief Bioinform* 2001; 2: 258-270. 2001/10/09. DOI: 10.1093/bib/2.3.258.
44. Jacob F and Monod J. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* 1961; 3: 318-356. 1961/06/01. DOI: 10.1016/s0022-2836(61)80072-7.
45. Weaver W. Science and complexity. *American Scientist* 1948; 36: 536-544.

46. Mossio M and Moreno A. Organisational Closure in Biological Organisms. *History and Philosophy of the Life Sciences* 2010; 32: 269-288.
47. Klir GJ. *Facets of Systems Science*. 2001.
48. Gross F and Green S. The Sum of the Parts: Large-Scale Modeling in Systems Biology. *Philosophy, Theory, and Practice in Biology* 2017; 9. DOI: 10.3998/ptb.6959004.0009.010.
49. Brigandt I and Love A. Reductionism in Biology. In: Zalta EN and Nodelman U (eds) *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2023.
50. Bechtel W and Richardson RC. *Discovering Complexity*. 2010.
51. Kitano H. Systems biology: a brief overview. *Science* 2002; 295: 1662-1664. 2002/03/02. DOI: 10.1126/science.1069492.
52. Kauffman SA. Origins of Order in Evolution: Self-Organization and Selection. *Understanding Origins*. 1992, pp.153-181.
53. Werner J. System properties, feedback control and effector coordination of human temperature regulation. *Eur J Appl Physiol* 2010; 109: 13-25. 2009/09/30. DOI: 10.1007/s00421-009-1216-1.
54. Walter MH, Abele H and Plappert CF. The Role of Oxytocin and the Effect of Stress During Childbirth: Neurobiological Basics and Implications for Mother and Child. *Front Endocrinol (Lausanne)* 2021; 12: 742236. 2021/11/16. DOI: 10.3389/fendo.2021.742236.
55. Krebs JR. Colonial Nesting and Social Feeding as Strategies for Exploiting Food Resources in the Great Blue Heron (Ardea Herodias). *Behaviour* 1974; 51: 99-134. DOI: 10.1163/156853974x00165.
56. Gross MR and MacMillan AM. Predation and the evolution of colonial nesting in bluegill sunfish (Lepomis macrochirus). *Behavioral Ecology and Sociobiology* 1981; 8: 163-174. DOI: 10.1007/bf00299826.
57. Wolkenhauer O and Mesarovic M. Feedback dynamics and cell function: Why systems biology is called Systems Biology. *Mol Biosyst* 2005; 1: 14-16. 2006/08/02. DOI: 10.1039/b502088n.
58. Kitano H. Biological robustness. *Nat Rev Genet* 2004; 5: 826-837. 2004/11/03. DOI: 10.1038/nrg1471.
59. Goldbeter A. Dissipative structures in biological systems: bistability, oscillations, spatial patterns and waves. *Philos Trans A Math Phys Eng Sci* 2018; 376 2018/06/13. DOI: 10.1098/rsta.2017.0376.
60. Prigogine I and Nicolis G. Self-Organisation in Nonequilibrium Systems: Towards A Dynamics of Complexity. *Bifurcation Analysis*. 1985, pp.3-12.
61. Rathkopf C. Network representation and complex systems. *Synthese* 2015; 195: 55-78. DOI: 10.1007/s11229-015-0726-0.
62. Pavlopoulos GA, Secrier M, Moschopoulos CN, et al. Using graph theory to analyze biological networks. *BioData Min* 2011; 4: 10. 2011/04/30. DOI: 10.1186/1756-0381-4-10.
63. Koutrouli M, Karatzas E, Paez-Espino D, et al. A Guide to Conquer the Biological Network Era Using Graph Theory. *Front Bioeng Biotechnol* 2020; 8: 34. 2020/02/23. DOI: 10.3389/fbioe.2020.00034.
64. Huber W, Carey VJ, Long L, et al. Graphs in molecular biology. *BMC Bioinformatics* 2007; 8 Suppl 6: S8. 2007/10/02. DOI: 10.1186/1471-2105-8-S6-S8.
65. Barabasi AL. Scale-free networks: a decade and beyond. *Science* 2009; 325: 412-413. 2009/07/25. DOI: 10.1126/science.1173299.

66. Erdős P and Rényi A. On the strength of connectedness of a random graph. *Acta Mathematica Hungarica* 1961.
67. Watts DJ and Strogatz SH. Collective dynamics of 'small-world' networks. *Nature* 1998; 393: 440-442. 1998/06/12. DOI: 10.1038/30918.
68. Barabasi AL and Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004; 5: 101-113. 2004/01/22. DOI: 10.1038/nrg1272.
69. Ideker T, Thorsson V, Ranish JA, et al. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 2001; 292: 929-934. 2001/05/08. DOI: 10.1126/science.292.5518.929.
70. Arrieta-Ortiz ML, Immanuel SRC, Turkarslan S, et al. Predictive regulatory and metabolic network models for systems analysis of Clostridioides difficile. *Cell Host Microbe* 2021; 29: 1709-1723 e1705. 2021/10/13. DOI: 10.1016/j.chom.2021.09.008.
71. Zheng F, Kelly MR, Ramms DJ, et al. Interpretation of cancer mutations using a multiscale map of protein systems. *Science* 2021; 374: eabf3067. 2021/10/01. DOI: 10.1126/science.abf3067.
72. Thompson JN. The evolution of species interactions. *Science* 1999; 284: 2116-2118. 1999/06/26. DOI: 10.1126/science.284.5423.2116.
73. Wootton JT and Emmerson M. Measurement of Interaction Strength in Nature. *Annual Review of Ecology, Evolution, and Systematics* 2005; 36: 419-444. DOI: 10.1146/annurev.ecolsys.36.091704.175535.
74. Zhang Z, Yan C and Zhang H. Mutualism between antagonists: its ecological and evolutionary implications. *Integr Zool* 2021; 16: 84-96. 2020/09/16. DOI: 10.1111/1749-4877.12487.
75. Banerji A. Acquisition of Adaptive Traits via Interspecific Association: Ecological Consequences and Applications. *Ecologies* 2021; 2: 43-70. DOI: 10.3390/ecologies2010004.
76. Sportes A, Heriche M, Boussageon R, et al. A historical perspective on mycorrhizal mutualism emphasizing arbuscular mycorrhizas and their emerging challenges. *Mycorrhiza* 2021; 31: 637-653. 2021/10/18. DOI: 10.1007/s00572-021-01053-2.
77. Mathis KA and Bronstein JL. Our Current Understanding of Commensalism. *Annual Review of Ecology, Evolution, and Systematics* 2020; 51: 167-189. DOI: 10.1146/annurev-ecolsys-011720-040844.
78. Naranjo C, Iriondo JM, Riofrio ML, et al. Evaluating the structure of commensalistic epiphyte-phorophyte networks: a comparative perspective of biotic interactions. *AoB Plants* 2019; 11: plz011. 2019/04/19. DOI: 10.1093/aobpla/plz011.
79. Couto DR, Francisco TM and Nascimento MT. Commensalistic epiphyte–phorophyte networks in woody vegetation of tropical inselbergs: Patterns of organization and structure. *Austral Ecology* 2022; 47: 911-927. DOI: 10.1111/aec.13173.
80. Johnson JB and Belk MC. Predators as Agents of Selection and Diversification. *Diversity* 2020; 12. DOI: 10.3390/d12110415.
81. Dubois F and Binning SA. Predation and parasitism as determinants of animal personalities. *J Anim Ecol* 2022; 91: 1918-1928. 2022/07/21. DOI: 10.1111/1365-2656.13781.
82. Courbin N, Loveridge AJ, Fritz H, et al. Zebra diel migrations reduce encounter risk with lions at night. *J Anim Ecol* 2019; 88: 92-101. 2018/10/04. DOI: 10.1111/1365-2656.12910.

83. Brown SP, Le Chat L and Taddei F. Evolution of virulence: triggering host inflammation allows invading pathogens to exclude competitors. *Ecol Lett* 2008; 11: 44-51. 2007/11/21. DOI: 10.1111/j.1461-0248.2007.01125.x.
84. Runge S and Rosshart SP. The Mammalian Metaorganism: A Holistic View on How Microbes of All Kingdoms and Niches Shape Local and Systemic Immunity. *Front Immunol* 2021; 12: 702378. 2021/07/20. DOI: 10.3389/fimmu.2021.702378.
85. Trueba G, Cardenas P, Romo G, et al. Reevaluating human-microbiota symbiosis: Strain-level insights and evolutionary perspectives across animal species. *Biosystems* 2024; 244: 105283. 2024/08/06. DOI: 10.1016/j.biosystems.2024.105283.
86. Zaiss MM and Harris NL. Interactions between the intestinal microbiome and helminth parasites. *Parasite Immunol* 2016; 38: 5-11. 2015/09/09. DOI: 10.1111/pim.12274.
87. Yadav A, Mohite SV, Behl A, et al. Gut Microbiota and Its Role in Human Metabolic Disorders. *Role of Microbes in Sustainable Development*. 2023, pp.313-334.
88. Simon JC, Marchesi JR, Mougel C, et al. Host-microbiota interactions: from holobiont theory to analysis. *Microbiome* 2019; 7: 5. 2019/01/13. DOI: 10.1186/s40168-019-0619-4.
89. Donia MS and Fischbach MA. HUMAN MICROBIOTA. Small molecules from the human microbiota. *Science* 2015; 349: 1254766. 2015/07/25. DOI: 10.1126/science.1254766.
90. Nascimento JS, Ceotto H, Nascimento SB, et al. Bacteriocins as alternative agents for control of multiresistant staphylococcal strains. *Lett Appl Microbiol* 2006; 42: 215-221. 2006/02/16. DOI: 10.1111/j.1472-765X.2005.01832.x.
91. Dabard J, Bridonneau C, Phillipe C, et al. Ruminococcin A, a new lantibiotic produced by a Ruminococcus gnavus strain isolated from human feces. *Appl Environ Microbiol* 2001; 67: 4111-4118. 2001/08/30. DOI: 10.1128/AEM.67.9.4111-4118.2001.
92. Hou K, Wu ZX, Chen XY, et al. Microbiota in health and diseases. *Signal Transduct Target Ther* 2022; 7: 135. 2022/04/25. DOI: 10.1038/s41392-022-00974-4.
93. Gomaa EZ. Human gut microbiota/microbiome in health and diseases: a review. *Antonie Van Leeuwenhoek* 2020; 113: 2019-2040. 2020/11/03. DOI: 10.1007/s10482-020-01474-7.
94. Zhao Q, Chen Y, Huang W, et al. Drug-microbiota interactions: an emerging priority for precision medicine. *Signal Transduct Target Ther* 2023; 8: 386. 2023/10/09. DOI: 10.1038/s41392-023-01619-w.
95. Schratzberger M, Holterman M, van Oevelen D, et al. A Worm's World: Ecological Flexibility Pays Off for Free-Living Nematodes in Sediments and Soils. *Bioscience* 2019; 69: 867-876. 2019/11/14. DOI: 10.1093/biosci/biz086.
96. Majdi N and Traunspurger W. Free-living nematodes in the freshwater food web: a review. *Journal of nematology* 2015; 47: 28-44.
97. van den Hoogen J, Geisen S, Routh D, et al. Soil nematode abundance and functional group composition at a global scale. *Nature* 2019; 572: 194-198. 2019/07/26. DOI: 10.1038/s41586-019-1418-6.
98. Blaxter M. Nematodes: the worm and its relatives. *PLoS Biol* 2011; 9: e1001050. 2011/04/29. DOI: 10.1371/journal.pbio.1001050.
99. Dirksen P, Marsh SA, Braker I, et al. The native microbiome of the nematode Caenorhabditis elegans: gateway to a new host-microbiome model. *BMC Biol* 2016; 14: 38. 2016/05/11. DOI: 10.1186/s12915-016-0258-1.
100. Felix MA and Braendle C. The natural history of Caenorhabditis elegans. *Curr Biol* 2010; 20: R965-969. 2010/11/26. DOI: 10.1016/j.cub.2010.09.050.

101. Zimmermann J, Obeng N, Yang W, et al. The functional repertoire contained within the native microbiota of the model nematode Caenorhabditis elegans. *ISME J* 2020; 14: 26-38. 2019/09/06. DOI: 10.1038/s41396-019-0504-y.
102. Berg M, Stenuit B, Ho J, et al. Assembly of the Caenorhabditis elegans gut microbiota from diverse soil microbial environments. *ISME J* 2016; 10: 1998-2009. 2016/01/23. DOI: 10.1038/ismej.2015.253.
103. Petersen C, Hamerich IK, Adair KL, et al. Host and microbiome jointly contribute to environmental adaptation. *ISME J* 2023; 17: 1953-1965. 2023/09/07. DOI: 10.1038/s41396-023-01507-9.
104. Lansdon P, Carlson M and Ackley BD. Wild-type Caenorhabditis elegans isolates exhibit distinct gene expression profiles in response to microbial infection. *BMC Genomics* 2022; 23: 229. 2022/03/25. DOI: 10.1186/s12864-022-08455-2.
105. Volkers RJ, Snoek LB, Hubar CJ, et al. Gene-environment and protein-degradation signatures characterize genomic and phenotypic diversity in wild Caenorhabditis elegans populations. *BMC Biol* 2013; 11: 93. 2013/08/21. DOI: 10.1186/1741-7007-11-93.
106. Yang W, Petersen C, Pees B, et al. The Inducible Response of the Nematode Caenorhabditis elegans to Members of Its Natural Microbiota Across Development and Adult Life. *Front Microbiol* 2019; 10: 1793. 2019/08/24. DOI: 10.3389/fmicb.2019.01793.
107. Hong RL and Sommer RJ. Pristionchus pacificus: a well-rounded nematode. *Bioessays* 2006; 28: 651-659. 2006/05/16. DOI: 10.1002/bies.20404.
108. Sommer RJ and McGaughran A. The nematode Pristionchus pacificus as a model system for integrative studies in evolutionary biology. *Mol Ecol* 2013; 22: 2380-2393. 2013/03/28. DOI: 10.1111/mec.12286.
109. Sommer RJ, Dardiry M, Lenuzzi M, et al. The genetics of phenotypic plasticity in nematode feeding structures. *Open Biol* 2017; 7 2017/03/17. DOI: 10.1098/rsob.160332.
110. Lenuzzi M, Witte H, Riebesell M, et al. Influence of environmental temperature on mouth-form plasticity in Pristionchus pacificus acts through daf-11-dependent cGMP signaling. *J Exp Zool B Mol Dev Evol* 2023; 340: 214-224. 2021/08/12. DOI: 10.1002/jez.b.23094.
111. Sanghvi GV, Baskaran P, Roseler W, et al. Life History Responses and Gene Expression Profiles of the Nematode Pristionchus pacificus Cultured on Cryptococcus Yeasts. *PLoS One* 2016; 11: e0164881. 2016/10/16. DOI: 10.1371/journal.pone.0164881.
112. Werner MS, Sieriebriennikov B, Loschko T, et al. Environmental influence on Pristionchus pacificus mouth form through different culture methods. *Sci Rep* 2017; 7: 7207. 2017/08/05. DOI: 10.1038/s41598-017-07455-7.
113. Namdeo S, Moreno E, Rodelsperger C, et al. Two independent sulfation processes regulate mouth-form plasticity in the nematode Pristionchus pacificus. *Development* 2018; 145 2018/07/04. DOI: 10.1242/dev.166272.
114. Lo WS, Han Z, Witte H, et al. Synergistic interaction of gut microbiota enhances the growth of nematode through neuroendocrine signaling. *Curr Biol* 2022; 32: 2037-2050 e2034. 2022/04/10. DOI: 10.1016/j.cub.2022.03.056.
115. Iatsenko I, Nikolov A and Sommer RJ. Identification of distinct Bacillus thuringiensis 4A4 nematicidal factors using the model nematodes Pristionchus pacificus and Caenorhabditis elegans. *Toxins (Basel)* 2014; 6: 2050-2063. 2014/07/16. DOI: 10.3390/toxins6072050.

116. Akduman N, Lightfoot JW, Roseler W, et al. Bacterial vitamin B(12) production enhances nematode predatory behavior. *ISME J* 2020; 14: 1494-1507. 2020/03/11. DOI: 10.1038/s41396-020-0626-2.
117. Lo WS, Sommer RJ and Han Z. Microbiota succession influences nematode physiology in a beetle microcosm ecosystem. *Nat Commun* 2024; 15: 5137. 2024/06/16. DOI: 10.1038/s41467-024-49513-5.
118. Akduman N, Rodelsperger C and Sommer RJ. Culture-based analysis of Pristionchus-associated microbiota from beetles and figs for studying nematode-bacterial interactions. *PLoS One* 2018; 13: e0198018. 2018/06/05. DOI: 10.1371/journal.pone.0198018.
119. Herrmann M, Mayer WE and Sommer RJ. Nematodes of the genus Pristionchus are closely associated with scarab beetles and the Colorado potato beetle in Western Europe. *Zoology (Jena)* 2006; 109: 96-108. 2006/04/18. DOI: 10.1016/j.zool.2006.03.001.
120. Renahan T, Lo WS, Werner MS, et al. Nematode biphasic 'boom and bust' dynamics are dependent on host bacterial load while linking dauer and mouth-form polyphenisms. *Environ Microbiol* 2021; 23: 5102-5113. 2021/02/16. DOI: 10.1111/1462-2920.15438.
121. Dieterich C, Clifton SW, Schuster LN, et al. The Pristionchus pacificus genome provides a unique perspective on nematode lifestyle and parasitism. *Nat Genet* 2008; 40: 1193-1198. 2008/09/23. DOI: 10.1038/ng.227.
122. Meyer JM, Baskaran P, Quast C, et al. Succession and dynamics of Pristionchus nematodes and their microbiome during decomposition of Oryctes borbonicus on La Reunion Island. *Environ Microbiol* 2017; 19: 1476-1489. 2017/02/16. DOI: 10.1111/1462-2920.13697.
123. Sun S, Rodelsperger C and Sommer RJ. Single worm transcriptomics identifies a developmental core network of oscillating genes with deep conservation across nematodes. *Genome Res* 2021; 31: 1590-1601. 2021/07/25. DOI: 10.1101/gr.275303.121.
124. Prabh N, Roeseler W, Witte H, et al. Deep taxon sampling reveals the evolutionary dynamics of novel gene families in Pristionchus nematodes. *Genome Res* 2018; 28: 1664-1674. 2018/09/21. DOI: 10.1101/gr.234971.118.
125. Khalturin K, Hemmrich G, Fraune S, et al. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet* 2009; 25: 404-413. 2009/09/01. DOI: 10.1016/j.tig.2009.07.006.
126. Sieriebriennikov B, Sun S, Lightfoot JW, et al. Conserved nuclear hormone receptors controlling a novel plastic trait target fast-evolving genes expressed in a single cell. *PLoS Genet* 2020; 16: e1008687. 2020/04/14. DOI: 10.1371/journal.pgen.1008687.
127. Caspi R, Altman T, Billington R, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* 2014; 42: D459-471. 2013/11/15. DOI: 10.1093/nar/gkt1103.

6. APPENDIX

SCIENTIFIC REPORTS
nature research

OPEN

Crowdsourcing and the feasibility of manual gene annotation: A pilot study in the nematode *Pristionchus pacificus*

Christian Rödelsperger*, Marina Athanasouli, Maša Lenuzzi, Tobias Theska, Shuai Sun, Mohannad Dardiry, Sara Wighard, Wen Hu, Devansh Raj Sharma & Ziduan Han

Nematodes such as *Caenorhabditis elegans* are powerful systems to study basically all aspects of biology. Their species richness together with tremendous genetic knowledge from *C. elegans* facilitate the evolutionary study of biological functions using reverse genetics. However, the ability to identify orthologs of candidate genes in other species can be hampered by erroneous gene annotations. To improve gene annotation in the nematode model organism *Pristionchus pacificus*, we performed a genome-wide screen for *C. elegans* genes with potentially incorrectly annotated *P. pacificus* orthologs. We initiated a community-based project to manually inspect more than two thousand candidate loci and to propose new gene models based on recently generated Iso-seq and RNA-seq data. In most cases, misannotation of *C. elegans* orthologs was due to artificially fused gene predictions and completely missing gene models. The community-based curation raised the gene count from 25,517 to 28,036 and increased the single copy ortholog completeness level from 86% to 97%. This pilot study demonstrates how even small-scale crowdsourcing can drastically improve gene annotations. In future, similar approaches can be used for other species, gene sets, and even larger communities thus making manual annotation of large parts of the genome feasible.

How well can biological knowledge be transferred across species? Are biological functions carried out by the same genes in different organisms? How fast do regulatory networks diverge? In order to address these fundamental questions, more than 20 years ago, the nematode *Pristionchus pacificus* has been introduced as a so-called "satellite" model organism to one of the most successful animal model systems, *Caenorhabditis elegans*[1,2]. Since then, several comparative studies in developmental and ecological contexts have highlighted the importance of developmental system drift as a concept in evolution[3] and have demonstrated that the divergence between *Pristionchus* and *Caenorhabditis* was accompanied by extensive chemical[4–6], genic[7–9], and morphological[10–12] innovations. The establishment of multiple genetic[13,14] and genomic tools and resources[15,16] by Sommer and colleagues motivated an increasing number of independent groups to adapt *P. pacificus* as a model system for comparative studies at a mechanistic level[17–21]. However, reverse genetic approaches based on candidate genes with known functions in *C. elegans*[22,23] have been hampered not only by the huge amount of lineage-specific duplications[23–26], but also by missing and incorrect gene annotations. Traditionally, protein-coding genes are annotated by gene prediction algorithms that model general sequence features of transcription and translation start and end sites, as well as splicing signals[27–29]. This can be complemented with evidence based approaches using transcriptomic and protein homology data[30,31]. While automated annotation pipelines perform reasonably well to be useful for genetic screens[32–34] and evolutionary genomic analyses[35–37], their outcomes by far do not meet the standards of the gene annotations from classical model organisms such as *C. elegans*, *Drosophila melanogaster*, and *Mus musculus* that have been curated over decades by a large research community[38]. In order to make the *P. pacificus* system more tractable for researchers without extensive genomic and phylogenetic expertise, we need to minimize the discrepancy in gene annotation quality between *C. elegans* and *P. pacificus*. To this end, we employed an integrative approach using comparative genomic and transcriptomic data combined with crowdsourcing to improve the *P. pacificus* annotations of *C. elegans* homologs and orthologs. First, we carry out a comparative assessment of 22

Max Planck Institute for Developmental Biology, Department for Integrative Evolutionary Biology, Max-Planck-Ring 9, 72076, Tübingen, Germany. *email: christian.roedelsperger@tuebingen.mpg.de

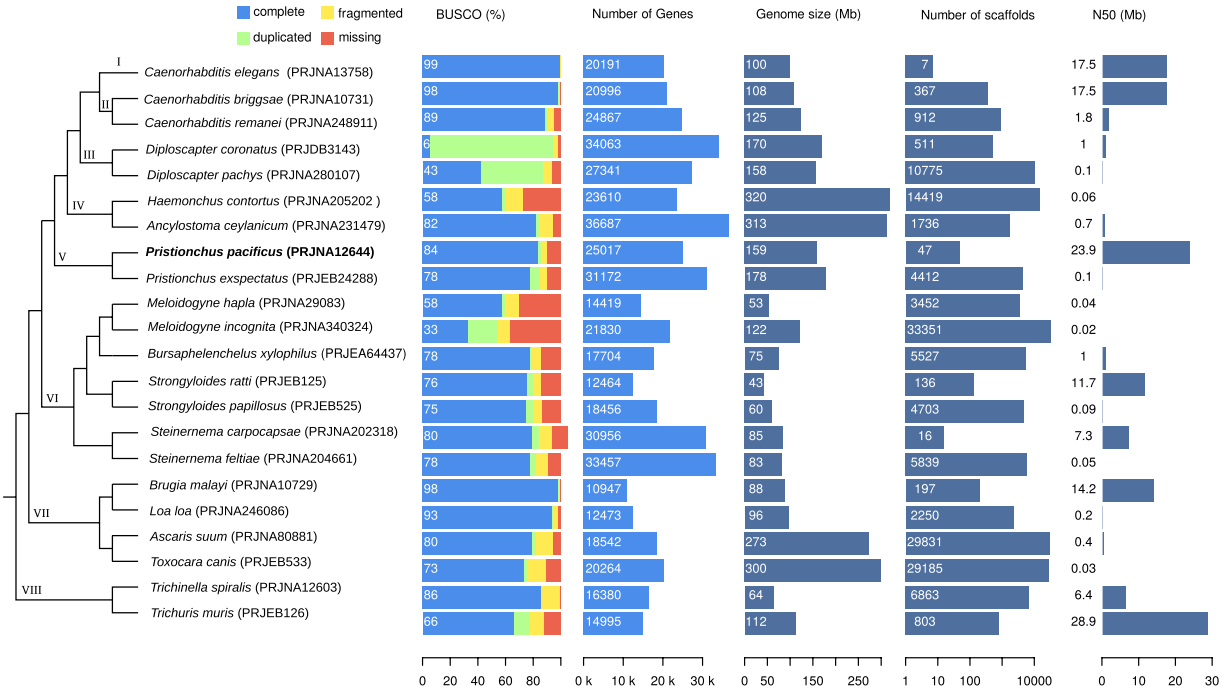


Figure 1. Comparative assessment of nematode genome quality. Genomic data for 22 nematode species was obtained from WormBase ParaSite (release WBPS13) and evaluated based on completeness level of gene annotations and genome assembly contiguity. The barplots show the results of a benchmarking of single copy orthologs (BUSCO[40]) analysis, the number of genes, genome sizes, number of scaffolds, and the N50 measure of assembly contiguity. The genome and annotations of *P. pacificus* exhibit an overall comparatively high quality. The schematic phylogeny is based on phylogenomic analysis of 108 nematodes[39], Roman numerals indicate phylostrata that are used for further analysis.

nematode genomes and demonstrate that *P. pacificus* has one of the best available nematode genomes. Second, we perform a genomewide screen for *C. elegans* genes where homologs and orthologs are not or incorrectly annotated in the *P. pacificus* genome. Third, a community-based manual curation of suspicious gene models reveals thousands of hidden orthologs and missing homologs. This pilot study can be extended to even larger gene sets and communities possibly employing citizen scientists, which would raise the quality of gene annotations to the next level[38].

Results

The quality of nematode draft genomes is highly heterogeneous. To obtain a general overview of the current status of nematode genome quality, we analyzed assemblies and gene annotations of 22 species (Fig. 1). The species were arbitrarily selected to span the diversity of the nematode phylum[39]. We will further use this taxon sampling to perform an analysis of gene age, i.e. phylostratigraphic analysis where each phylostratum is defined by at least two outgroup species to minimize the effect of species-specific gene loss. Nematode genomes range in size between 43 and 320 Mb and contain between 11 and 37 thousand annotated protein-coding genes (Fig. 1). Analyses of assembly features and gene annotations indicate a wide range of qualitative variability. Some genomes are assembled and scaffolded to the level of chromosomes with high degrees of contiguity (the N50 value which is a measure of genome assembly contiguity is up to 29 Mb) whereas others are largely fragmented into up to 33 thousand scaffolds with N50 values below 0.1 Mb (Fig. 1). Similarly, analyses of completeness levels based on benchmarking univeral single copy orthologs (BUSCO[40]) reveal substantial amount of either missing or duplicated genes and it is not totally clear to what extent these differences are of biological or technical nature[41]. In the case of *Diploscapter coronatus*, the apparent high fraction of duplicated genes could either be explained by hybridization of two divergent lineages or a whole genome duplication[42]. The genome of *P. pacificus*, which was generated by assembly from single-molecule, long-read sequencing data and scaffolding with the help of a genetic linkage map[15], shows one of the highest levels of contiguity (47 scaffolds, N50 = 24 Mb). Gene annotations were generated by the MAKER2 pipeline[30,31] which combined gene prediction algorithms, transcriptome data, and protein homology data from other *Pristionchus* species[11,15,43]. The completeness level of gene annotations (BUSCO completeness: 84%) is in the upper range when compared to most other nematode genomes (median 78%, interquartile range (IQR): 68–85%, Fig. 1). This demonstrates the relatively high quality of the current *P. pacificus* assembly and gene annotations.

Complementary genome and transcriptomes reveal potentially missing gene models. The completeness analysis as implemented in the software BUSCO[40] can also be applied to the raw genome assembly of *P. pacificus*. This yielded a combined completeness value of 93% (complete single copy and duplicates) as

Data set	BUSCO (%)				Ref
	Complete Single Copy (+Duplicates)	Duplicate	Fragmented	Missing	
Genome assembly (El Paco assembly)	91.6 (92.9)	1.3	3.1	4.0	[15]
El Paco annotation v1/WS268	84.0 (85.8)	1.8	4.3	9.9	[15]
de novo transcriptome assembly	59.1 (97.1)	38.0	2.6	0.3	[16]
Iso-Seq assembly	48.0 (73.3)	25.3	10.9	15.8	[44]
El Paco annotation v2	95.4 (97.1)	1.7	2.0	0.9	this study

Table 1. Completeness analysis of different *P. pacificus* data set. The high level of duplicates in the two transcriptomic data sets is due to the presence of isoforms.

compared to 86% for the *P. pacificus* gene annotations and indicates towards the presence of incorrectly annotated or missing *C. elegans* orthologs in the genome of *P. pacificus*. Moreover, the fact that a recent *de novo* transcriptome assembly that was based on a strand-specific RNA-seq data set exhibited an even higher combined completeness level of 97% (Table 1) demonstrates even further room for improvement[16]. Finally, single-molecule, long-read transcriptome sequencing data were recently generated for *P. pacificus* which allows a much more accurate definition of gene structures from reference alignments of single reads[44]. However, neither transcriptomic data set was available when the existing gene annotations (version: El Paco annotation v1/WormBase release: WS268) were generated and they could still be used for further improvement.

To systematically identify potentially missing genes in the *P. pacificus* genome, we searched for *C. elegans* genes lacking homologs in the current *P. pacificus* gene annotations (BLASTP e-value $< 10^{-5}$) but having a matching open reading frame in the *de novo* transcriptome assembly (Fig. 2a). While 12,504 (62%) *C. elegans* genes had BLASTP hits in both data sets, 634 (3%) *C. elegans* genes showed only BLASTP hits against the current gene annotations suggesting that these genes are properly annotated but are expressed so weakly that they were not captured in the transcriptome assembly of mixed-stage cultures[45]. Similarly, we identified 526 (3%) *C. elegans* genes that were only found in the transcriptome assembly and therefore represent candidates for missing gene annotations.

Community-based curation identifies missing genes in the *P. pacificus* genome. In order to improve the existing gene annotations, we chose to manually inspect and classify all 526 missing gene candidates in the *P. pacificus* genome browser (http://www.pristionchus.org). Thereby, we recruited and trained colleagues as community annotators, who would be capable to classify a genomic locus and to propose a correction to the existing gene models (see *Methods*). Lists of missing gene candidates were shared in online spreadsheets and documents, which allowed multiple annotators to inspect and correct candidate loci in parallel. 119 (25%) of the 486 non-redundant *P. pacificus* loci were classified as missing genes in predicted UTRs of annotated genes (Fig. 2b). We would speculate that this is caused by the fact that nematode genomes are compact and UTR regions can frequently overlap[45]. This can cause artificial fusion of transcripts during the assembly of RNA-seq data. Consequently, only the largest ORF of such a gene is annotated as protein-coding and the rest is classified as 3′ and 5′ UTR. Alternatively, this problem could arise when a fused gene prediction from the sister species is used as homology information but MAKER2 fails to generate a complete gene model out of it. The *C. elegans* gene C29H12.2 is one example of a missing gene model residing in the UTR of a *P. pacificus rars-2* homolog (Fig. 2c). The corresponding *P. pacificus* locus is spanned by two assembled transcripts that are homologous two C29H12.2 and *rars-2*, respectively. Both transcripts are also well supported by Iso-seq data and exhibit different expression levels[44,46]. In such a case, we would propose a replacement of the old *P. pacificus* gene model by the two distinct transcripts.

After manual inspection of all 526 missing gene candidates, 201 (41%) of the 486 non-redundant *P. pacificus* loci were classified as missing genes (Fig. 2b). Presumably this kind of error could arise when the gene annotation pipeline is mostly dependent on gene prediction algorithms which fail to predict all genes in gene dense regions (e.g. operon structures) as the intergenic distances might span only a few hundred nucleotides, which could be too small for triggering the initiation of a new gene model. The *C. elegans* gene *apn-1* is one example of a missed gene model in a gene dense region (Fig. 2d). Given that the *P. pacificus* homolog of *apn-1* has good transcriptomic support, the correction in this case would simply add the transcript to the existing gene models. Other instances of missing homologs are due to borderline cases in the BLASTP searches where one search resulted in an e-value slightly below the e-value threshold (10^{-5}) and the result of the other BLASTP search was slightly above the threshold. In total, we encountered 87 of such cases which we termed 'weak similarity' (Fig. 2b). For such cases no correction was proposed. In summary, we compiled corrections for 280 *P. pacificus* genes which were replaced by 714 new gene models. All these changes were submitted to WormBase and were incorporated in the release WS272.

Artificial gene fusions mask thousands of hidden orthologs. A small number of *C. elegans* genes with missing homologs in the current gene annotations (version: El Paco v1/WS268) of *P. pacificus* were classified as located in fused gene models (Fig. 2b). One potential explanation could be that an artificially fused gene prediction from the sister species is taken as homology data to annotate the orthologous locus in *P. pacificus*, but small errors cause parts of the gene model to be either incompletely or incorrectly annotated in *P. pacificus* resulting in a loss of detectable homology (Fig. 2c). Even if the homolog of a *C. elegans* gene is incorporated in the correct ORF within an artificially fused gene model, this could still cause a loss of one-to-one orthology as the

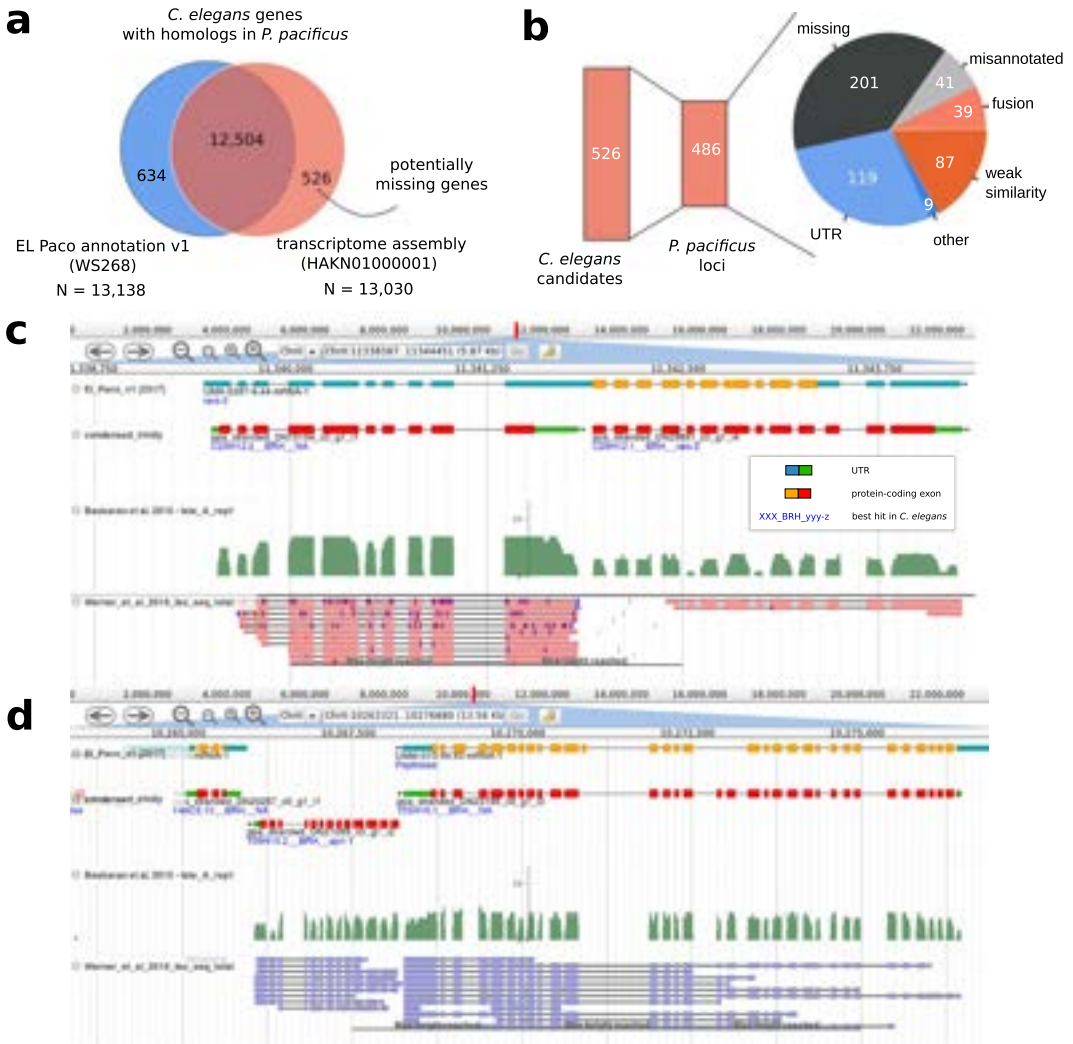


Figure 2. Identification of missing genes. (**a**) 526 potentially missing genes were identified based on *C. elegans* genes with homologs in the transcriptome assembly but not in current gene annotations. (**b**) The 526 missing gene candidates were located in 486 *P. pacificus* loci that were classified based on community annotators. (**c**) The genome browser screenshot shows a homolog of *C. elegans* C29H12.2 which is located in the annotated 5′UTR of a *P. pacificus* gene. This locus harbors two *P. pacificus* transcripts with different expression levels and well supported as non-overlapping transcripts based on RNA-seq and Iso-seq data. (**d**) A homolog of apn-1 is completely missing from current gene annotations.

corresponding *P. pacificus* gene can only be identified as one-to-one ortholog of a single *C. elegans* gene. Thus, we performed a second screen for *C. elegans* genes that had a predicted one-to-one ortholog (best-reciprocal hit) in the transcriptome assembly but not in current gene annotations (Fig. 3a). In total, 6075 (93%) of *C. elegans* genes with a predicted one-to-one ortholog (based on best-reciprocal hits) in current gene annotations, also had a predicted one-to-one ortholog against the *de novo* transcriptome assembly (Fig. 3a). Nevertheless, we found 2075 *C. elegans* genes that only had predicted one-to-one orthologs in the *de novo* transcriptome assembly. Excluding *C. elegans* genes that were identified already in the previous screen for missing homologs, this resulted in 1692 *C. elegans* genes with predicted one-to-one orthologs in the *de novo* transcriptome assembly but not in the current set of gene annotations (version: El Paco v1/WS268). Community-based classification and curation of the 1281 corresponding *P. pacificus* loci classified 912 (71%) cases as artificial gene fusions (Fig. 3b). One such an example is the *C. elegans* gene D1053.3. Its putative ortholog is fused with the *P. pacificus mvb-12* ortholog (Fig. 3c). Apart from being orthologous to two different *C. elegans* genes, both *P. pacificus* genes are supported as non-overlapping transcripts by RNA-seq and Iso-seq, and are expressed at different levels. This confirmed the interpretation of an artificially fused annotation. The proposed correction in this case would be a replacement of the old gene model by the two non-overlapping transcripts. In total, we updated 1241 *P. pacificus* gene models and replaced them with 3305 new models. These updates were submitted to WormBase and will be released following curation. The new *P. pacifcus* gene annotation (version: El Paco v2) with 28,036 gene models is also available on http://www.pristionchus.org/download. The results of the BUSCO analysis (Complete and Single Copy: 95.4%, Duplicated:

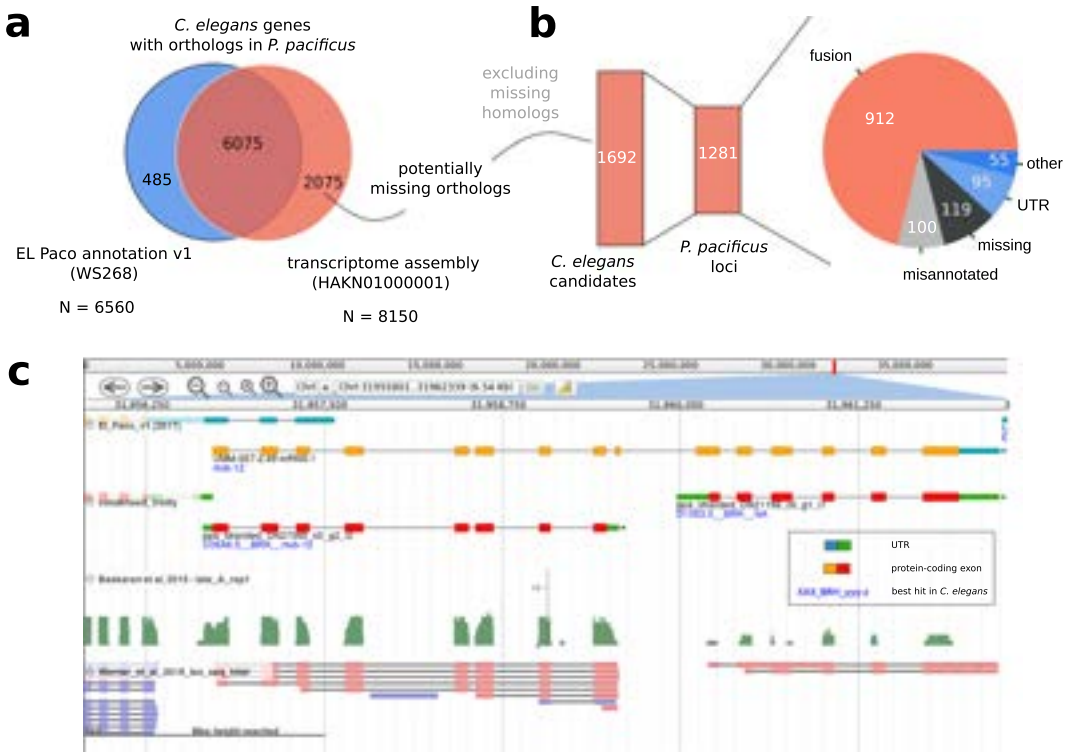


Figure 3. Community-based curation of hidden orthologs. (**a**) We identified 2075 putative *C. elegans* one-to-one orthologs that were specific to the *P. pacificus* transcriptome assembly. (**b**) Community-based curation classified most of the corresponding gene loci as artificial gene fusions. (**c**) Non-overlapping transcripts corresponding to *P. pacificus* orthologs of *mvb-12* and D1053.3 are artificially fused in a current gene model. This prohibits the detection of a one-to-one ortholog of D1053.3 based on a genome-wide approach such as best reciprocal hits.

1.7%, Fragmented: 2.0%, Missing: 0.9%) indicate that the new annotation represents a substantial improvement over the previous annotations[15] (Table 1).

Improved gene annotations facilitate the establishment of a catalog of *C. elegans* homologs and orthologs in the *P. pacificus* genome. Since our primary focus was to improve the annotation of *C. elegans* orthologs in the *P. pacificus* genome, we wanted to use the updated gene annotation to generate a catalog of predicted orthologs between *C. elegans* and *P. pacificus*. As the identification of orthologs typically requires sufficient genomic and phylogenetic knowledge to retrieve relevant protein data sets and to perform reconstruction of gene trees[24,45,46], a genome-wide catalog of orthologs would be highly useful as a starting point for researchers without sufficient expertise. Previous comparisons between *C. elegans* and *P. pacificus* identified putative one-to-one orthologs for roughly 6000–8000 genes[44,46]. To further characterize *C. elegans* genes without orthologs in *P. pacificus*, we additionally carried out a phylostratigraphic analysis[47] to estimate their relative age. Basically, phylostratigraphy uses absence-presence patterns of a gene to map its origin to an internal branch in a species tree[47]. Our analysis revealed that 5258 (26%) of *C. elegans* genes do not have BLAST hits in *Pristionchus* or more distantly related species (Phylostrata I–IV, Supplemental Table 1). This strongly suggests that they are younger than the common ancestor between *C. elegans* and *P. pacificus* and consequently have no orthologs. Next, we applied two different approaches to predict orthologs between *C. elegans* and *P. pacificus*: best reciprocal hits and Markov clustering as implemented in the software orthAgogue[48]. Computation of best reciprocal hits is a standard approach for predicting one-to-one orthologs across species[49,50]. In order to capture more complex orthology relationships (e.g. many-to-many), more general approaches such as Markov clustering have been widely applied[48,51]. Based on best reciprocal hits, we identified 8348 predicted one-to-one orthologs between both species (Supplemental Table 1) whereas the orthAgogue pipeline identified 7643 orthologous clusters, of which only 3345 corresponded to one-to-one orthologs. The large majority (98%) of these predicted one-to-one orthologs was also identified as best reciprocal hits and in 3260 (99%) cases, the same *P. pacificus* gene was predicted as one-to-one ortholog. The large discrepancy between the total number of best reciprocal hits and one-to-one orthologs defined by orthAgogue could be explained by the fact that best reciprocal hits do not take inparalogs into account[49]. However, only 1049 (21%) of *C. elegans* genes that were not identified as one-to-one orthologs by orthAgogue could be explained by the presence of lineage-specific inparalogs, suggesting that orthAgogue with default settings might be too conservative for this analysis. This is further supported by the reanalysis of 57 one-to-one orthologous pairs that were previously confirmed by phylogenetic analysis[46]. While 53 of the previously confirmed one-to-one orthologs were captured as best reciprocal hits, only 33 were also identified

by orthAgogue. Taken together, the improved gene annotation facilitated the prediction of substantially more one-to-one orthologs (Fig. 3a, Supplemental Table 1). This resource can be taken as a starting point to identify candidate genes in *P. pacificus*.

Discussion

With *C. elegans*, *C. briggsae*, and *P. pacificus*, three genetically tractable and free living nematode model organisms have been well established and can be used to study the evolution of gene function at various time-scales[2,3,52]. For example, recent reverse genetic approaches in *P. pacificus* have revealed functional divergence of genes with known roles in *C. elegans* dauer formation[22,23,53]. In addition, mutant screens in *P. pacificus* for social behaviours have uncovered multiple orthologous *C. elegans* genes for which a behavioral phenotype had been overlooked previously[33,54]. Together with complementary studies of the functional importance of novel genes[7,32,55], this makes nematodes an extremely powerful system to study genome evolution and gene function at a mechanistic level.

In order to facilitate fruitful functional studies across multiple model organisms, it is crucial to generate genomic resources (e.g. assemblies, annotations) and experimental genetic toolkits (e.g. forward and reverse genetics) of comparable quality. The chromosome-scale assembly of the *P. pacificus* genome[15] was a major step towards making this species more tractable for other groups. In our study, we aimed to minimize the discrepancy between automatically generated gene annotations for *P. pacificus* and heavily curated annotations for *C. elegans*. To this end, we incorporated recently generated Iso-seq and RNA-seq data into current gene annotations by manual curation of suspicious candidate loci that were identified by comparative genomic analysis. While application of alternative annotation pipelines can generate overall better gene annotations[29,41], they cannot guarantee that gene annotations will only improve. In certain cases, new annotation pipelines will also cause new errors. In contrast, during manual inspection, each community curator has the choice to not propose any change of gene models in case of uncertainty. Thus, manual inspection should only lead to removal of errors and thus improve annotation quality without introducing biases elsewhere. While manual annotation is an incredibly tedious task that is probably not scalable to complete genomes[38], we minimized the workload by focusing on a small gene set of *C. elegans* orthologs, recruiting colleagues as community curators, and restricting the task just to the selection of alternative gene models that were generated from transcriptomic data[16,44] or previous rounds of gene prediction[56,57]. In our opinion, the most crucial aspect of this community project is a good training of new annotators. We achieved this by personal training sessions between experienced and new annotators and the possibility to always discuss cases of uncertainty with other curators. For larger projects, initial training could be achieved by comprehensive online tutorials and communication via email, but this will likely be less efficient. In the case of the *P. pacificus* gene annotations, our study raised the gene count from 25,517 to 28,036 and increased the single copy ortholog completeness level from 86% to 97%. In the *P. pacificus* genome, the greatest source of error was the artificial fusion of neighboring genes. This type of error might be more prevalent in nematodes where genomes are compact[9] and genes frequently overlap[37,45]. Consequently, manual annotation of restricted gene sets has been proposed and applied previously to circumvent this problem[58]. Given that nematode genomes tend to be pretty compact (Fig. 1), we anticipate that misannotation due to overlapping gene models should be much less pronounced in large vertebrate or plant genomes. Nevertheless, it would be interesting to apply similar screens for gene annotation artifacts to other systems and eventually this could reveal some incorrect annotations in the genomes of classical model organisms.

While this study was restricted to *P. pacificus* genes with putative orthologs in *C. elegans*, we cannot reliably estimate the fraction of erroneous gene models across the whole genome. Our results would suggest that the fraction of missing genes is around one percent (Fig. 2a,b) and the amount of gene models affected by artificial fusions may be up to 15% (Fig. 3a). However, as the *P. pacificus* genome has a higher gene density and a higher concentration of old genes at the chromosome centers[8,15], we hypothesize that errors due to artificial gene fusions should be much less pronounced at chromosome arms. To test this, an unbiased quantification of error rates across genomic segments would be needed. In future, we also plan to focus on large gene families and lineage-specific orphan genes[55] that were not explicit subjects of this study. Artificial fusions in these classes of genes could be identified by screens for unexpectedly long gene models or unusual protein domain content. For orphan genes abundant RNA-seq studies of different developmental stages[22,46], tissues[10,46], environmental conditions[59], sexes[16], and genetic backgrounds[60,61] could be used to detect non-overlapping transcripts that exhibit anticorrelated expression within a single locus. Thus, while our study has demonstrated that community-based curation of gene annotations is feasible and can lead to substantial improvements, continued effort is needed to lift its quality to a level that would be similar to classical model organisms.

Methods

Comparative assessment of nematode genomes. We downloaded 22 nematode genomes and corresponding protein sequences from WormBase ParaSite (release WBPS13). For *Steinernema carpocapsae*, the latest version at WBPS14 was used. In case of multiple isoforms, we selected the longest isoform for further analysis. We ran BUSCO (version 3.0.1) in protein mode (option: -m prot) against the nematode_odb9 data set (N = 982 genes) to evaluate the completeness level of available protein sequences.

Genome browser integration of transcriptomic resources. To allow community annotators to propose alternative gene models, we integrated recent raw read alignments and reference guided transcript assemblies of Iso-seq data[44] and a *de novo* assembly of strand-specific RNA-seq data[16] into our genome browser (implemented in jbrowse[62]) on our webserver (http://www.pristionchus.org). Genomic coordinates for the *de novo* transcriptome assembly were generated by alignment to the *P. pacificus* reference genome (version El Paco) with the program exonerate[63] (version: 2.2.0, options: -m est2genome – dnawordlen 20 – maxintron 20000). To reduce the complexity of this data set, a condensed version of the *de novo* transcriptome assembly (selection of the isoform with the longest ORF as single representative isoform per gene, minimum peptide length of 60 amino acids, removal of single exon transcripts) with annotated best-reciprocal hits and best hits (BLASTP, e-value $< 10^{-5}$) in

C. elegans was also incorporated into our jbrowse instance. In addition, predicted protein sequences of previous versions of *P. pacificus* annotations (Hybrid1[56] and TAU2011[57]) were mapped against the *P. pacificus* assembly by exonerate (version: 2.2.0, options: -m protein2genome – dnawordlen 20 – maxintron 20000). All data sets are available under the gene annotation track of the El Paco reference assembly in our genome browser. To evaluate the quality of the two recent transcriptome assemblies, we ran BUSCO (version 3.0.1, options -m trans) against the nematode_odb9 data set (N = 982 genes) for completeness assessment (Table 1).

Identification of missing and fused gene models in current gene annotations. We ran bidirectional BLASTP (e-value $< 10^{-5}$) searches between *C. elegans* (version: WS260, longest isoform per gene) and two different *P. pacificus* data sets: the annotated proteins (version: El Paco v1, WS268) and the *de novo* transcriptome assembly[16]. For the *de novo* transcriptome, we reduced the redundancy resulting from different isoforms by selecting the longest ORFs per gene. Based on the different BLASTP searches, we first screened for *C. elegans* proteins with BLASTP hits against ORFs in the *de novo* transcriptome assembly but not against the currently annotated proteins. This yielded 526 candidate genes. In a second phase, we screened for *C. elegans* proteins with putative orthologs, defined by best-reciprocal BLASTP relationships, in the *de novo* transcriptome assembly but not in the annotated proteins, resulting in 2075 candidate genes.

Community-based manual curation of candidate loci. All *C. elegans* genes together with their candidate homologs and orthologs in the *P. pacificus de novo* transcriptome assembly were stored in a shared online spreadsheet. Community annotators were trained to find the corresponding locus in the genome browser by entering the transcript identifier and to manually inspect the surrounding regions that were defined by the encompassing *P. pacificus* gene model. The candidate locus was then classified as untranslated region (UTR) (the query transcript overlapped exons that were annotated as UTR), missing gene (the query transcript did not overlap any annotated exon), gene fusion (the query transcript did overlap protein-coding exons and homology was detected by BLASTP), misannotation (the query transcript did overlap protein-coding exons but no BLASTP hit was found due incorrect reading frame annotation or minimal overlap) or inconclusive. After classification, a correction was proposed that either added new genes (identifiers could be selected from the *de novo* assembled transcripts, Iso-seq assemblies, or previous versions of gene annotations) or replaced an existing gene model by one or more new genes. In such a case the objective was to lose as little annotated coding sequence as possible. Thus, new genes were selected from the above mentioned data set in order to cover as much coding sequence of the initial gene model as possible. If parts of the old gene model were not covered, BLAST searches against *C. elegans* and other *Pristionchus* species were used to split the old gene model into several parts with sequence matches to distinct *C. elegans* genes, or to extract partial protein sequences of the old gene model that were not covered. Such protein sequence stretches were given a pseudo identifier and were stored in a shared online document. All these sequences were later automatically reannotated by mapping them against the reference genome with the help of exonerate. In case that an existing gene model was replaced by multiple new gene models, we additionally selected one of the new gene models to inherit the WormBase identifier of the old gene model to allow WormBase to record the history of a given gene model. Usually, either the most conserved or the longest new gene model was chosen. Due to the fact that a single artificially fused gene could cause missing homologs and orthologs for multiple *C. elegans* genes, some loci were curated multiple times. We randomly picked some of these cases to compare the classifications and the corresponding corrections from multiple curators, which turned out to be largely consistent. In case of redundant curations, one out of many possible curations for a given locus was chosen based on the following criteria: preference towards higher number of new models, experience of the curator (number of curated loci), and transcriptional evidence over gene prediction.

Phylostratigraphy and orthology predictions. Outgroup data sets were defined by concatenating all protein sequence data from different species in the ladder-like phylogeny leading to *C. elegans* (Fig. 1). More precisely, we pooled all data from species in an induced subtree defined by branches with roman numbers in Fig. 1. We then ran a BLASTP search (e-value < 0.001) of *C. elegans* proteins (longest isoform per gene) against each of the outgroup data sets. Starting from the *C. elegans* genes with homologs in the most distant outgroup set (VIII), we iteratively defined phylostrata by comparison with the next, more closely related outgroup set. The results of this analysis are summarized in Supplemental Table 1. *C. elegans* specific genes are assigned to phylostratum I, whereas genes that are present in the most divergent outgroups are assigned to phylostratum VIII. Orthologs were defined after performing all pairwise BLASTP searches including self-searches (e-value $< 10^{-5}$) between *C. elegans* and *P. pacificus* and extracting best reciprocal hits from the BLAST output files. Simultaneously, the program orthAgogue was run with default setting on the same input files[48].

Data availability

The strand-specific *de novo* transcriptome was submitted to the European Nucleotide Archive under the accession HAKN01000001[16] and the Iso-seq data was submitted to the European Nucleotide Archive under the accessions ERX2315712 and ERX2315713[44]. All data sets are also available at http://www.pristionchus.org/download. The initial set of *P. pacificus* gene annotations corresponds to WormBase WS268. Corrections from this study were submitted to WormBase and will be released following curation.

Received: 1 October 2019; Accepted: 20 November 2019;
Published online: 11 December 2019

References

1. Sommer, R. J., Carta, L. K., Kim, S.-Y. & Sternberg, P. W. Morphological, genetic and molecular description of *Pristionchus pacificus* sp. n. (Nematoda: Neodiplogastridae). *Fundam. Appl. Nematol.* **19**, 511–521 (1996).
2. Sternberg, P. W. Why *Caenorhabditis elegans* is great and *Pristionchus pacificus* might be better. In *Pristionchus pacificus* (ed. Sommer, R. J.) **11**, 1–17 (BRILL).
3. Sommer, R. J. Evolution of regulatory networks: nematode vulva induction as an example of developmental systems drift. *Adv. Exp. Med. Biol.* **751**, 79–91 (2012).
4. Bose, N. *et al*. Complex small-molecule architectures regulate phenotypic plasticity in a nematode. *Angew. Chem. Int. Ed Engl.* **51**, 12438–12443 (2012).
5. Yim, J. J., Bose, N., Meyer, J. M., Sommer, R. J. & Schroeder, F. C. Nematode signaling molecules derived from multimodular assembly of primary metabolic building blocks. *Org. Lett.* **17**, 1648–1651 (2015).
6. Falcke, J. M. *et al*. Linking genomic and metabolomic natural variation uncovers nematode pheromone biosynthesis. *Cell Chem Biol* **25**, 787–796.e12 (2018).
7. Mayer, M. G., Rödelsperger, C., Witte, H., Riebesell, M. & Sommer, R. J. The orphan gene *dauerless* regulates dauer development and intraspecific competition in nematodes by copy number variation. *PLoS Genet.* **11**, e1005146 (2015).
8. Prabh, N. *et al*. Deep taxon sampling reveals the evolutionary dynamics of novel gene families in *Pristionchus* nematodes. *Genome Res.* **28**, 1664–1674 (2018).
9. Rödelsperger, C., Streit, A. & Sommer, R. J. Structure, function and evolution of the nematode genome. In: *eLS*. John Wiley & Sons, Ltd: Chichester, https://doi.org/10.1002/9780470015902.a0024603 (2013).
10. Lightfoot, J. W., Chauhan, V. M., Aylott, J. W. & Rödelsperger, C. Comparative transcriptomics of the nematode gut identifies global shifts in feeding mode and pathogen susceptibility. *BMC Res. Notes* **9**, 142 (2016).
11. Susoy, V. *et al*. Large-scale diversification without genetic isolation in nematode symbionts of figs. *Sci Adv* **2**, e1501031 (2016).
12. Bumbarger, D. J., Riebesell, M., Rödelsperger, C. & Sommer, R. J. System-wide rewiring underlies behavioral differences in predatory and bacterial-feeding nematodes. *Cell* **152**, 109–119 (2013).
13. Witte, H. *et al*. Gene inactivation using the CRISPR/Cas9 system in the nematode *Pristionchus pacificus*. *Dev. Genes Evol.* **225**, 55–62 (2015).
14. Srinivasan, J. *et al*. A bacterial artificial chromosome-based genetic linkage map of the nematode *Pristionchus pacificus*. *Genetics* **162**, 129–134 (2002).
15. Rödelsperger, C. *et al*. Single-molecule sequencing reveals the chromosome-scale genomic architecture of the nematode model organism *Pristionchus pacificus*. *Cell Rep.* **21**, 834–844 (2017).
16. Rödelsperger, C. *et al*. Phylotranscriptomics of *Pristionchus* nematodes reveals parallel gene loss in six hermaphroditic lineages. *Curr. Biol.* **28**, 3123–3127.e5 (2018).
17. Namai, S. & Sugimoto, A. Transgenesis by microparticle bombardment for live imaging of fluorescent proteins in *Pristionchus pacificus* germline and early embryos. *Dev. Genes Evol.* **228**, 75–82 (2018).
18. Lo, T.-W. *et al*. Precise and heritable genome editing in evolutionarily diverse nematodes using TALENs and CRISPR/Cas9 to engineer insertions and deletions. *Genetics* **195**, 331–348 (2013).
19. Bui, L. T. & Ragsdale, E. J. Multiple plasticity regulators reveal targets specifying an induced predatory form in nematodes. *Mol. Biol. Evol*, https://doi.org/10.1093/molbev/msz171 (2019).
20. Ishita, Y., Chihara, T. & Okumura, M. Serotonergic modulation of feeding behavior in *Caenorhabditis elegans* and other related nematodes. *Neurosci. Res*, https://doi.org/10.1016/j.neures.2019.04.006 (2019).
21. Liu, Z. *et al*. Predator-secreted sulfolipids induce defensive responses in *C. elegans*. *Nature Communications* **9** (2018).
22. Moreno, E. *et al*. DAF-19/RFX controls ciliogenesis and influences oxygen-induced social behaviors in *Pristionchus pacificus*. *Evol. Dev.* **20**, 233–243 (2018).
23. Markov, G. V. *et al*. Functional conservation and divergence of *daf-22* paralogs in *Pristionchus pacificus* dauer development. *Mol. Biol. Evol.* **33**, 2506–2514 (2016).
24. Markov, G. V., Baskaran, P. & Sommer, R. J. The same or not the same: lineage-specific gene expansions and homology relationships in multigene families in nematodes. *J. Mol. Evol.* **80**, 18–36 (2015).
25. Namdeo, S. *et al*. Two independent sulfation processes regulate mouth-form plasticity in the nematode. *Development* **145** (2018).
26. Rödelsperger, C. Comparative genomics of gene loss and gain in *Caenorhabditis* and Other Nematodes. In *Methods in Molecular Biology* 419–432 (2018).
27. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
28. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–9 (2006).
29. Hoff, K. J., Lomsadze, A., Borodovsky, M. & Stanke, M. Whole-genome annotation with BRAKER. *Methods Mol. Biol.* **1962**, 65–95 (2019).
30. Cantarel, B. L. *et al*. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).
31. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
32. Lightfoot, J. W. *et al*. Small peptide-mediated self-recognition prevents cannibalism in predatory nematodes. *Science* **364**, 86–89 (2019).
33. Moreno, E. *et al*. Regulation of hyperoxia-induced social behaviour in *Pristionchus pacificus* nematodes requires a novel cilia-mediated environmental input. *Sci. Rep.* **7**, 17550 (2017).
34. Kieninger, M. R. *et al*. The nuclear hormone receptor NHR-40 acts downstream of the sulfatase EUD-1 as part of a developmental plasticity switch in *Pristionchus*. *Curr. Biol.* **26**, 2174–2179 (2016).
35. Baskaran, P. & Rödelsperger, C. Microevolution of duplications and deletions and their impact on gene expression in the Nematode *Pristionchus pacificus*. *PLoS One* **10**, e0131136 (2015).
36. Weller, A. M., Rödelsperger, C., Eberhardt, G., Molnar, R. I. & Sommer, R. J. Opposing forces of A/T-biased mutations and G/C-biased gene conversions shape the genome of the nematode *Pristionchus pacificus*. *Genetics* **196**, 1145–1152 (2014).
37. Prabh, N. & Rödelsperger, C. Divergence, and mixed origin contribute to the emergence of orphan genes in nematodes. *G3* **9**, 2277–2286 (2019).
38. Salzberg, S. L. Next-generation genome annotation: we still struggle to get it right. *Genome Biol.* **20**, 92 (2019).
39. Smythe, A. B., Holovachov, O. & Kocot, K. M. Improved phylogenomic sampling of free-living nematodes enhances resolution of higher-level nematode phylogeny. *BMC Evolutionary Biology* **19** (2019).
40. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
41. McLean, F., Berger, D., Laetsch, D. R., Schwartz, H. T. & Blaxter, M. Improving the annotation of the *Heterorhabditis bacteriophora* genome. *Gigascience* **7** (2018).
42. Hiraki, H. *et al*. Genome analysis of Diploscapter coronatus: insights into molecular peculiarities of a nematode with parthenogenetic reproduction. *BMC Genomics* **18**, 478 (2017).
43. Rödelsperger, C. *et al*. Characterization of genetic diversity in the nematode *Pristionchus pacificus* from population-scale resequencing data. *Genetics* **196**, 1153–1165 (2014).

44. Werner, M. S. *et al*. Young genes have distinct gene structure, epigenetic profiles, and transcriptional regulation. *Genome Res.* **28**, 1675–1687 (2018).
45. Rödelsperger, C., Menden, K., Serobyan, V., Witte, H. & Baskaran, P. First insights into the nature and evolution of antisense transcription in nematodes. *BMC Evol. Biol.* **16**, 165 (2016).
46. Baskaran, P. *et al*. Ancient gene duplications have shaped developmental stage-specific expression in *Pristionchus pacificus*. *BMC Evol. Biol.* **15**, 185 (2015).
47. Domazet-Loso, T., Brajković, J. & Tautz, D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* **23**, 533–539 (2007).
48. Ekseth, O. K., Kuiper, M. & Mironov, V. orthAgogue: an agile tool for the rapid prediction of orthology relations. *Bioinformatics* **30**, 734–736 (2014).
49. Remm, M., Storm, C. E. & Sonnhammer, E. L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314**, 1041–1052 (2001).
50. Tatusov, R. L. A Genomic Perspective on Protein Families. *Science* **278**, 631–637 (1997).
51. Li, L., Stoeckert, C. J. Jr & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
52. Verster, A. J., Ramani, A. K., McKay, S. J. & Fraser, A. G. Comparative RNAi Screens in *C. elegans* and *C. briggsae* Reveal the Impact of Developmental System Drift on Gene Function. *PLoS Genetics* **10**, e1004077 (2014).
53. Sieriebriennikov, B., Markov, G. V., Witte, H. & Sommer, R. J. The Role of DAF-21/Hsp90 in Mouth-Form Plasticity in *Pristionchus pacificus*. *Mol. Biol. Evol.* **34**, 1644–1653 (2017).
54. Moreno, E. & Sommer, R. J. A cilia-mediated environmental input induces solitary behaviour in *Caenorhabditis elegans* and *Pristionchus pacificus* nematodes. *Nematology* **20**, 201–209 (2018).
55. Prabh, N. & Rödelsperger, C. Are orphan genes protein-coding, prediction artifacts, or non-coding RNAs? *BMC Bioinformatics* **17**, 226 (2016).
56. Borchert, N. *et al*. Proteogenomics of *Pristionchus pacificus* reveals distinct proteome structure of nematode models. *Genome Res.* **20**, 837–846 (2010).
57. Sinha, A., Sommer, R. J. & Dieterich, C. Divergent gene expression in the conserved dauer stage of the nematodes *Pristionchus pacificus* and *Caenorhabditis elegans*. *BMC Genomics* **13**, 254 (2012).
58. Stoltzfus, J. D., Minot, S., Berriman, M., Nolan, T. J. & Lok, J. B. RNAseq analysis of the parasitic nematode *Strongyloides stercoralis* reveals divergent regulation of canonical dauer pathways. *PLoS Negl. Trop. Dis.* **6**, e1854 (2012).
59. Sanghvi, G. V. *et al*. Life history responses and gene expression profiles of the nematode *Pristionchus pacificus* cultured on *Cryptococcus* yeasts. *PLoS One* **11**, e0164881 (2016).
60. Serobyan, V. *et al*. Chromatin remodelling and antisense-mediated up-regulation of the developmental switch gene *eud-1* control predatory feeding plasticity. *Nat. Commun.* **7**, 12337 (2016).
61. Moreno, E., McGaughran, A., Rödelsperger, C., Zimmer, M. & Sommer, R. J. Oxygen-induced social behaviours in *Pristionchus pacificus* have a distinct evolutionary history and genetic regulation from *Caenorhabditis elegans*. *Proc. Biol. Sci.* **283**, 20152263 (2016).
62. Buels, R. *et al*. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.* **17**, 66 (2016).
63. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).

Acknowledgements

The authors would like to thank the complete *Pristionchus* community for their long-term interest in studying *P. pacificus* and thus motivating this work. Further thanks to Bogdan Sieriebriennikov for providing additional manual curations and to all members of the Sommer lab for general discussions. Finally, special thanks to Michael Paulini for incorporating the updated gene models into WormBase. This work was funded by the Max Planck Society.

Author contributions

Conceptualization, C.R.; Investigation, C.R., M.A., M.L., T.T., S.S., M.D., S.W., W.H., D.R.S. and Z.H.; Writing – Original Draft, C.R.; Writing – Review & Editing, C.R., M.A., M.L., T.T., S.S., M.D., S.W., W.H., D.R.S. and Z.H.; Supervision, C.R.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-55359-5.

Correspondence and requests for materials should be addressed to C.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Athanasouli et al. BMC Genomics (2020) 21:708
https://doi.org/10.1186/s12864-020-07100-0

BMC Genomics

RESEARCH ARTICLE Open Access

Comparative genomics and community curation further improve gene annotations in the nematode *Pristionchus pacificus*

Marina Athanasouli, Hanh Witte, Christian Weiler, Tobias Loschko, Gabi Eberhardt, Ralf J. Sommer and Christian Rödelsperger*

Abstract

Background: Nematode model organisms such as *Caenorhabditis elegans* and *Pristionchus pacificus* are powerful systems for studying the evolution of gene function at a mechanistic level. However, the identification of *P. pacificus* orthologs of candidate genes known from *C. elegans* is complicated by the discrepancy in the quality of gene annotations, a common problem in nematode and invertebrate genomics.

Results: Here, we combine comparative genomic screens for suspicious gene models with community-based curation to further improve the quality of gene annotations in *P. pacificus*. We extend previous curations of one-to-one orthologs to larger gene families and also orphan genes. Cross-species comparisons of protein lengths, screens for atypical domain combinations and species-specific orphan genes resulted in 4311 candidate genes that were subject to community-based curation. Corrections for 2946 gene models were implemented in a new version of the *P. pacificus* gene annotations. The new set of gene annotations contains 28,896 genes and has a single copy ortholog completeness level of 97.6%.

Conclusions: Our work demonstrates the effectiveness of comparative genomic screens to identify suspicious gene models and the scalability of community-based approaches to improve the quality of thousands of gene models. Similar community-based approaches can help to improve the quality of gene annotations in other invertebrate species, including parasitic nematodes.

Keywords: Genome, Evolution, *Caenorhabditis elegans*, Parasitic nematodes, Orphan genes

Background

The nematode *Pristionchus pacificus* was initially introduced as a satellite model organism for comparing developmental processes to *Caenorhabditis elegans* [1, 2]. More recently, it has emerged as an independent model organism for studying the genetics of phenotypic plasticity [3–5] and behavior [6–8], interactions between host and microbes [9–11], and genome evolution [12–14]. Central to all these studies was the genome sequence of *P. pacificus*, which has undergone continuous improvements over time [15–17]. However, until recently, its gene annotations were almost exclusively based on automated pipelines that combined gene predictions and evidence-based annotations [18–20]. As a consequence, the gene annotations of *P. pacificus* did not match the quality of the highly curated *C. elegans* genome. This made it difficult for researchers from the *C. elegans* field to adapt *P. pacificus* for comparative studies, even though the availability of genetic toolkits including transgenic reporter lines and gene knockouts makes *P. pacificus* ideally suited for comparative studies of gene function [8, 21, 22]. Therefore, we

* Correspondence: christian.roedelsperger@tuebingen.mpg.de
Department for Integrative Evolutionary Biology, Max Planck Institute for Developmental Biology, Max-Planck-Ring 9, 72076 Tübingen, Germany

have recently started to combine comparative genomic screens for suspicious gene models with community-based manual curation to improve the quality of the gene annotations in *P. pacificus* [23]. This pilot study screened for missing one-to-one orthologs of *C. elegans* genes in *P. pacificus*. Community-based curation of these candidate gene loci resulted in a substantial improvement of the *P. pacificus* gene annotations (version: El Paco annotation V2). Precisely, when assessed by benchmarking of universally conserved single copy orthologs (BUSCO) [24], the completeness level increased from 86 to 97%. Most missing orthologs were due to fused gene models some of which had long untranslated regions (UTRs) that actually contained complete genes. These errors could be corrected by manual inspection of the suspicious gene loci under the consideration of two recent transcriptome assemblies that were generated from strand-specific RNA-seq [25, 26] and Iso-seq data [27].

Here, we employ comparative genomic approaches to screen for further errors in other gene classes including large gene families that have undergone lineage-specific duplications [28] and species-specific orphan genes (SSOGs) [29] that were not the focus of our previous study [23]. Candidate loci are then curated by community-based manual inspection and eventually, corrections were proposed mainly based on available transcriptome assemblies. Overall, we investigated 4311 suspicious gene models and implemented 2946 corrections. This resulted in a further improved set of gene annotations for *P. pacificus*. Similar community-based curation approaches can help approving gene annotations in other nematode genomes including those of animal and plant parasites [23].

Results

Protein length comparison of orthologs identify hundreds of suspicious gene models

In our previous study, we focused on the identification of missing one-to-one orthologous genes in the *P. pacificus* genome and the identification of artificial fusions between two adjacent *P. pacificus* genes both of which have one-to-one orthologous genes [23]. Here, we aim to further improve the quality of one-to-one orthologous genes by finding and curating *P. pacificus* genes that are either unusually large or small with regard to their *C. elegans* counterpart. We performed a comparison of protein length of 8348 one-to-one orthologs between *C. elegans* and *P. pacificus* (Fig. 1a-c). Protein lengths between one-to-one orthologs are well correlated (Pearson's r = 0.83, Fig. 1a). However, there are slight differences in the length distributions (Fig. 1b,c) and using an arbitrary cutoff of a two-fold difference in protein length, we defined 532 *P. pacificus* genes as candidates for manual inspection. For example, in the case of the *P. pacificus* gene PPA00494 (ortholog of *C. elegans lev-8*), its predicted protein sequence encompasses 1094 amino acids, which is more than twice as long as *C. elegans* LEV-8 (531 amino acids) (Fig. 1d). Also, BLASTP analysis against the *C. elegans* proteins (version WS277) shows that the N-terminal part of PPA00494 is homologous to another *C. elegans* protein, Y73B6BL.37 (Fig. 1d), suggesting that it could represent an artificial gene fusion. Subsequent inspection in the genome browser showed two transcripts that were assembled from strand-specific RNA-seq data [26], which span the PPA00494 locus (Fig. 1e). This strongly supports that PPA00494 should be split by replacing it with the two assembled transcripts. After community-based curation, 309 (57%) corrections were proposed. The remaining cases were judged as either inconclusive (due to the lack of transcriptomic support) or correct. These results demonstrate that protein length comparisons between one-to-one orthologs are an effective way to identify suspicious gene models and to further improve the quality of one-to-one orthologs.

Analysis of protein domains identifies further artificial gene fusions

Our previous study showed that the combination of incorrectly predicted gene boundaries and overlapping UTRs between neighboring genes in regions with high gene density most likely caused artificial gene fusions. In order to screen for further cases of artificial gene fusions, we applied a comparative genomic approach to identify proteins with atypical domain combinations that do not exist in other nematodes such as *C. elegans*, and more distantly related *Bursaphelenchelus xylophilus* [30], and *Strongyloides ratti* [31]. This yielded 1589 *P. pacificus* candidates (Table 1) for further inspection. Note, that such atypical domain combinations are not necessarily artifacts. For example, the same screen in the highly curated *C. elegans* genome, identified 932 genes with atypical domain combinations. Manual inspection of these gene models and available transcriptome assemblies in the WormBase genome browser (WS177) combined with BLASTP analysis against *C. briggsae* revealed three candidates for putatively incorrect annotation in *C. elegans*, which deserve closer inspection (Additional file 1, Figure S1). After community curation of the *P. pacificus* candidates, corrections were proposed for 695 (44%) candidates. Next, we defined 1388 unusually small or long members of 25 highly abundant gene families as further candidates for manual inspection (Fig. 2a). After community curation, corrections were proposed for 420 (32%) of these candidates. The three described screens partially identify the same candidates (Fig. 2b), yet the presence of hundreds of candidate genes that are specific to each method indicates how complementary these different approaches are.

Gene prediction artifacts are a likely source of SSOGs

A previous analysis of *P. pacificus* orphan genes revealed that the majority of SSOGs had no transcriptomic support

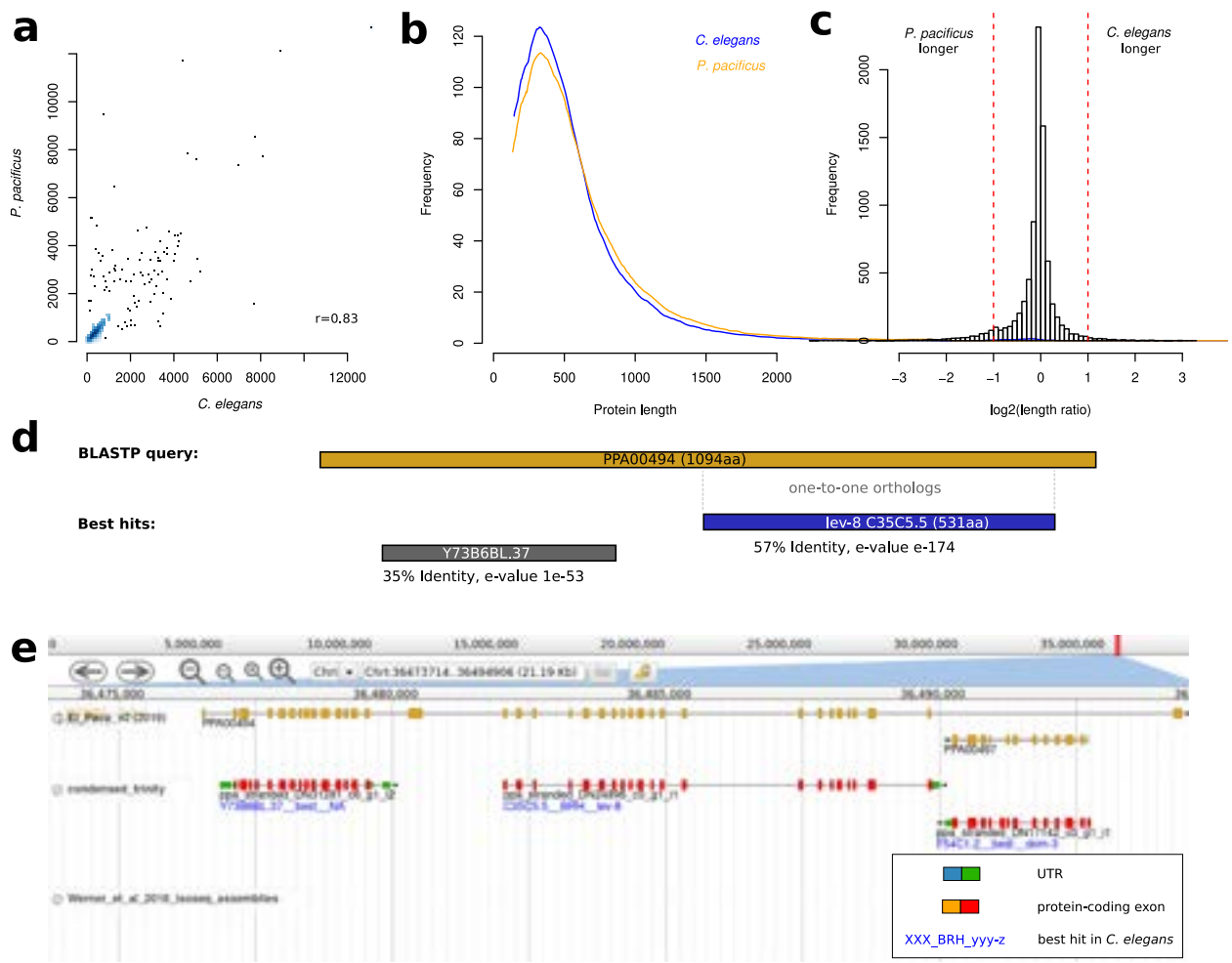


Fig. 1 Comparison of protein lengths between one-to-one orthologs. **a** One-to-one orthologous genes between *C. elegans* and *P. pacificus* have highly similar protein lengths (Pearson's r = 0.83). **b** Size distributions of one-to-one orthologs show a peak at around 300 amino acids. **c** *P. pacificus* genes with more than two-fold length difference were considered for manual curation. **d** The *P. pacificus* one-to-one ortholog (PPA0494) of *C. elegans lev-8*, is more than twice as long as LEV-8. BLAST analysis showed that the N-terminal region has similarity to another *C. elegans* gene (Y37B6BL.37) suggesting that it represents an artificial gene fusion. **e** Manual inspection of the PPA0494 in the genome browser shows that there are two assembled RNA-seq transcripts (red) that cover most of the original gene model and further support that PPA0494 is an artificially fused gene model

[14]. Based on the reanalysis of the current gene annotations with available phylogenomic and phylotranscriptomic data [26, 32], we identified 1988 (7%) *P. pacificus* SSOGs of which 314 were classified as having transcriptomic support. Manual inspection of the remaining SSOGs classified 678 (41%) of candidates as not having any transcriptional support (Fig. 2c), even when considering additional transcriptomic data sets such as iso-seq or dauer-specific transcriptomes [27, 33]. Further 196 (12%) of SSOG candidates showed some transcriptional activity, but this expression data was mostly not sufficient to support their gene structure. Strikingly, we found 704 (42%) and 46 (3%) SSOGs, which overlapped existing gene models on the antisense strand of protein-coding exons and UTRs, respectively (Fig. 2c and 3a,b). However, visual inspection of transcriptomic data only supported the sense gene as opposed to the antisense SSOG. As there is neither protein homology nor transcriptional data supporting these antisense SSOGs, we would tend to argue that these spurious antisense gene models most likely derive from the contribution of gene prediction softwares SNAP and AUGUSTUS during the process of the original gene annotation [17, 19, 20]. Thus, manual curation removed 1515 of the unsupported SSOGs, mainly from the "no support", "Antisense prediction", and "UTR" categories (Fig. 2c), as their lack of transcriptional evidence makes it difficult to conclusively study the process of novel gene formation [14, 29].

New *P. pacificus* gene annotations show increased homogeneity and better reflect existing RNA-seq data

In total, we visually inspected 4311 suspicious gene models and proposed corrections for 2946 (68%). The

Table 1 Comparative assessment of different *P. pacificus* gene annotations

Category	*P. pacificus* El Paco gene annotations	
	V2	V3
Number of genes	28,036	28,896
Protein-coding sequence (Mb)	35.3	35.3
BUSCO Completeness (%)	97.1	97.6
BUSCO Duplicated (%)	1.7	1.8
BUSCO Fragmented (%)	2.0	2.0
BUSCO Missing (%)	0.9	0.4
Number of 1–1 orthologs (BRHs)	8348	8607
Number of 1–1 orthologs with variable protein length (%)	532	265
Number of proteins with atypical domain combinations	1589	1137
Number of protein family length outlier	1388	1201

The table shows an overview about general characteristics of different *P. pacificus* gene annotations

most common errors were artificial fusions and unsupported SSOGs that are putative annotations artifacts. This led to the correction of 1367 gene models by replacing them with one or more alternative gene models and to the removal of 1579 gene models. We implemented all proposed corrections into a new *P. pacificus* gene annotation (version: El Paco gene annotation V3), which comprises 28,896 gene models and spans 35.2 Mb of protein-coding sequence with a BUSCO completeness level of 97.6% (Table 1). As expected, the numbers of one-to-one orthologs with length differences, the number of genes with atypical domain combinations, and the number of gene family outliers went down by 10–50%. To additionally test if the new set of gene annotations better captures RNA-seq data sets, we reanalyzed 15 RNA-seq data sets from four different studies [9, 13, 34, 35] and quantified the percentage of reads that could be assigned to features of the gene annotations. The new set of gene annotations consistently captures 2% more of the RNA-seq alignments (Table 2). Despite the fact that the number of genes was increased by ~ 3%, the total amount of annotated protein-coding sequence remained almost unaltered (Table 1). Thus, the new set of gene annotations better reflects RNA-seq data.

Discussion

In the early genomic era, gene annotation was heavily dependent on automated gene finding algorithms that tried to recognize gene structures based on statistical sequence properties of exons, introns, and splicing sites [19, 20]. This was highly suited when functional data, e.g. expressed sequence tags and cDNAs, were scarce and the only way to annotate a complete genome was to extract informative sequence features from a limited test set and extrapolate them to the whole genome. However, with the dramatic improvement of sequencing protocols and technologies, it became feasible to generate evidence-based gene annotations from transcriptome and homology data [18, 36]. Under the consideration that related genomes at an optimal evolutionary distance to a focal organism and transcriptomic evidence for all genes are rarely available, this still justifies the usage of gene prediction tools. In the case of the *P. pacificus*, previous versions of gene annotations that were completely based on the results of gene prediction tools were suited to perform evolutionary genomic analysis and genetic screens [37, 38]. Subsequently, we employed the widely used MAKER2 pipeline to generate a more comprehensive gene annotation by integration of large-scale transcriptomic and protein homology data as well as gene predictions [17, 36]. Comparative analysis of genome quality for 22 nematode species revealed that already these gene annotations (version: El Paco annotation V1) were of relatively high quality (86% BUSCO completeness) [23]. Nevertheless, the question of how good gene annotations need to be will depend on what researchers want to do with them. Reverse genetic studies in nematodes with well established genetic toolkits are extremely powerful systems for comparative studies of gene function [8, 22] and the evolution of the nervous system and associated behaviors [39, 40]. Yet, the identification of *P. pacificus* orthologs for candidate *C. elegans* genes with known function is complicated by the widespread abundance of lineage-specific duplications [28, 33], but also by the difference in the quality of gene annotations. Facilitating the easy adaptation of *P. pacificus* as a comparative model system for *C. elegans* researchers, who are used to working with one of the best and well-characterized genomes, is one of our main motivations for this study. The chromosome-scale assembly of *P. pacificus* has already been a major step to minimize the disparity between the genomic resources of both species [17]. Lifting up the quality of gene annotations to a comparable level will thus further increase the attractiveness of the *P. pacificus* system for evolutionary studies.

Another motivation for continuous efforts in improving the quality of gene annotations is our focus on the origin and evolution of orphan genes in *P. pacificus* [41–43]. Initially, around one-third of the *P. pacificus* gene repertoire was defined as orphan genes without homology in the genomes of other nematode families [16, 37]. Unbiased genetic screens have identified orphan genes that control important biological processes such as developmental decisions and predatory behavior [6, 42]. Phylogenomic investigation of ten diplogastrid genomes revealed the evolutionary dynamics of

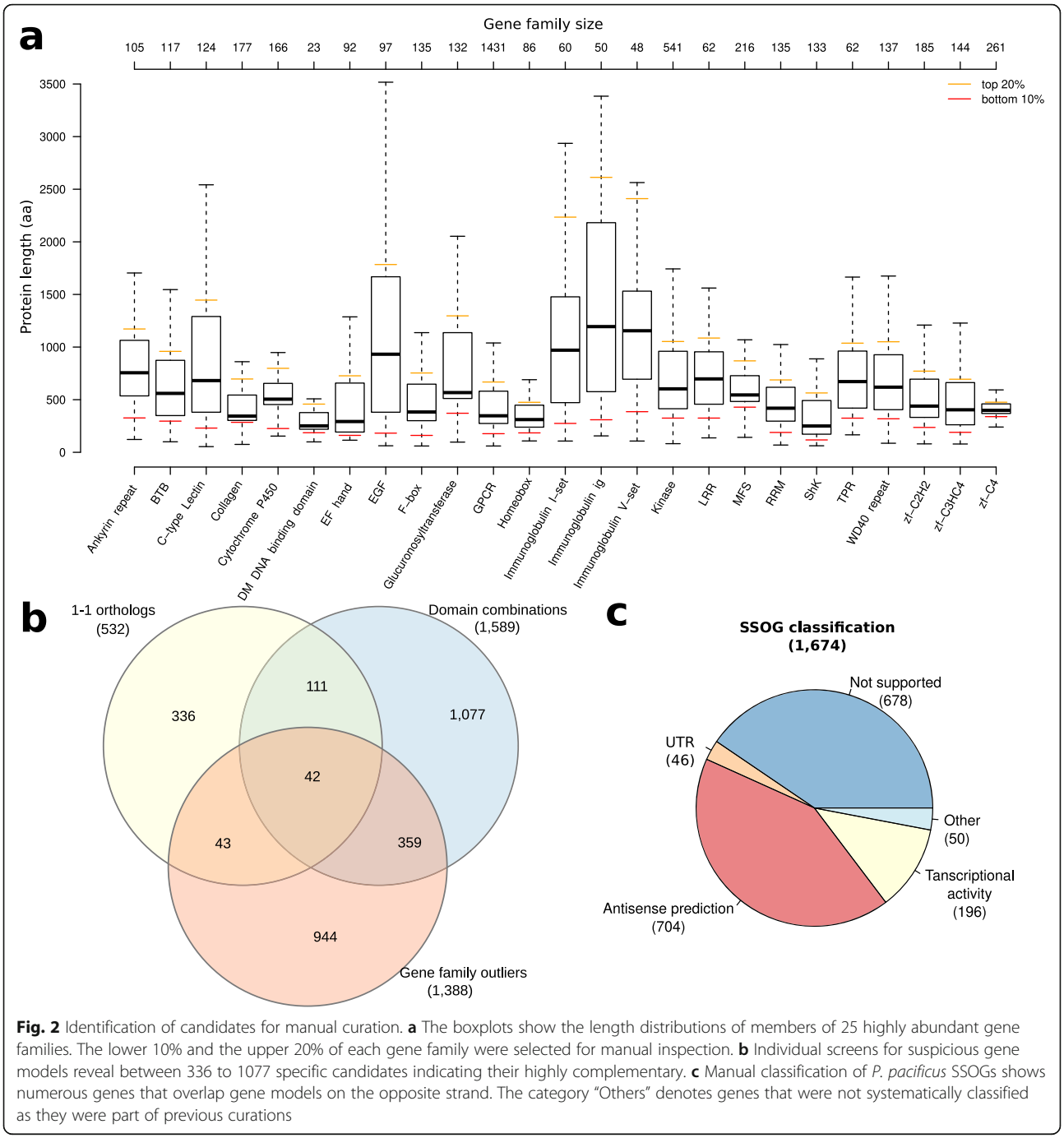


Fig. 2 Identification of candidates for manual curation. **a** The boxplots show the length distributions of members of 25 highly abundant gene families. The lower 10% and the upper 20% of each gene family were selected for manual inspection. **b** Individual screens for suspicious gene models reveal between 336 to 1077 specific candidates indicating their highly complementary. **c** Manual classification of *P. pacificus* SSOGs shows numerous genes that overlap gene models on the opposite strand. The category "Others" denotes genes that were not systematically classified as they were part of previous curations

these novel genes and built the framework to dissect the diversity of mechanisms of origin [14, 32]. When we screened for high quality SSOG candidates for origin analysis, we found that the majority of SSOGs had no transcriptomic support. Together with the finding that SSOGs constitute an unusually large age class (phylostratum), this made us wonder to what extent this gene class might possibly be inflated by gene annotation artifacts [14]. Therefore, we revisited 1674 candidates and confirmed that most of them indeed show no evidence of transcription. In addition, we found 704 SSOGs, which overlapped other gene models on the antisense strand and whenever available, strand-specific RNA-seq did not support the SSOG gene model. Even though SSOGs are expected to show little or no evidence of expression and we cannot conclusively argue that these gene models are annotations artifacts (they represent coding potential that might be

Fig. 3 Examples of unsupported SSOGs. **a** The *P. pacificus* SSOG PPA46345 overlaps exons of two other gene models that are well supported by transcriptome assemblies from strand-specific RNA-seq and Iso-seq data. **b** The *P. pacificus* SSOG PPA4618 overlaps the UTR of a well supported gene model. The absence of strand-specific transcriptomic support indicates that *P. pacificus* SSOGs PPA46345 and PPA4618 are likely gene prediction artifacts

used under some conditions), for practical reasons we chose to remove most of the unsupported SSOGs to allow future investigations of orphan origin to start with a set of well supported candidate SSOGs. Thus, we hope that the community-based curation of the *P. pacificus* gene annotations will help future studies in many aspects of evolutionary biology.

Conclusions

Our work demonstrates that even for non-classical model organisms with small research communities, manual inspection and curation of thousands of genes can be achieved. Thereby numerous comparative genomic screens can be applied to enrich the candidate set for suspicious gene models that actually need to be corrected. The example of the highly curated *P. pacificus* genome emphasizes the effectiveness and scalability of manual curation for many other genome projects including those of nematode animal and plant parasites.

Methods

Candidate identification based on length comparison of orthologous proteins

We obtained 8348 one-to-one orthologs between *C. elegans* and *P. pacificus* that were predicted based on best reciprocal BLASTP hits in a previous study [23]. We then calculated the protein length ratio between the *P. pacificus* and *C. elegans* one-to-one orthologs. In case of multiple isoforms for a given gene, we chose the isoform with the longest protein sequence (WormBase release WS260). Based on an arbitrary cutoff of a two-fold difference in protein length between the two species, we identified 532 *P. pacificus* candidates for manual curation.

Candidate identification based on protein domain content

We ran the hmmsearch program of the HMMER package (version 3.0, e-value < 0.001, profiles from PFAM-A.hmm) on protein sets of *C. elegans* (WS260), *P. pacificus* (El Paco

Table 2 Comparison of RNA-seq read alignability

P. pacificus RNA-seq samples		Successfully assigned alignments (%)		Reference
Accession	Description	V2	V3	
ERR777792	Mixed-stage on *E. coli* OP50	74.8	76.8	[13]
ERR777793	Mixed-stage on *E. coli* OP50	74.9	76.6	[13]
ERR777794	Mixed-stage on *E. coli* OP50	74.4	76.1	[13]
SRR4017216	Adults on *E. coli* OP50	79.8	81.7	[34]
SRR4017217	Adults on *E. coli* OP50	80.3	82.2	[34]
SRR4017218	Adults on *Cryptococcus* C3	79.6	81.6	[34]
SRR4017219	Adults on *Cryptococcus* C3	79.2	81.1	[34]
SRR4017220	Adults on *Cryptococcus* C5	79.9	81.8	[34]
SRR4017221	Adults on *Cryptococcus* C5	80.7	82.6	[34]
ERR3421261	Adults on *E. coli* OP50	79.7	81.6	[9]
ERR3421262	Adults on *E. coli* OP50	79.5	81.3	[9]
ERR3421263	Adults on *Novosphingobium* L76	79.6	81.5	[9]
ERR3421264	Adults on *Novosphingobium* L76	79.5	81.5	[9]
SRR2142256	Adults on *E. coli* OP50	77.8	79.8	[35]
SRR2142257	Intestines	72.5	74.2	[35]

The table shows the percentage of assigned reads from 15 RNA-seq experiments for different *P. pacificus* gene annotations

annotation V2), *B. xylophilus* (WS248), *S. ratti* (WS260). We counted occurrences of protein domains and defined as candidates, domain combinations that are unique to *P. pacificus* and occur at low frequencies (less than ten times). This yielded 1589 candidates with atypical protein domain combinations. Next, we selected 25 highly abundant gene families such as collagens and C-type lectins that were defined by a PFAM domain and classified further candidate proteins if their length fell under the first or above the eighth decile of the length distribution of all members of a given gene family. This identified 1388 candidate genes for manual curation.

Identification and curation of *P. pacificus* species-specific orphan genes

We defined *P. pacificus* SSOGs by BLASTP searches of the *P. pacificus* proteins (version: El Paco annotation V2) against annotated protein sets and predicted open reading frames (ORFs) in assembled transcripts of *P. exspectatus*, *P. arcanus*, *P. maxplancki,* and *P. japonica* [26, 32]. This identified 1988 (7%) *P. pacificus* SSOGs without a BLASTP hit in any of the reference data sets (e-value < 0.001). Three hundred fourteen SSOGs showed transcriptomic support as they had a BLASTP hit in ORFs of the *P. pacificus* transcriptome assembly. The remaining 1674 were defined as SSOGs without transcriptomic support and were thus considered as candidates for manual curation.

Community-based manual curation of gene models

Community-based gene curation was performed as described in our pilot study [23]. In short, candidate lists were shared in online spreadsheets and individual genes were visually inspected in the jbrowse genome browser instance on http://www.pristionchus.org [44]. Based on available transcriptomic resources, which include RNA-seq data from different developmental stages, strand-specific transcriptome assemblies from mixed-stage cultures [25, 26], and iso-seq data [27], community curators were trained to evaluate whether a locus was well covered by transcriptomic data and in case of evidence for an artificial gene fusion to propose the replacement of the original gene model by assembled transcripts. If the genomic neighborhood of the candidate genes showed obvious inconsistencies between original gene models and transcriptome data, we eventually curated such neighboring genes. However, we omitted any gene that was curated in our previous study, as these changes were not yet fully implemented in the latest WormBase release WS177 of *P. pacificus* and we wanted to avoid version conflicts. While for most candidate genes, we did not propose any correction in case that available transcriptomic data was insufficient to make a conclusive statement, in the case of SSOGs, we typically removed the gene model if not at least some RNA-seq data supported the gene structure.

Quality assessment of gene annotations

In order to evaluate the quality of gene annotations, we ran the BUSCO program (version 3.0.1) in protein mode

(option: -m prot) against the nematode_odb9 data set ($N = 982$ orthologs) [24]. To test whether the new set of gene annotations better captures RNA-seq data, we downloaded 15 RNA-seq data sets from the European Nucleotide Archive and aligned these data sets against the *P. pacificus* reference genome (version: El Paco) with the help of the STAR aligner (version: 2.5.4b, default options, reference was the *P. pacificus* genome without any gene annotation) [45]. Next, we quantified the percentage of alignments that could be assigned to gene annotations using the featureCounts function of the Rsubread library in R (version 4.0.0).

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-020-07100-0.

Additional file 1: Figure S1. Candidate genes for incorrect gene annotations in *C. elegans*.

Abbreviations

BUSCO: Benchmarking Universal Single Copy Orthologs; ORF: Open Reading Frame; SSOG: Species-specific Orphan Gene; UTR : Untranslated Region

Acknowledgements

The authors would like to thank the complete *Pristionchus* community for their long-term interest in studying *P. pacificus* and thus motivating this work.

Authors' contributions

Conceptualization, C.R.; Investigation, C.R., M.A., H.W., C.W., T.L. and G.E.; Data curation, C.R., M.A., H.W., C.W., T.L. and G.E.; Visualization, C. R.; Writing original draft, C.R.; Writing – review & editing, C.R., and R.J.S.; Project administration, C. R.; Supervision, C.R. and R.J.S.; Funding acquisition, R.J.S. The author(s) read and approved the final manuscript.

Funding

This work was funded by the Max Planck Society. Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

The new set of *P. pacificus* gene annotations (version: El Paco gene annotation V3) was submitted as an update of the existing whole-genome shotgun project at ENA/Genbank: ABKE00000000. The annotations were also submitted to WormBase where they will be published following further curation and they are publicly available at http://www.pristionchus.org/download/.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

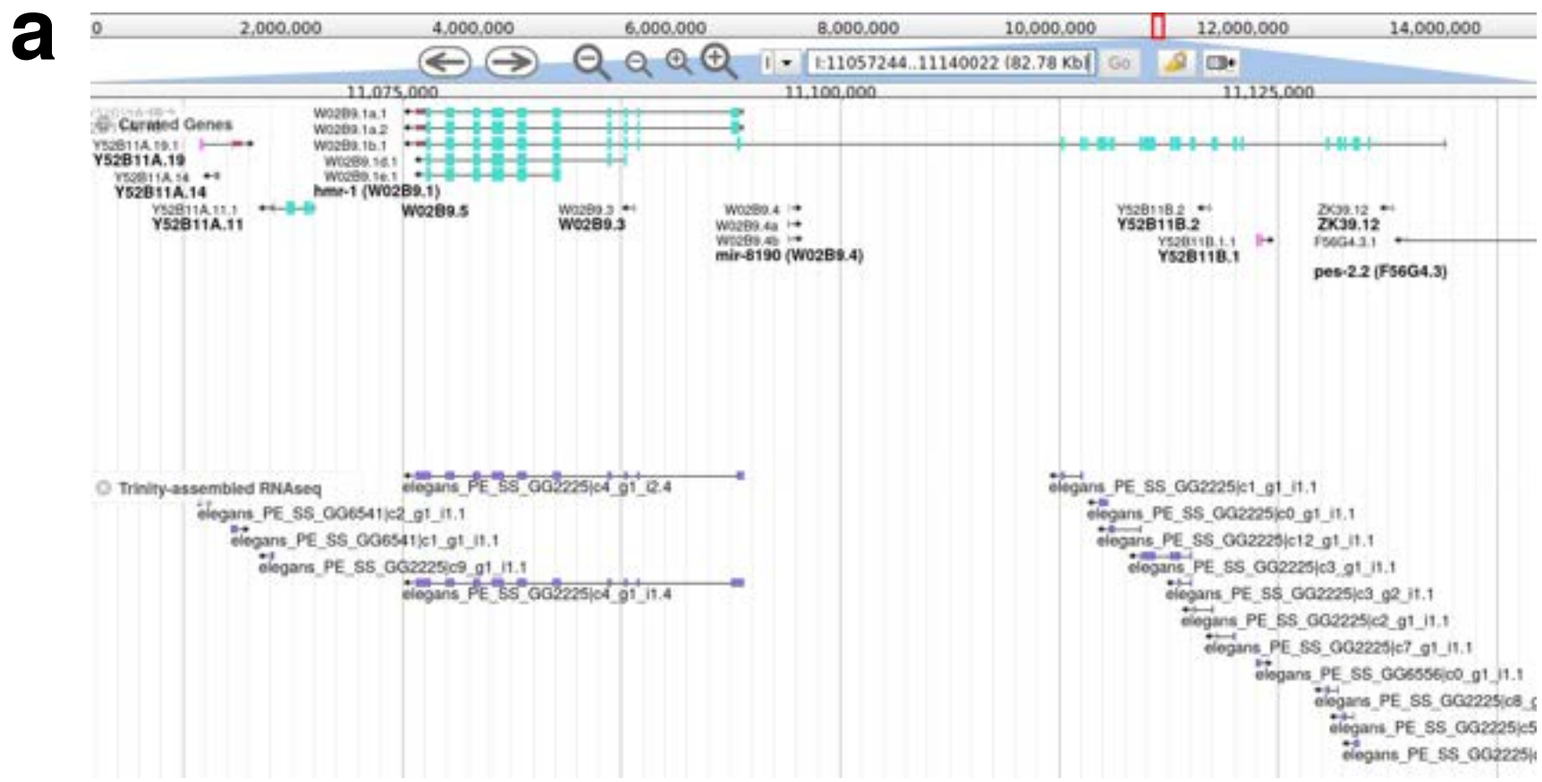
Received: 3 August 2020 Accepted: 23 September 2020
Published online: 12 October 2020

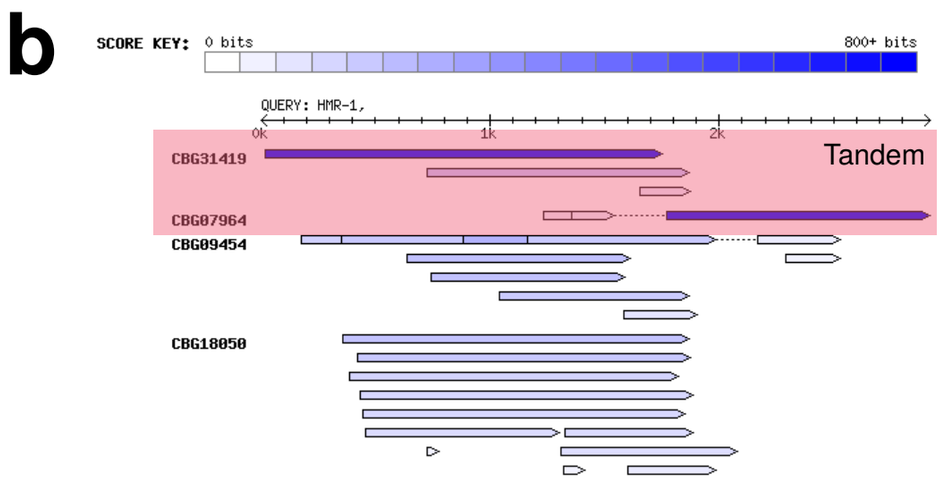
References

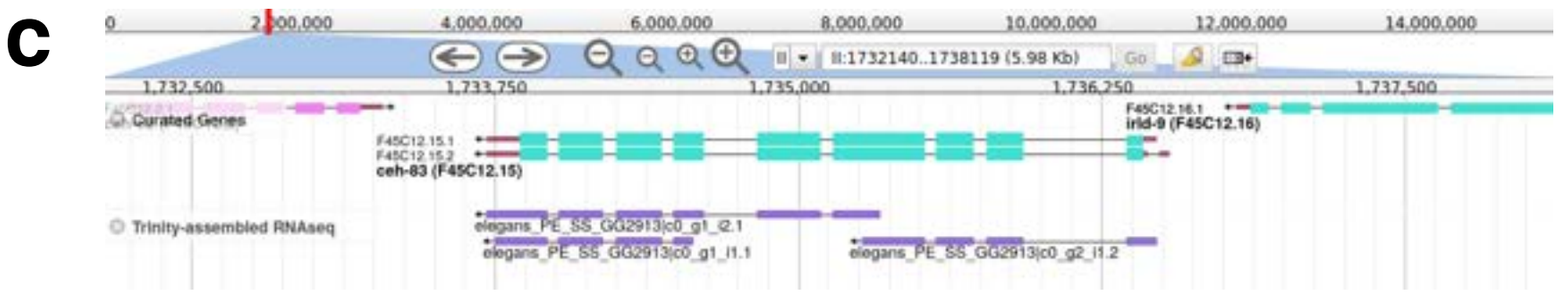
1. Sommer RJ, Carta L, Kim S-Y, Sternberg PW. Morphological, genetic and molecular description of *Pristionchus pacificus* sp. n.(Nematoda: Neodiplogasteridae). Fundam Appl Nematol. 1996;19:511–22.
2. Sommer RJ. The future of evo–devo: model systems and evolutionary theory. Nat Rev Genet. 2009;10:416–22. https://doi.org/10.1038/nrg2567.
3. Kieninger MR, Ivers NA, Rödelsperger C, Markov GV, Sommer RJ, Ragsdale EJ. The nuclear hormone receptor NHR-40 acts downstream of the sulfatase EUD-1 as part of a developmental plasticity switch in *Pristionchus*. Curr Biol. 2016;26:2174–9.
4. Sieriebriennikov B, Prabh N, Dardiry M, Witte H, Röseler W, Kieninger MR, et al. A developmental switch generating phenotypic plasticity is part of a conserved multi-gene locus. Cell Rep. 2018;23:2835–43.e4.
5. Sieriebriennikov B, Sun S, Lightfoot JW, Witte H, Moreno E, et al. Conserved nuclear hormone receptors controlling a novel plastic trait target fast-evolving genes expressed in a single cell. PLoS Genet. 2020;16:e1008687. https://doi.org/10.1371/journal.pgen.1008687.
6. Lightfoot JW, Wilecki M, Rödelsperger C, Moreno E, Susoy V, Witte H, et al. Small peptide–mediated self-recognition prevents cannibalism in predatory nematodes. Science. 2019;364:86–9. https://doi.org/10.1126/science.aav9856.
7. Moreno E, McGaughran A, Rödelsperger C, Zimmer M, Sommer RJ. Oxygen-induced social behaviours in *Pristionchus pacificus* have a distinct evolutionary history and genetic regulation from *Caenorhabditis elegans*. Proc Biol Sci. 2016;283:20152263.
8. Moreno E, Sieriebriennikov B, Witte H, Rödelsperger C, Lightfoot JW, Sommer RJ. Regulation of hyperoxia-induced social behaviour in *Pristionchus pacificus* nematodes requires a novel cilia-mediated environmental input. Sci Rep. 2017;7:17550.
9. Akduman N, Lightfoot JW, Röseler W, Witte H, Lo W-S, et al. Bacterial vitamin B12 production enhances nematode predatory behavior. ISME J. 2020;14:1494–507. https://doi.org/10.1038/s41396-020-0626-2.
10. Meyer JM, Baskaran P, Quast C, Susoy V, Rödelsperger C, Glöckner FO, et al. Succession and dynamics of *Pristionchus* nematodes and their microbiome during decomposition of *Oryctes borbonicus* on La Réunion Island. Environ Microbiol. 2017;19:1476–89. https://doi.org/10.1111/1462-2920.13697.
11. Akduman N, Rödelsperger C, Sommer RJ. Culture-based analysis of Pristionchus-associated microbiota from beetles and figs for studying nematode-bacterial interactions. PLoS One. 2018;13:e0198018.
12. Weller AM, Rödelsperger C, Eberhardt G, Molnar RI, Sommer RJ. Opposing forces of a/T-biased mutations and G/C-biased gene conversions shape the genome of the nematode *Pristionchus pacificus*. Genetics. 2014;196:1145–52. https://doi.org/10.1534/genetics.113.159863.
13. Baskaran P, Rödelsperger C. Microevolution of duplications and deletions and their impact on gene expression in the nematode *Pristionchus pacificus*. PLoS One. 2015;10:e0131136.
14. Prabh N, Rödelsperger C. De novo, divergence, and mixed origin contribute to the emergence of orphan genes in *Pristionchus* nematodes. G3: Genes|Genomes|. Genetics. 2019;9:2277–86. https://doi.org/10.1534/g3.119.400326..
15. Dieterich C, Clifton SW, Schuster LN, Chinwalla A, Delehaunty K, Dinkelacker I, et al. The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism. Nat Genet. 2008;40:1193–8.
16. Borchert N, Dieterich C, Krug K, Schütz W, Jung S, Nordheim A, et al. Proteogenomics of *Pristionchus pacificus* reveals distinct proteome structure of nematode models. Genome Res. 2010;20:837–46.
17. Rödelsperger C, Meyer JM, Prabh N, Lanz C, Bemm F, Sommer RJ. Single-molecule sequencing reveals the chromosome-scale genomic architecture of the nematode model organism *Pristionchus pacificus*. Cell Rep. 2017;21:834–44.
18. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008;18:188–96.
19. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004;5:59.
20. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006;34(Web Server issue):W435–9.
21. Markov GV, Meyer JM, Panda O, Artyukhin AB, Claaßen M, Witte H, et al. Functional conservation and divergence of daf-22 paralogs in *Pristionchus pacificus* dauer development. Mol Biol Evol. 2016;33:2506–14.
22. Okumura M, Wilecki M, Sommer RJ. Serotonin drives predatory feeding behavior via synchronous feeding rhythms in the nematode. G3. 2017;7:3745–55.
23. Rödelsperger C, Athanasouli M, Lenuzzi M, Theska T, Sun S, Dardiry M, et al. Crowdsourcing and the feasibility of manual gene annotation: a pilot study in the nematode *Pristionchus pacificus*. Sci Rep. 2019;9:18789.
24. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.

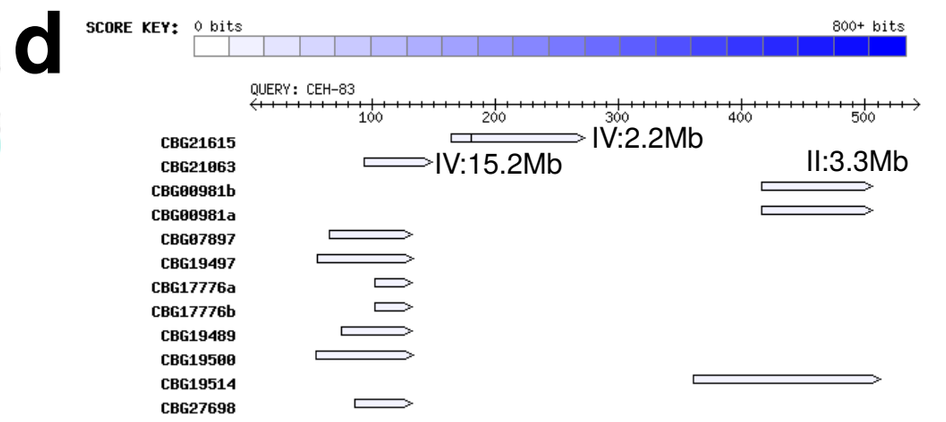
25. Rödelsperger C, Menden K, Serobyan V, Witte H, Baskaran P. First insights into the nature and evolution of antisense transcription in nematodes. BMC Evol Biol. 2016;16:165.
26. Rödelsperger C, Röseler W, Prabh N, Yoshida K, Weiler C, Herrmann M, et al. Phylotranscriptomics of *Pristionchus* nematodes reveals parallel gene loss in six hermaphroditic lineages. Curr Biol. 2018;28:3123–7.e5.
27. Werner MS, Sieriebriennikov B, Prabh N, Loschko T, Lanz C, Sommer RJ. Young genes have distinct gene structure, epigenetic profiles, and transcriptional regulation. Genome Res. 2018;28:1675–87.
28. Markov GV, Baskaran P, Sommer RJ. The same or not the same: lineage-specific gene expansions and homology relationships in multigene families in nematodes. J Mol Evol. 2015;80:18–36.
29. Rödelsperger C, Prabh N, Sommer RJ. New gene origin and deep taxon phylogenomics: opportunities and challenges. Trends Genet. 2019;35:914–22.
30. Kikuchi T, Cotton JA, Dalzell JJ, Hasegawa K, Kanzaki N, McVeigh P, et al. Genomic insights into the origin of parasitism in the emerging plant pathogen *Bursaphelenchus xylophilus*. PLoS Pathog. 2011;7:e1002219.
31. Hunt VL, Tsai IJ, Coghlan A, Reid AJ, Holroyd N, Foth BJ, et al. The genomic basis of parasitism in the *Strongyloides* clade of nematodes. Nat Genet. 2016;48:299–307.
32. Prabh N, Roeseler W, Witte H, Eberhardt G, Sommer RJ, Rödelsperger C. Deep taxon sampling reveals the evolutionary dynamics of novel gene families in *Pristionchus* nematodes. Genome Res. 2018;28:1664–74.
33. Baskaran P, Rödelsperger C, Prabh N, Serobyan V, Markov GV, Hirsekorn A, et al. Ancient gene duplications have shaped developmental stage-specific expression in *Pristionchus pacificus*. BMC Evol Biol. 2015;15:185.
34. Sanghvi GV, Baskaran P, Röseler W, Sieriebriennikov B, Rödelsperger C, Sommer RJ. Life history responses and gene expression profiles of the nematode *Pristionchus pacificus* cultured on *Cryptococcus* yeasts. PLoS One. 2016;11:e0164881. https://doi.org/10.1371/journal.pone.0164881.
35. Lightfoot JW, Chauhan VM, Aylott JW, Rödelsperger C. Comparative transcriptomics of the nematode gut identifies global shifts in feeding mode and pathogen susceptibility. BMC Res Notes. 2016;9:142.
36. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011;12:491.
37. Rödelsperger C, Sommer RJ. Computational archaeology of the *Pristionchus pacificus* genome reveals evidence of horizontal gene transfers from insects. BMC Evol Biol. 2011;11. https://doi.org/10.1186/1471-2148-11-239.
38. Rae R, Witte H, Rödelsperger C, Sommer RJ. The importance of being regular: *Caenorhabditis elegans* and *Pristionchus pacificus* defecation mutants are hypersusceptible to bacterial pathogens. Int J Parasitol. 2012;42:747–53. https://doi.org/10.1016/j.ijpara.2012.05.005.
39. Bumbarger DJ, Riebesell M, Rödelsperger C, Sommer RJ. System-wide rewiring underlies behavioral differences in predatory and bacterial-feeding nematodes. Cell. 2013;152:109–19.
40. Hong RL, Riebesell M, Bumbarger DJ, Cook SJ, Carstensen HR, Sarpolaki T, et al. Evolution of neuronal anatomy and circuitry in two highly divergent nematode species. Elife. 2019;8. https://doi.org/10.7554/eLife.47155.
41. Rödelsperger C. Comparative genomics of gene loss and gain in *Caenorhabditis* and other nematodes. Comparative Genomics. 2018:419–32. https://doi.org/10.1007/978-1-4939-7463-4_16.
42. Mayer MG, Rödelsperger C, Witte H, Riebesell M, Sommer RJ. The orphan gene dauerless regulates dauer development and intraspecific competition in nematodes by copy number variation. PLoS Genet. 2015;11:e1005146.
43. Prabh N, Rödelsperger C. Are orphan genes protein-coding, prediction artifacts, or non-coding RNAs? BMC Bioinformatics. 2016;17:226.
44. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. JBrowse: a next-generation genome browser. Genome Res. 2009;19:1630–8. https://doi.org/10.1101/gr.094607.109.
45. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

Publisher's Note









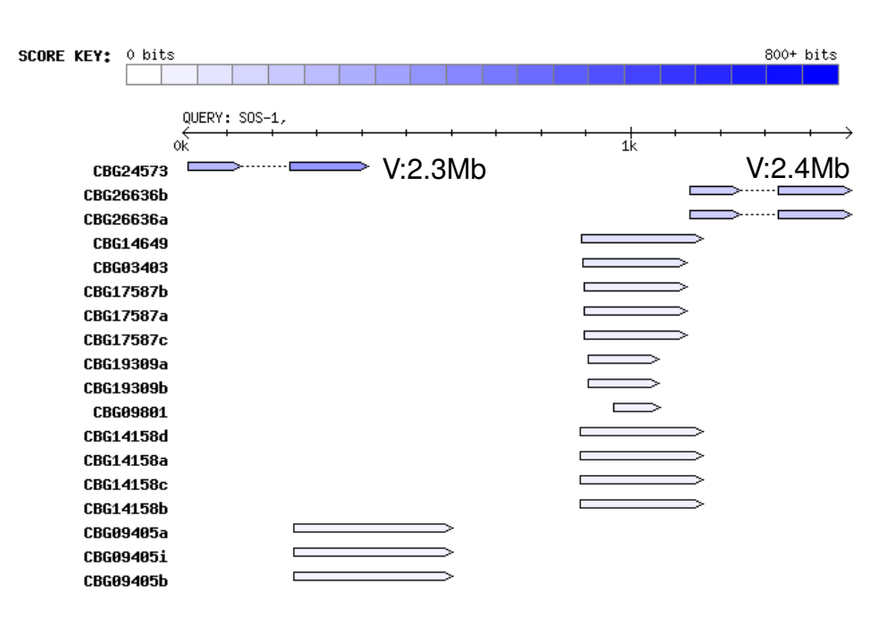


Figure S1 Putative candidates of annotation errors in *C. elegans*. 932 *C. elegans* candidates that were defined based on atypical protein domain combinations were manually inspected in the WormBase genome browser (WS177). **a** The longest isoform of the *C. elegans* gene *hmr-1* is not fully supported by transcripts that were assembled from RNA-seq data. **b** BLASTP analysis against the *C. briggsae* proteins identified two separate genes, located in a tandem configuration, that share partial matches with the longest isoform of *hmr-1*, suggesting potential annoation problems in either of the two species. **c** The gene *ceh-83* is also only partially supported by the transcriptome assembly. **d** BLASTP analysis identifies partial matches against multiple *C. briggsae* genes, which are scattered across the genome. Therefore, the gene model of *C. elegans ceh-83* could represesent an artificial gene fusion. **e,f** Incomplete support by transcriptome assemblies and BLASTP hits to separate *C. briggsae* proteins point towards potential annotation problems of *C. elegans sos-1*.

Athanasouli and Rödelsperger
BMC Genomics (2022) 23:523
https://doi.org/10.1186/s12864-022-08731-1

BMC Genomics

RESEARCH Open Access

Analysis of repeat elements in the *Pristionchus pacificus* genome reveals an ancient invasion by horizontally transferred transposons

Marina Athanasouli and Christian Rödelsperger*

Abstract

Background: Repetitive sequences and mobile elements make up considerable fractions of individual genomes. While transposition events can be detrimental for organismal fitness, repetitive sequences form an enormous reservoir for molecular innovation. In this study, we aim to add repetitive elements to the annotation of the *Pristionchus pacificus* genome and assess their impact on novel gene formation.

Results: Different computational approaches define up to 24% of the *P. pacificus* genome as repetitive sequences. While retroelements are more frequently found at the chromosome arms, DNA transposons are distributed more evenly. We found multiple DNA transposons, as well as LTR and LINE elements with abundant evidence of expression as single-exon transcripts. When testing whether transposons disproportionately contribute towards new gene formation, we found that roughly 10–20% of genes across all age classes overlap transposable elements with the strongest trend being an enrichment of low complexity regions among the oldest genes. Finally, we characterized a horizontal gene transfer of Zisupton elements into diplogastrid nematodes. These DNA transposons invaded nematodes from eukaryotic donor species and experienced a recent burst of activity in the *P. pacificus* lineage.

Conclusions: The comprehensive annotation of repetitive elements in the *P. pacificus* genome builds a resource for future functional genomic analyses as well as for more detailed investigations of molecular innovations.

Keywords: Evolution, Nematode, Comparative genomics, *Caenorhabditis elegans*, Repetitive, Zisupton

Background

Repetitive DNA describes sequence motifs repeated from hundreds to thousands of times within a genome. Repetitive DNA represents a large fraction of eukaryotic genomes, hampering genome assembly and annotation [1]. The fraction of repetitive sequences in a genome varies across species, from 12% in *Caenorhabditis elegans* to 80% in some plants [2, 3]. While their role and significance is not fully understood, the origin for the majority of these repeats has been traced to transposable elements (TEs) due to their mobility and their ability to increase their copy-number rapidly [2, 4]. The current hierarchical classification system for TEs was proposed by Wicker et al. in 2007 and takes into consideration the structural characteristics of TEs as well as their mode of replication [5]. Based on this system, TEs are classified into retrotransposons utilizing a RNA intermediate for mobilization (class I) and transposons with a DNA intermediate (class II). Class I TEs are further divided into Long Terminal Repeats (LTRs) and non-LTR sequences

*Correspondence: christian.roedelsperger@tuebingen.mpg.de

Max Planck Institute for Biology, Department for Integrative Evolutionary Biology, Max-Planck-Ring 9, 72076 Tübingen, Germany

while class II includes DNA and rolling circle (RC) elements. Initially labelled as selfish elements, TEs have been linked to metazoan genome evolution and regulation of processes associated with development and diseases [5, 6].

Until now, little is known about the impact of TEs in nematode genome evolution. In plant-parasitic nematodes, the high frequency of TEs has been associated with polyploidy and is thought to affect genome adaptation [7–9]. In *C. elegans* where 12% of the genome is estimated to be covered by TEs, experimental evidence of transposon activity is sparse with the exception of the DNA transposon superfamily Tc1/Mariner [10]. The free living nematode *Pristionchus pacificus* was introduced as a satellite model organism to *Caenorhabditis elegans* but has since been established as an independent model organism for studying phenotypic plasticity and genome evolution due to novel traits not observed in *C. elegans* [11–13]. *P. pacificus* has an established genetic toolkit and a chromosome-scale genome assembly [14]. The combination of comparative genomics and subsequent manual curation of the gene predictions produced by automated pipelines has generated a high quality gene annotation for *P. pacificus* [15, 16]. However, the current annotation does not include a dataset for repetitive sequences and specifically TEs. A comprehensive characterization of repetitive sequences in *P. pacificus* is of particular importance for us as this may complement current studies to understand the origin and evolution of new genes [17]. Previous studies have shown that new genes show substantial contributions by transposons [18]. In addition, a recent study in *P. pacificus* demonstrated that repetitive sequences can cause homology detection failures leading to misclassifications as species-specific genes [19]. Therefore, we want to test whether new genes preferentially show overlaps with such repetitive sequences.

In this study, we provide the first complete repeat dataset for *P. pacificus*. To that end, we applied different approaches to identify and annotate repetitive sequences in this nematode. The resulting datasets were evaluated and RepeatModeler2 was chosen for further analysis due to the high coverage, agreement with the other methods, as well as the TE classification it provides. Subsequently we utilized the available transcriptomic, phylostratigraphic and gene annotation data to screen for evidence of active transcription of TEs. We found multiple candidates for active transposons while simple repeats were overrepresented in protein-coding genes. Contrary to our expectation, we do not see a strong trend towards an enrichment of repetitive elements among young genes. We actually found an opposing trend with the strongest signal being an overrepresentation of simple repeats among old gene classes. Finally, we identified distant homologs of the Zisupton DNA transposon superfamily which is absent in *C.elegans* but present in fishes, fungi and other metazoans and attributed their presence in *P. pacificus* to horizontal gene transfer.

Results

There is little agreement between different repeat finders

We initially used RepeatMasker (A.F.A. Smit, R. Hubley & P. Green RepeatMasker, http://repeatmasker.org) with *C. elegans* as a reference to identify repeats in *P. pacificus*. This approach masked 3.8 Mb (2.4%) of the *P. pacificus* genome, a small portion considering the fact that *P. pacificus*' genome is larger than *C. elegans*' combined with previous knowledge regarding the amount of repetitive elements in *C. elegans* [3, 10]. The failure of RepeatMasker could be attributed to divergence between the two genomes and horizontal gene transfer which deemed *C. elegans* an insufficient reference and de novo repeat detection necessary. For this purpose, we chose 11 additional tools representing a variety of approaches, ranging from de novo identification based on machine-learning to library-based detection [20–30]. The repeat finders applied to this study are listed in Table 1. To compare methods, we split the *P. pacificus* genome in consecutive 1-kb windows. We then encoded repeat information as 1 if a 1-kb window contained repeats by a given method and 0 otherwise. Subsequently we used hierarchical clustering and analysis of most abundant patterns to compare different methods. Similar approaches clustered together as is the case for the RED/RepeatModeler2 and Dustmasker/sDust pairs (Fig. 1A). Furthermore, software tools like MiteFinderII and mReps which masked a small percentage of the genome were separated from the other tools, an indication of lower effectiveness in identifying TE in the nematode's genome. This is mirrored in the most abundant patterns of 1-kb windows (Fig. 1B). Out of the 117,134 most common 1-kb windows between the different tools, RepeatModeler2 and RED shared the majority ($N = 77{,}679$ 1-kb windows) while all of the approaches except LTRharvest, MiteFinderII and mReps shared the top 12,641 1-kb windows (Fig. 1B). LTRharvest and MiteFinderII regions were completely absent throughout the most abundant patterns (Fig. 1B). On the contrary, coverage by Tantan was present in all windows of the presence/absence heatmap which spans approximately 123 Mb. These comparisons revealed little congruence between the different datasets and we thus wanted to decide which dataset to use for further analysis.

The RepeatModeler2 dataset provides high coverage and TE annotation

To decide which repeat annotation to accept as the most representative, we considered the total genomic coverage

Table 1 The repeat finders used to detect repeat sequences incorporate a variety of approaches, ranging from library-based masking (e.g. RepeatMasker) to machine-learning using the reference genome (e.g. RED)

Software	Type of repeat	Method	References
RED	Tandem repeats, TEs	Machine-learning	[20]
LTRharvest	Long terminal repeat retrotransposons	Signature-based	[21]
Tallymer	Tandem repeats, TEs (plants)	De novo based on k-mers	[22]
MiteFinderII	Miniature inverted repeat TEs	De novo based on k-mers	[23]
TRF	Low-complexity regions, tandem repeats	De novo	[24]
Tantan	Low-complexity regions, short tandem repeats	De novo	[25]
MsDetector	Microsatellites	Learning-based	[26]
RepeatModeler2	Tandem repeats, TEs	Consensus	[27]
sDust	Low-complexity regions, tandem repeats	De novo	[28]
Dustmasker	Low-complexity regions, tandem repeats	De novo	[29]
mReps	Low-complexity regions, tandem repeats	De novo	[30]
RepeatMasker	Tandem repeats, TEs	Library-based	A.F.A. Smit, R. Hubley & P. Green RepeatMasker, http://repeatmasker.org

for each method. We elected a baseline equal to 12% of the genome due to the repetitive content of *C. elegans.* This threshold is the lower estimate regarding the span of repeats in nematodes as shown in relevant work [4, 9, 31, 32]. For a comprehensive annotation we arbitrarily decided to accept possible false positives and therefore focus on approaches which identified around 12% or even a larger fraction of the *P. pacificus* genome as repetitive. Compared to the other methods, RED masked the highest portion of the genome with 38.2 Mb of repeats identified, closely followed by RepeatModeler2 (Fig. 1C). On the contrary, MiteFinderII and mReps produced low coverage masking data. We performed pairwise comparisons for the masked regions between each of the other software tools and RepeatModeler2 to determine the level of agreement between RepeatModeler2 and the remaining methods. As expected, the RepeatModeler2 dataset incorporated almost all the genomic loci identified by RepeatMasker (using the *C. elegans* repeat library available) and the majority of the repeats identified by RED and LTRharvest (Fig. 1D). RepeatModeler2's dataset was chosen for further investigation due to the high agreement with the majority of the approaches as well as the annotation it provides, with 9.8 Mb out of the total 33.8 classified (Fig. 1E).

Retrotransposons are the most abundant TE class, accounting for 50% of annotated repeats

According to the RepeatModeler2 classification, the most abundant repetitive sequences in the *P. pacificus* genome are Long Interspersed Nuclear Elements (LINEs), with simple repeats and DNA transposons following closely (Fig. 1E, F). Penelope elements, LTRs and RC/Helitrons make up a small portion of the RepeatModeler2 dataset. It is worth noting that RepeatModeler2 could not distinguish Penelope elements from LINEs, offering instead a unified classification as Penelope/LINE. RepeatModeler2 did not annotate any of the identified repetitive sequences as a SINE, in contrast to RepeatMasker which yielded 30 SINEs with *C. elegans* as a reference. In order to improve classification, we tested DeepTE [33] and the RFSB classifier from transposonUltimate [34] against the annotated dataset of RepeatModeler2. Both classifiers differed from the RepeatModeler2 homology-based classification with LTR retrotransposons and DNA transposons as the main sources of the discrepancies (Additional file 1, Fig. S1). DNA transposons and LTRs were predominant in the DeepTE classification with 50% of DNA transposons identified by RepeatModeler2 classified as LTR and vice versa (Additional file 1, Fig. S1A). Furthermore, LINEs and Helitrons appeared underrepresented in DeepTE. RFSB reclassified less than 25% of DNA transposons as LTRs but recognized the majority of RC/Helitrons as LTRs and the rest as SINEs (Additional file 1, Fig. S1B). Inconsistent classification of repetitive elements could likely be due to substantial sequence divergence or nested insertions. In order to compare the classifications on a cleaner data set, we focused on single exon transcripts which completely overlap TEs and are therefore our best candidates for active transposons. The corresponding sequences should be less degenerated and the probability of nested insertions should be minimized. Based on a manual inspection of classifications of 200 randomly chosen sequences, we found that in 73.3% of cases, all three methods agreed on the TE class (DNA transposon/retrotransposon). In addition,

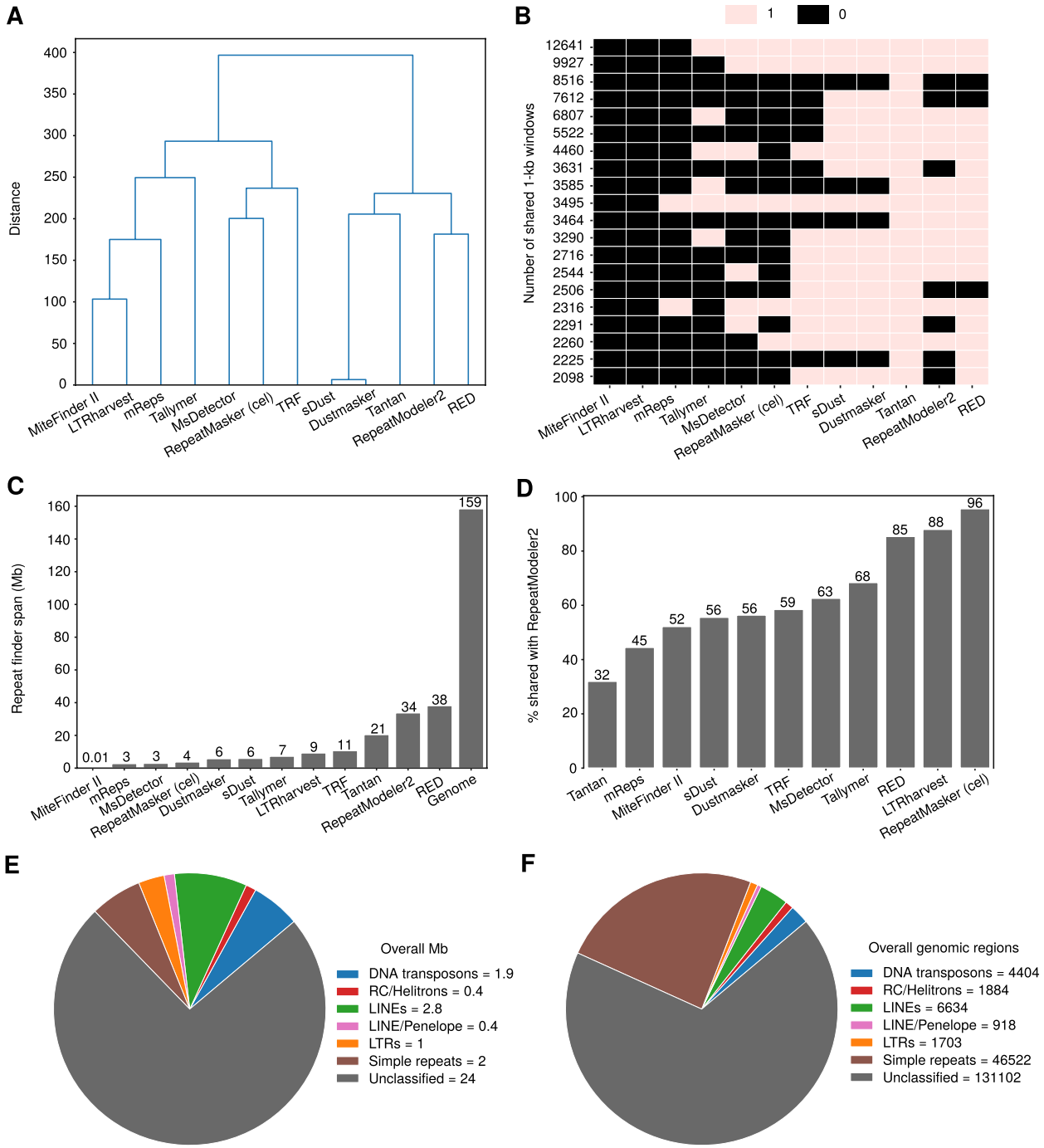


Fig. 1 **A** Hierarchical clustering of the tools used for de novo repeat detection in *P. pacificus* based on 1-kb non-overlapping windows identified repeat finders with similar performance. The y-axis reflects the Euclidean distance between the binary vectors of the 1-kb windows. **B** The 20 most abundant patterns of common 1-kb windows reflected the clustering results. **C** The sum of nucleotides each repeat finder spanned in Mb was calculated and compared to the entire genome of *P. pacificus*. RED masked the most genomic regions with RepeatModeler2 following. **D** The comparison between each of the repeat finders and RepeatModeler2 showed that the level of agreement varies, ranging from 95.6% with RepeatMasker (*C. elegans* as a template) to 32% with Tantan. **E** The pie chart shows the amount of repetitive regions identified by RepeatModeler2. Class I TEs and specifically LINEs, LTRs and Penelope elements consist the majority of annotated TEs in the *P. pacificus* genome. **F** The pie chart shows the number of identified RepeatModeler2 regions for each type. The majority of annotated masked genomic regions is assigned to simple repeats

RepeatModeler2 showed the lowest error rate when compared to the other two classification methods (Additional file 1, Fig. S2). Therefore, we decided to use the RepeatModeler2 classifications for further analysis.

The distribution of TE across chromosomes depends on the class

The general distribution of repeats across chromosomes has been analyzed previously showing lower repeat density at chromosome centers [14]. Note that that *P. pacificus* chromosome I has two center-like regions that are also defined by high gene density and low sequence diversity. To investigate the chromosomal distribution of the different TE subclasses, we calculated the fraction of coverage by DNA transposons, LINEs and LTRs per 5-kb window. In chromosome I DNA transposons were roughly evenly distributed in contrast to LINEs and LTRs which showed higher TE densities at the arms and the middle of the chromosome I (Fig. 2). LTRs and LINEs followed a similar distribution pattern in chromosomes II, III and V, with a noticeable decline in center-like regions and enrichment at the chromosome arms (Fig. 2). Enrichment towards the chromosomal arms was also observed in chromosome IV for all three types of elements. The distribution of DNA transposons, LTRs and LINEs was more even on chromosome X. Furthermore, we examined the distribution of the repeat datasets produced by RepeatModeler2, RED and Tantan. The selected datasets represent the three methods that masked the highest percentage of the genome. RepeatModeler2 and RED exhibited almost identical distribution across all chromosomes (Additional file 1, Fig. S3). In summary, the lower repeat density at the chromosome center holds true for LINEs and LTRs while DNA transposons exhibit mostly a more even distribution.

DNA transposons and LINEs show evidence of expression as single-exon transcripts

To examine the expression of TEs and simple repeats, we identified single-exon (SE) genes from an existing transcriptome assembly [35]. We initially investigated the repeats fully overlapping SE genes to gather evidence for active transcription of TEs. In total, 14% of single exon genes have their exons fully covered by TEs. Out of a total of 897 TE fully covering SE genes, 281 were LINEs, 269 DNA transposons, 151 LTRs, 118 RC/Helitrons, 49 Penelope elements and only 29 Simple repeats (Fig. 3A). To determine the best candidates for active TEs, we set a cutoff of at least 20 single-exon genes covered by a single superfamily. Among the LINEs with evidence of expression, the CR1 and RTE superfamilies were the most abundant while the most overrepresented DNA transposons superfamilies were Sola-3 and TcMar-Tc1 (Fig. 3B). The Gypsy and Pao superfamilies were the two most abundant contributors regarding LTRs with evidence of transcription (Fig. 3B). Thus CR1, Sola-3, RTE, TcMar-Tc1, Gypsy and Pao are the best candidates for active transposons in *P. pacificus*.

Protein-coding genes of all age classes exhibit contributions from TEs

Previous analysis in *P. pacificus* has shown that repeats can lead to homology detection failures, thereby contributing to the classification of coding sequences as orphan genes [19]. Furthermore, around half of primate-specific orphan genes show traces of TEs [18]. To assess the contribution of TEs and repeats in protein-coding genes, we screened for overlaps between TEs/repeats and the complete gene annotation for *P. pacificus* [16]. For this purpose, we changed the overlap threshold to a minimum of 50% exon coverage by repeats. Simple repeats accounted for 1126 out of 2569 genes overlapping repeats (Fig. 4A). On the contrary, TEs did not exhibit a high number of overlaps as was the case with SE genes (Fig. 4A). The trend for the overrepresented DNA transposon superfamilies was similar to SE genes with Sola-3 and TcMar-m44 being heavily overrepresented (Additional file 1, Fig. S4). Compared to LINE-associated exons in SE genes, the CR1 superfamily remains the predominant one.

To test whether transposons disproportionately contribute towards new gene formation, we quantified the overlap between repeat elements and protein-coding genes across different age classes (minimum 50% of an exon). These age classes were defined based on phylostratigraphic analysis of ten diplogastrid genomes that form a ladder-like phylogeny [36]. Genes were assigned to age classes based on the presence of BLASTP hits (e-value< 0.00001) in the most distantly related genome, with *P. pacificus*-specific genes being assigned to age class 0 and genes with homologs in the genome of *M. japonica* being assigned to age class 9. The majority of genes in all age classes did not show an overlap with TEs or simple repeats (Fig. 4B). We found that simple repeats were overrepresented across all age classes with the oldest age classes exhibiting the highest percentage.

We performed a Gene ontology (GO) term overrepresentation analysis for the oldest genes with repetitive sequences. Specifically, we employed the David webtool (https://david.ncifcrf.gov/summary.jsp) to test for enriched GO terms in *C. elegans* orthologs of genes with simple repeats or low complexity regions against a background set of all *C. elegans* orthologs. This identified 'nucleus' (GO:0005634, corrected $P = 1.3 \times 10^{-13}$), 'nucleic acid binding' (GO:0003676, corrected $P = 3.2 \times 10^{-8}$) and 'DNA binding' (GO:0003677, corrected $P = 1.3 \times 10^{-8}$) as the most significantly enriched

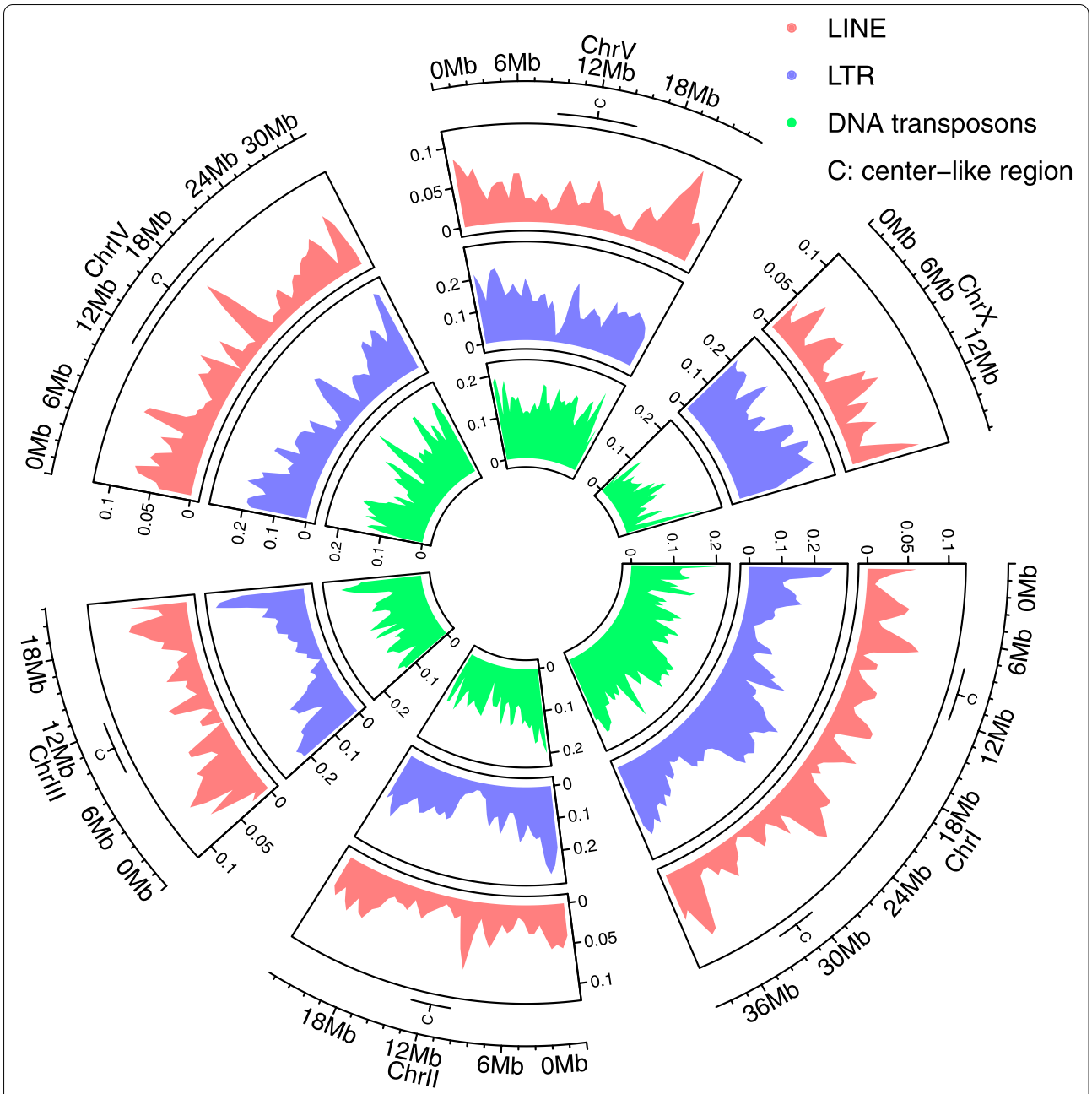


Fig. 2 Distribution of DNA transposons, LINEs and LTRs as identified by RepeatModeler2 across the chromosomes of *P. pacificus* revealed TE enrichment in the arms of the autosomes and a depletion in the chromosome centers

terms, followed by terms like 'locomotion' (GO:0040011, corrected $P=1.5\times10^{-5}$) and 'hermaphrodite genitalia development'[1] (GO:0040035, corrected $P=7.7\times10^{-5}$). One example of these genes is the ortholog of *C. elegans dpy-22*, which functions as a transcriptional coactivator [37]. Multiple simple repeats span protein-coding exons of the *P. pacificus* ortholog. Protein translations of these repetitive sequences result in a glutamine-rich C-terminal region that is also found in *C. elegans dpy-22* (Additional file 1, Fig. S5).

[1] As was pointed out by an anonymous reviewer and was also confirmed by other colleagues, hermaphroditic nematodes have female genitalia.

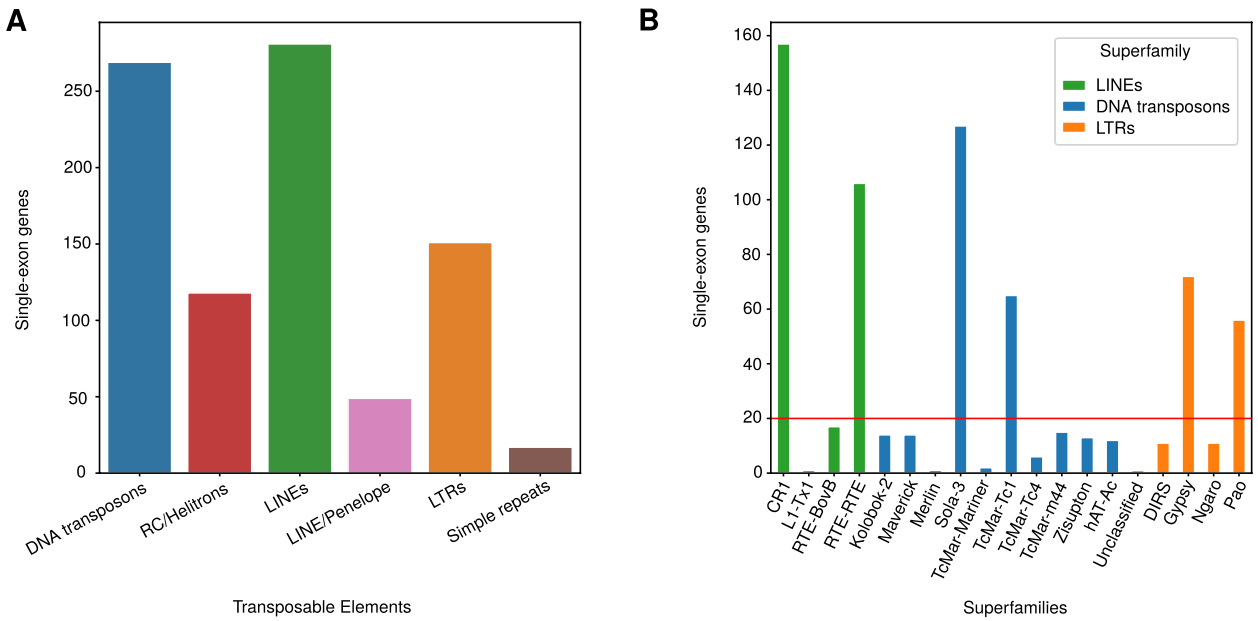


Fig. 3 Expression of TEs and simple repeats. **A** Investigation of single-exonic transcripts fully overlapping repeat sequences revealed the best candidates for active transposons. **B** Further analysis identified the CR1, Sola-3, RTE, TcMar-Tc1, Gypsy and Pao superfamilies as the most overrepresented superfamilies among DNA transposons, LINEs and LTRs with 100% exon coverage in single-exon transcripts. The threshold for determining the most abundant superfamilies was set at a coverage of at least 20 single-exon transcripts

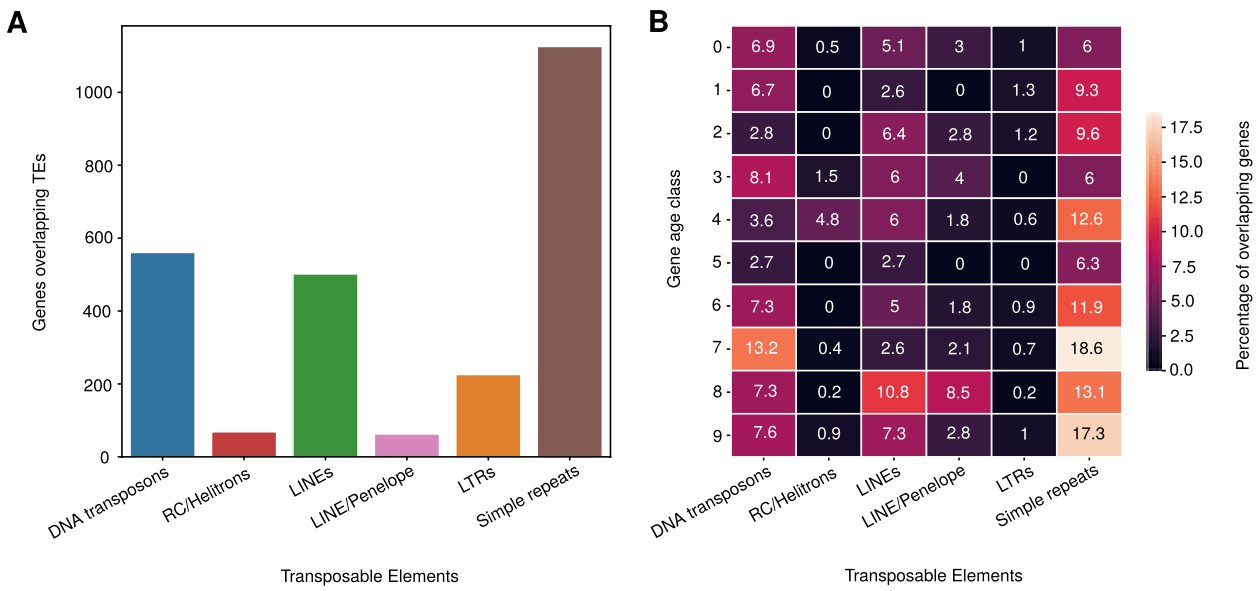


Fig. 4 **A** Coverage analysis of repeat sequences with the current gene annotations showed that protein-coding genes are enriched with simple repeats contrary to single-exon genes. **B** Tandem repeats remained the most overrepresented repeat sequences after inspection of protein-coding genes spanning the nine gene age classes (see *Methods*). The fraction of genes overlapping simple repeats gradually increased from 6% in *P. pacificus* exclusive genes to 17.3 and 18.6% of genes conserved with *M. japonica* and *P. fissidentatus*, respectively with a noticeable depletion of all repeat sequences in age class five

Horizontal gene transfer has led to an ancient invasion by DNA transposons into diplogastrid genomes

The comparison of transposon annotations with current gene models (Fig. 4A) revealed a large number of DNA transposons (Fig. 3). Fifty-eight of these sequences correspond to regions that were classified as Zisupton transposons by RepeatModeler2. Zisupton denotes a class of multi-exonic DNA transposons that were initially characterized in fishes, but are also present in fungi and algae, which suggested horizontal gene transfers [38]. In *P. pacificus*, we identified two orthologous gene families comprising more than 40 genes that overlap annotated Zisupton regions. Similar to their homologs in fishes, the corresponding proteins are up to 1400 amino acids in length. BLASTP searches against the NCBI nr database identified the best hits in green algae, fungi and other metazoans such as lancelets (*Branchiostoma floridae*), sea stars (*Patiria miniata*) and mussels (Fig. 5A). Complementary BLASTP searches against 147 nematodes (excluding diplogastrids) on WormBase ParaSite (version WBPS16) [39] identified only hits (e-value < 0.001) in the nematode *Plectus sambesii* [40]. However, phylogenetic analysis indicated that Zisupton sequences from *Pristionchus* and *Plectus sambesii* do not form a monophyletic clade. This suggests that they derived from independent horizontal gene transfers. More detailed analysis of the two orthologous families shows that one family (OG000357) has arisen only recently in the *Pristionchus* genus whereas the second family (OG00158) seems to be much older as orthologs exist in almost all *Pristionchus* species. Additional BLASTP searches could identify a homologous sequence in *Micoletzkya japonica* and *Diplogasteroides magnus* (Fig. 5A), which indicates that the initial invasion presumably occurred in the diplogastrid family. Moreover, both orthologous gene families have undergone recent expansions in the *P. pacificus* lineage (Fig. 5B). For both recently expanded orthogroups, we identified a core region of nearly perfect sequence identity. This core region spanned 8099 and 8870 nucleotides in the orthogroups OG00158 and OG000357, respectively. The corresponding protein products differed substantially in their protein length and gene structure, ranging from 1166 amino acids and 18 exons for OG00158 to 1367 amino acids and 31 exons for OG000357 (Additional file 1, Fig. S6A). In addition, alignment of the 5′ and 3′ non-coding regions showed evidence of terminal inverted repeats (Additional file 1, Fig. S6B) but we could not detect any target site duplications.

To identify more direct evidence of recent transposon activity, we screened conserved syntenic regions between the three most closely related *Pristionchus* species for a *P. pacificus*-specific insertion of a Zisupton sequence. Figure 5C shows an example of a *P. pacificus*-specific Zisupton insertion in a conserved syntenic region between *P. exspectatus* and *P. arcanus*. Thus, our analysis suggests that horizontal gene transfer has led to an ancient invasion of DNA transposon into the diplogastrid family and these transposons have undergone a recent wave of increased activity along the *P. pacificus* lineage.

There is no general trend of higher repeat content in *P. pacificus*

The previous analysis showed an increased activity of putative DNA transposons in the *P. pacificus* lineage. This lineage also represents a transition of the reproductive mode from a gonochoristic ancestor (females, males) to androdioecious species (hermaphrodites, males). One consequence of the evolution of hermaphroditism in nematodes is the ability to reproduce by selfing. Previous studies demonstrated that the degree of selfing can impact the activity of TEs [41, 42]. To test whether *P. pacificus* shows evidence for a generally increased transposon activity, we compared the repeat content in *P. pacificus* with four close relatives, *P. exspectatus*, *P. arcanus*, *P. maxplancki*, and *P. japonicus* [36]. For better comparability with the short-read assemblies of these gonochoristic species, we have included an alternative short read assembly of *P. pacificus* (version Pinocchio) in this comparison (Fig. 6). This analysis shows that *P. pacificus* has a higher repeat content than its closest relatives *P. exspectatus* and *P. arcanus* (Fig. 6A). However, the genome of *P. maxplancki* has the overall highest repeat content (Fig. 6A). Further inspection of transposon classes shows that only LINE elements (CR1 and RTE-RTE) are much more abundant in *P. pacificus* when compared to *P. exspectatus* and *P. arcanus* (Fig. 6B, C). Moreover, SINE elements which appear to be missing in *P. pacificus* are present in the other genomes (Fig. 6B). The more distantly related genome of *P. maxplancki* has much higher levels of LTRs and SINEs, when compared to *P. pacificus*. Thus, the androdioecious genome of *P. pacificus* does not generally have the highest content of repeats and TEs.

Discussion

How well do we know our genomes? Certainly, we have gained tremendous knowledge over the last twenty years after the sequencing of the first metazoan genomes. With constantly developing technologies, genome sequencing and functional genomic studies allowed us to identify disease associations, to gain evolutionary insights, and

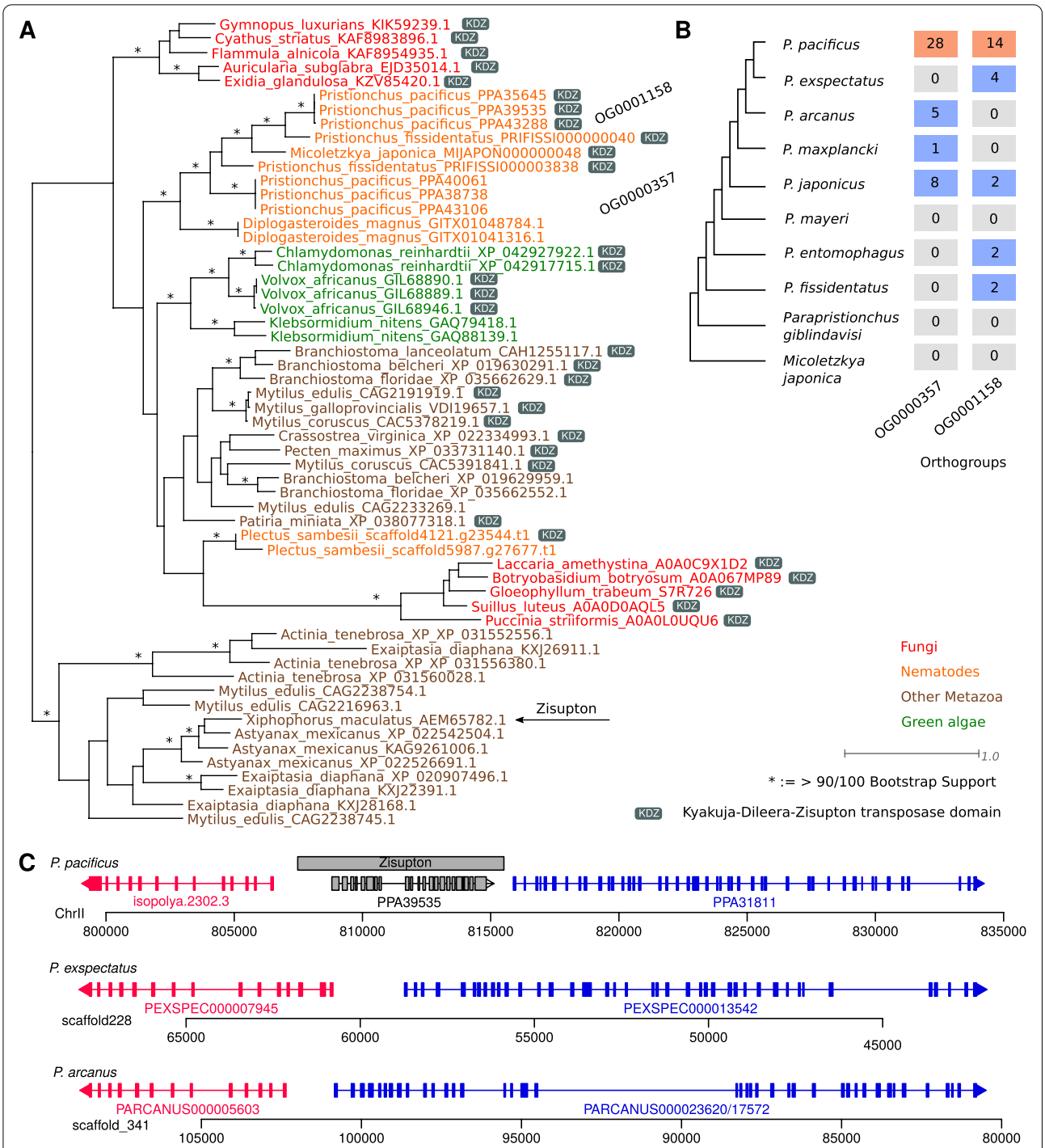


Fig. 5 **A** Homology searches of putative Zisupton transposons in *P. pacificus* identified most closely related sequences in other metazoan phyla as well as green algae and fungi. The phylogenetic tree was reconstructed from a protein alignment of the Kyakuja-Dileera-Zisupton transposase domain (PF18758). The distinct grouping of the two orthologous families in *P. pacificus* from a homolog in the nematode *Plectus sambesii* suggests independent horizontal gene transfers. **B** Orthology analysis across ten diplogastrid genomes elucidates the recent evolutionary history of Zisupton activity. The plot shows the numbers of orthologs for both families across the schematic phylogeny. While the orthologous family OG0000357 seems to have arisen only recently, both families showed rapid expansions in the *P. pacificus* lineage. **C** Syntenic analysis identified evidence of recent transposon activity in *P. pacificus*. The plot shows gene structures of syntenic regions in *P. pacificus*, *P. exspectatus*, and *P. arcanus*. Orthologous genes are indicated by identical colors. While synteny is conserved between the two outgroup species, the *P. pacificus* genome shows a specific insertion of a putative Zisupton transposon

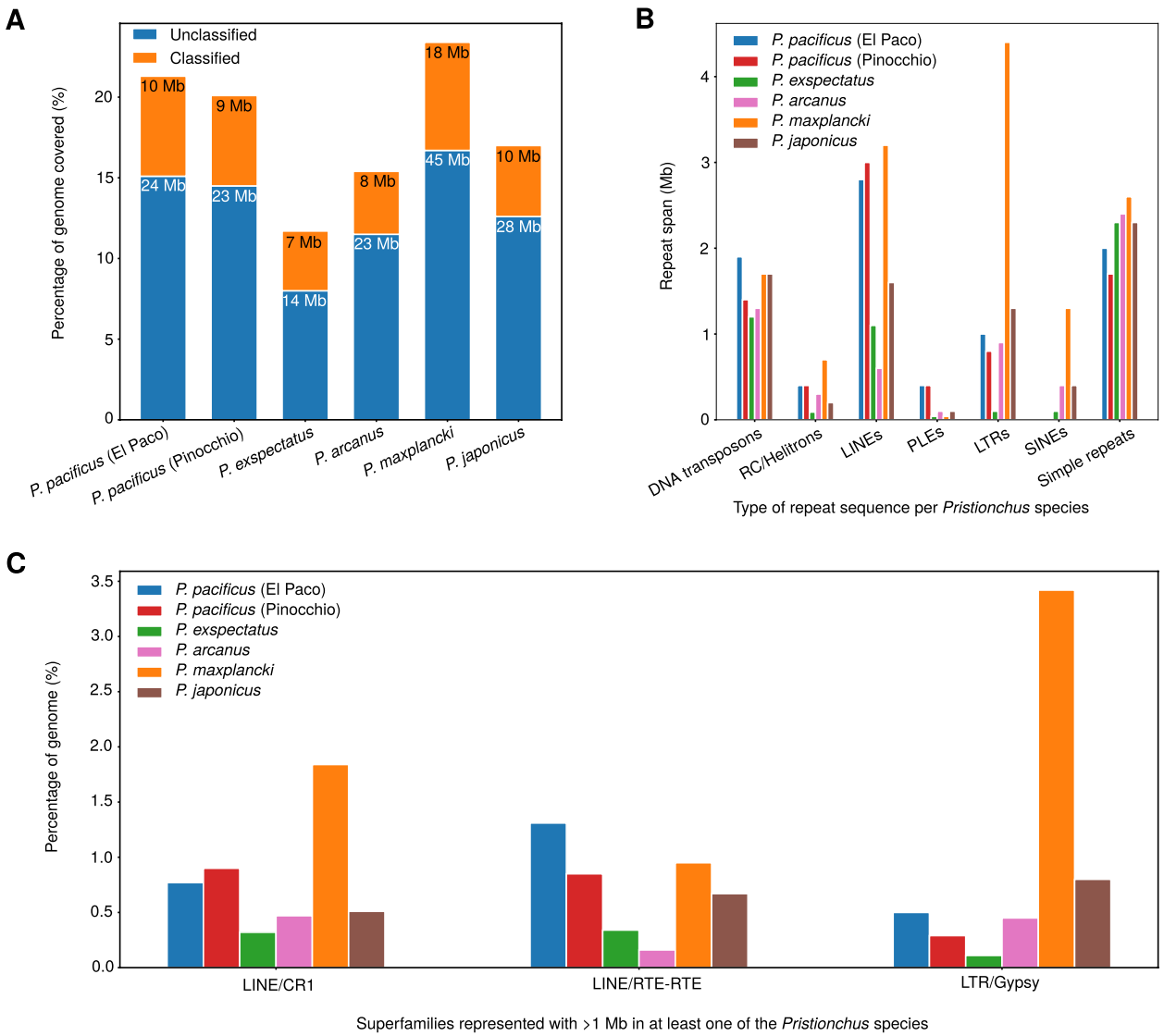


Fig. 6 **A** The bars show the percentage of the genome assemblies that could be annotated as repetitive regions by Repeatmodeler and Repeatmasker. **B** The plots show the distribution of classified repetitive elements across the *Pristionchus* genomes. **C** Superfamilies spanning at least 1 Mb in any *Pristionchus* genome are shown for different species

to characterize various mechanisms of gene regulation. However, even for an extensively studied genome such as *C. elegans*, more than 40% of genes lack functional annotation [43]. For a more exotic model organism such as *P. pacificus*, only dozens of genes have been experimentally characterized [44–46] and the inference of functional annotations based on homology is hampered by the fact that around one third of genes are classified as orphan genes without detectable homologs outside the diplogastrid family [19]. Hereby, we are ignoring the fact that largest parts of the genomes are not protein-coding. Thus, there still seems to be a long way to go before we understand how gene expression levels are regulated and which parts of the genomes are functional and which not. Ironically, with more and more sequencing data, it seems to become less clear how we define genes in the first place and what is biological function [47–49].

The primary objective of the current study was to extend our knowledge of the *P. pacificus* genome by characterizing its repetitive regions. In order to capture the full diversity of repetitive sequences ranging from low complexity regions to DNA transposons and retrotransposons, we applied multiple different computational approaches. We would argue that the large-scale

differences in their predictions are mostly due to their specific objectives for identifying different classes of repetitive elements. However, these differences together with problems in classification also suggest that comprehensive annotation of repetitive elements in divergent genomes is not straightforward. In the end, we focused on the predictions by RepeatModeler2, because it is a unified approach to identify all types of repetitive elements, it annotated a similar fraction of the *P. pacificus* genome in comparison with *C. elegans*, it is able to classify TEs, and it also showed fewest classification errors in our evaluation. We then used these annotations to screen for evidence of active transposons in available transcriptome data. Future analysis of divergent *P. pacificus* strains could be used to support that the transposon activity is not only limited to the transcriptional level but actually results in transposition events.

A second major objective of our study was to investigate the impact of repetitive sequences and TEs in the formation of novel genes. Numerous studies have shown that transposons show substantial contributions to novel genes and that individual domains can be coopted to serve new functions in the host organism [50]. While we do not see an obvious signal for an enrichment of TE derived sequences in very young genes, there is a consistent fraction of roughly 10–20% with contributions of transposon-derived sequences across all age classes. The strongest signal seems to be the large fraction of simple repeats in old gene classes. We would speculate that these low complexity regions form structural motifs such as coiled coils where the whole region is constrained to exhibit specific structural properties but little selection acts on individual amino acids [51]. Finally, we identified homologs of a class of DNA transposons that is absent in *C. elegans* and most other nematodes. These sequences likely invaded the ancestor of diplogastrid nematodes by horizontal gene transfer from another eukaryote. Horizontal gene transfer of transposons seems to be frequent and has been described in another type of transposons even in *P. pacificus* [52, 53]. The Zisupton elements are unusual in a way that they are multi-exonic which makes them superficially look like typical protein-coding genes. In *P. pacificus*, we found dozens of instances of Zisupton homologs, which seems to be a result of a recent burst of transposon activity after the switch to hermaphroditism. As these sequences are technically taxon-restricted orphan genes, they constitute another example that in addition to sequence divergence, and de novo formation, also horizontal gene transfer contributes to the emergence of novel genes [54].

Materials and methods

De novo repeat detection

We used eleven tools for de novo repeat identification in *Pristionchus pacificus* in order to represent the diversity of approaches with regard to the type of repeat they detect. The methods used for locating TE can be classified in library-based, signature-based, learning-based, homology-based, de novo and consensus while the detection of tandem repeats includes library-based, learning-based and de novo methods [20]. We utilized RepeatMasker (A.F.A. Smit, R. Hubley & P. Green RepeatMasker at http://repeatmasker.org) with *C. elegans* as a reference from the library-based methods as well as the standalone version of LTRharvest [21] with the index produced by Tallymer [22] and MiteFinderII [23] (version 1.0.006, parameters: -threshold 0.6) as signature-based programs. From the de novo detection methods available, we applied Tallymer [22] (genometools suite version 1.6.1, Suffixator: -dna -pl -tis -suf -lcp -v -parts 4, Tallymer occratio: -scan -output unique relative -minmersize 8 -maxmersize 20, parameters for -scan -mersize 19 -minocc 40 -counts -pl), RED [20] (version 2.0, parameters: -frm 2), mReps (version 2.6, parameters: -res 5 -exp 3.0), Tandem Repeat Finder [24] (version 4.09, parameters optimized for *C. elegans*: 2 5 5 80 10 402,000 -f -d -m) and TANTAN [25] (version 23, parameters: -r 0.02). We also selected MsDetector [26] (MsDetectorOptimized64, version 1.2) for locating tandem repeats and RepeatModeler2 [27] as the consensus method for both TE and TR detection (version 2.0.1, parameters: -LTRstruct). Low-complexity regions were identified using Dustmasker and sDust [28] with default parameters. We fragmented the nematode genome in consecutive 1-kb windows with the BEDTools suite [55] (option: make windows) in order to count the number of TEs and tandem repeats overlaps per 1 kb for each class. We created a binary vector based on the coverage of the 1-kb windows by each repeat finder and performed hierarchical clustering (Euclidean distance, complete-linkage). For the rest of the analysis, SINEs were excluded as RepeatModeler2 did not assign any of the identified TE to this order.

Evaluation of classification algorithms

To test whether TEs labelled as "Unknown" by RepeatModeler2 could be classified with DeepTE [33] and the RFSB classifier from transposonUltimate [34], we tested both classifiers with the labelled TE from RepeatModeler2. DeepTE was used with default parameters for metazoans (−sp M) and RFSB was run on -mode classify. Both methods were compared to RepeatModeler2. In addition, single exon transcripts from the transcriptome assembly were overlapped with annotated repeats

from RepeatModeler2. Under the assumption that such transcripts represent active transposons, classification should be easier as these sequences should be less degenerated and nested insertions should not occur. We therefore chose a small subset of 200 single exon transcripts for manual comparison of classification accuracy between RepeatModeler2, RFSB, and DeepTE. The data set comprised 45 transcripts, where we could annotate protein domains in complete or partial ORFs using the hmmsearch program HMMER package (version 3.3, e-value< 0.001) with the Pfam database (version 3.1b2) [56, 57]. The remaining transcripts were randomly chosen. If available, we used protein domain information as an additional source to classify a putative TE. This was done according to the classification scheme proposed by Wicker et al. (2007) [5]. If no protein domain information was available, classification was done based on the majority vote between all three methods.

Distribution of TEs across the *P. pacificus* genome

To investigate the distribution of the three most abundant TEs (LINEs, DNA transposons and LTRs) in *P. pacificus* we divided the genome in 5-kb continuous windows. Subsequently we calculated the coverage by each type of TE as the fraction of window length using the BEDTools coverage option. The distribution of TEs across the genome was determined with the circlize package from R (function: circus.genomicDensity()).

Phylostratigraphic analysis

Protein coding genes from *P. pacificus* (version El Paco 3) were classified into age classes (phylostrata) based on the presence of most distant homologs in the phylogenomic data set of nine other diplogastrid nematodes. Age classes 0 defined *P. pacificus* specific genes and older age classes were defined based on the presence of homologs in *P. expectatus* (Age class 1), *P. arcanus* (2), *P. maxplancki* (3), *P. japonicus* (4), *P. mayeri* (5), *P. entomophagus* (6), *P. fissidentatus* (7), *Parapristionchus giblindavisi* (8), and *Micoletzkya japonica* (9). Homologs were identified based on one-directional BLASTP searches using the *P. pacificus* proteins as queries (version 2.10.1, e-value< 0.00001).

Expression analysis

The strand-specific transcriptome assembly of the *P. pacificus* reference strain PS312 (European Nucleotide Archive: HAKN01000000, [35]) was aligned to the *P. pacificus* genome assembly (version: El Paco [14]) with the help of the exonerate est2genome program (version 2.2.0, [58]). Subsequently, the PPCAC pipeline (version: 1.0) was adjusted to select one representative transcript per 100-bp window without any restriction on exon number or protein sequence length [59]. This resulted in a set of 48,605 non-redundant sequences with evidence of active transcription. In addition, we obtained the current set of gene annotations for *P. pacificus* (version: El Paco gene annotations 3, [16]). For the expression analysis we only selected the TEs classified by RepeatModeler2. We filtered the transcriptome assembly [35] to select only genes with a single exon and used the BEDTools intersect option (parameters: -f 1, −wb) to identify the TEs fully overlapping exons. We created a non-redundant transcriptomic dataset by merging isoforms of the same gene, excluded single-exon genes and searched for TEs covering at least 50% of the exons. Furthermore, we checked the exon overlap distribution for each TE class with the latest *P. pacificus* annotation [16] (minimum 50% coverage of the exon by a TE) and the phylostratigraphic data of *P. pacificus* and nine closely related diplogastrids [36].

Comparative genomic analysis of putative Zisupton sequences

Putative Zisupton transposons in *P. pacificus* were extracted as gene models [16] that are located in regions that were annotated as Zisupton DNA transposons by RepeatModeler2. The corresponding protein sequences were searched by BLASTP (e-value < 0.001) against the NCBI nr database, against four diplogastrid genomes from Casasa et al. [60], against nine other diplogastrid genomes on http://www.pristionchus.org, and against 147 nematode genomes (excluding diplogastrids) from WormBase ParaSite (version: WBPS16, [39]). The highest sequence similarity was found in regions that corresponded to the Kyakuja-Dileera-Zisupton transposase domain (PF18758). Representative sequences from major taxonomic groups were compiled into a fasta file. This set of sequences was complemented by the original Zisupton sequence [38] and the most closely related sequences from the PF18758 in the Pfam database. A multiple sequence alignment was generated by the MUSCLE aligner (version 3.8.31, [61]). A Maximum-likelihood tree was computed using the pml, optim.pml and bootstrap.pml functions of the phangorn package in R (version 3.4.4, model="LG", optNNi = TRUE, optBf = TRUE, optInv = TRUE [62]). Orthologous clustering of ten diplogastrid genomes ([16, 36]) was done using OrthoFinder (version: 2.5.2 [63]) and conserved syntenic blocks were identified by pairwise gene order alignments between *P. pacificus* and either *P. exspectatus* or *P. arcanus* using the Cyntenator software [64].

Abbreviations

CR1: Chicken repeat 1; LINE: Long interspersed nuclear element; LTR: Long terminal repeat; RC: Rolling circle; RTE: Retrotransposable element; SE: Single exon; TE: Transposable element.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-022-08731-1.

Additional file 1: Fig. S1. TE classification comparison of DeepTE and RFSB to RepeatModeler2. **Fig. S2.** Manual inspection and comparison of classification between RepeatModeler2, DeepTE and RFSB. **Fig. S3.** Circos plots of TEs identified by RepeatModeler2, RED and Tantan by order. **Fig. S4.** Superfamilies present in the most abundant expressed TEs. **Fig. S5.** Example of simple repeats overlapping a gene conserved between *P. pacificus* and *C. elegans*. **Fig. S6.** Comparison of gene structures for representative genes of both Zisupton-related orthogroups.

Acknowledgements
We would like to thank Oliver Weichenrieder and the whole Sommer lab for helpful discussions.

Authors' contributions
Conceptualization, M.A.; Investigation, M.A. and C. R.; Data curation, M.A. and C. R.; Visualization, M.A. and C. R.; Writing original draft, M.A.; Writing – review & editing, M.A. and C.R.; Project administration, M.A. and C. R.; Supervision, C.R. The author(s) read and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. This work was funded by the Max Planck Society. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials
The *P. pacificus* genome is available at NCBI Genbank under the accession number ABKE04000000. Repeat annotations are available at http://www.pristionchus.org/download.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The author declares that he has no competing interests.

Received: 29 March 2022 Accepted: 1 July 2022
Published online: 19 July 2022

References

1. Biscotti MA, Olmo E, Heslop-Harrison JS. Repetitive DNA in eukaryotic genomes. Chromosom Res. 2015;23(3):415–20. https://doi.org/10.1007/s10577-015-9499-z.
2. Makałowski W, Gotea V, Pande A, Makałowska I. Transposable elements: classification, identification, and their use as a tool for comparative genomics. In: Anisimova M, editor. Evolutionary genomics: statistical and computational methods. New York: Springer; 2019. p. 177–207. https://doi.org/10.1007/978-1-4939-9074-0_6.
3. C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science. 1998;282(5396):2012–8. https://doi.org/10.1126/science.282.5396.2012.
4. Chénais B, Caruso A, Hiard S, Casse N. The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. Gene. 2012;509(1):7–15. https://doi.org/10.1016/j.gene.2012.07.042.
5. Wicker T, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8(12):973–82. https://doi.org/10.1038/nrg2165.
6. Ansaloni F, Scarpato M, Di Schiavi E, Gustincich S, Sanges R. Exploratory analysis of transposable elements expression in the *C. elegans* early embryo. BMC Bioinformatics. 2019;20(S9):484. https://doi.org/10.1186/s12859-019-3088-7.
7. Rodriguez F, Arkhipova IR. Transposable elements and polyploid evolution in animals. Curr Opin Genet Dev. 2018;49:115–23. https://doi.org/10.1016/j.gde.2018.04.003.
8. Rodriguez-Terrones D, Torres-Padilla M-E. Nimble and ready to mingle: transposon outbursts of early development. Trends Genet. 2018;34(10):806–20. https://doi.org/10.1016/j.tig.2018.06.006.
9. Kozlowski DKL, Hassanaly-Goulamhoussen R, Da Rocha M, Koutsovoulos GD, Bailly-Bechet M, Danchin EGJ. Movements of transposable elements contribute to the genomic plasticity and species diversification in an asexually reproducing nematode pest. Evol Appl. 2021;14(7):1844–66. https://doi.org/10.1111/eva.13246.
10. Laricchia KM, Zdraljevic S, Cook DE, Andersen EC. Natural variation in the distribution and abundance of transposable elements across the *Caenorhabditis elegans* species. Mol Biol Evol. 2017;34(9):2187–202. https://doi.org/10.1093/molbev/msx155.
11. R. J. Sommer, L. K. Carta, S. Kim, and P. W. Sternberg, "Morphological, genetic and molecular description of *Pristionchus pacificus* sp. n.(Nematoda: Neodiplogasteridae)," Fundam Appl Nematol, vol. 19, no. 6, pp. 511–5252, Jan. 1996.
12. Sommer RJ. The future of evo–devo: model systems and evolutionary theory. Nat Rev Genet. 2009;10(6):416–22. https://doi.org/10.1038/nrg2567.
13. Sieriebriennikov B, et al. A developmental switch generating phenotypic plasticity is part of a conserved multi-gene locus. Cell Rep. 2018;23(10):2835–2843.e4. https://doi.org/10.1016/j.celrep.2018.05.008.
14. Rödelsperger C, Meyer JM, Prabh N, Lanz C, Bemm F, Sommer RJ. Single-molecule sequencing reveals the chromosome-scale genomic architecture of the nematode model organism Pristionchus pacificus. Cell Rep. 2017;21(3):834–44. https://doi.org/10.1016/j.celrep.2017.09.077.
15. Rödelsperger C, et al. Crowdsourcing and the feasibility of manual gene annotation: a pilot study in the nematode *Pristionchus pacificus*. Sci Rep. 2019;9(1):18789. https://doi.org/10.1038/s41598-019-55359-5.
16. Athanasouli M, et al. Comparative genomics and community curation further improve gene annotations in the nematode *Pristionchus pacificus*. BMC Genomics. 2020;21(1):708. https://doi.org/10.1186/s12864-020-07100-0.
17. Rödelsperger C, Prabh N, Sommer RJ. New gene origin and deep taxon Phylogenomics: opportunities and challenges. Trends Genet. 2019;35(12):914–22. https://doi.org/10.1016/j.tig.2019.08.007.
18. Toll-Riera M, et al. Origin of primate orphan genes: a comparative genomics approach. Mol Biol Evol. 2008;26(3):603–12. https://doi.org/10.1093/molbev/msn281.
19. Prabh N, Rödelsperger C. *De Novo*, Divergence, and Mixed Origin Contribute to the Emergence of Orphan Genes in *Pristionchus* Nematodes. G3 Genes|Genomes|Genetics. 2019;9(7):2277–86. https://doi.org/10.1534/g3.119.400326.
20. Girgis HZ. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. BMC Bioinformatics. 2015;16(1):227. https://doi.org/10.1186/s12859-015-0654-5.
21. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics. 2008;9(1):18. https://doi.org/10.1186/1471-2105-9-18.
22. Kurtz S, Narechania A, Stein JC, Ware D. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics. 2008;9(1):517. https://doi.org/10.1186/1471-2164-9-517.
23. Hu J, Zheng Y, Shang X. MiteFinderII: a novel tool to identify miniature inverted-repeat transposable elements hidden in eukaryotic genomes. BMC Med Genet. 2018;11(S5):101. https://doi.org/10.1186/s12920-018-0418-y.
24. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27(2):573–80. https://doi.org/10.1093/nar/27.2.573.

25. Frith MC. A new repeat-masking method enables specific detection of homologous sequences. Nucleic Acids Res. 2011;39(4):e23. https://doi.org/10.1093/nar/gkq1212.
26. Girgis HZ, Sheetlin SL. MsDetector: toward a standard computational tool for DNA microsatellites detection. Nucleic Acids Res. 2013;41(1):e22–2. https://doi.org/10.1093/nar/gks881.
27. Flynn JM, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020;117(17):9451–7. https://doi.org/10.1073/pnas.1921046117.
28. Morgulis A, Gertz EM, Schäffer AA, Agarwala R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J Comput Biol. 2006;13(5):1028–40. https://doi.org/10.1089/cmb.2006.13.1028.
29. Camacho C, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10(1):421. https://doi.org/10.1186/1471-2105-10-421.
30. Kolpakov R, Bana G, Kucherov G. Mreps: efficient and flexible detection of tandem repeats in DNA. Nucleic Acids Res. 2003;31(13):3672–8. https://doi.org/10.1093/nar/gkg617.
31. Woodruff GC, Teterina AA. Degradation of the repetitive genomic landscape in a close relative of Caenorhabditis elegans. Mol Biol Evol. 2020;37(9):2549–67. https://doi.org/10.1093/molbev/msaa107.
32. Koutsovoulos GD, et al. Genome assembly and annotation of Meloidogyne enterolobii, an emerging parthenogenetic root-knot nematode. Sci Data. 2020;7(1):324. https://doi.org/10.1038/s41597-020-00666-0.
33. Yan H, Bombarely A, Li S. DeepTE: a computational method for de novo classification of transposons with convolutional neural network. Bioinformatics. 2020;36(15):4269–75. https://doi.org/10.1093/bioinformatics/btaa519.
34. Riehl K, Riccio C, Miska EA, Hemberg M. TransposonUltimate: software for transposon classification, annotation and detection. Bioinformatics. 2021. https://doi.org/10.1101/2021.04.30.442214.
35. Rödelsperger C, et al. Phylotranscriptomics of *Pristionchus* Nematodes reveals parallel gene loss in six hermaphroditic lineages. Curr Biol. 2018;28(19):3123–3127.e5. https://doi.org/10.1016/j.cub.2018.07.041.
36. Prabh N, Roeseler W, Witte H, Eberhardt G, Sommer RJ, Rödelsperger C. Deep taxon sampling reveals the evolutionary dynamics of novel gene families in *Pristionchus* nematodes. Genome Res. 2018;28(11):1664–74. https://doi.org/10.1101/gr.234971.118.
37. Moghal N, Sternberg PW. A component of the transcriptional mediator complex inhibits RAS-dependent vulval fate specification in C. elegans. Development. 2003;130(1):57–69. https://doi.org/10.1242/dev.00189.
38. Bohne A, et al. Zisupton--A Novel Superfamily of DNA Transposable Elements Recently Active in Fish. Mol Biol Evol. 2012;29(2):631–45. https://doi.org/10.1093/molbev/msr208.
39. Howe KL, Bolt BJ, Shafie M, Kersey P, Berriman M. WormBase ParaSite — a comprehensive resource for helminth genomics. Mol Biochem Parasitol. 2017;215:2–10. https://doi.org/10.1016/j.molbiopara.2016.11.005.
40. Beltran T, et al. Comparative Epigenomics reveals that RNA polymerase II pausing and chromatin domain organization control nematode piRNA biogenesis. Dev Cell. 2019;48(6):793–810.e6. https://doi.org/10.1016/j.devcel.2018.12.026.
41. Wright SI, Schoen DJ. Transposon dynamics and the breeding system. Genetica. 1999;107(1–3):139–48.
42. Boutin TS, Le Rouzic A, Capy P. How does selfing affect the dynamics of selfish transposable elements? Mob DNA. 2012;3:5. https://doi.org/10.1186/1759-8753-3-5.
43. Petersen C, Dirksen P, Schulenburg H. Why we need more ecology for genetic models such as C. elegans. Trends Genet. 2015;31(3):120–7. https://doi.org/10.1016/j.tig.2014.12.001.
44. Namdeo S, Moreno E, Rödelsperger C, Baskaran P, Witte H, Sommer RJ. Two independent sulfation processes regulate mouth-form plasticity in the nematode *Pristionchus pacificus*. Development. 2018;145(13):dev166272. https://doi.org/10.1242/dev.166272.
45. Falcke JM, et al. Linking genomic and Metabolomic natural variation uncovers nematode pheromone biosynthesis. Cell Chem Biol. 2018;25(6):787–796.e12. https://doi.org/10.1016/j.chembiol.2018.04.004.
46. Sieriebriennikov B, et al. Conserved nuclear hormone receptors controlling a novel plastic trait target fast-evolving genes expressed in a single cell. PLoS Genet. 2020;16(4):e1008687. https://doi.org/10.1371/journal.pgen.1008687.
47. Gerstein MB, et al. What is a gene, post-ENCODE? History and updated definition. Genome Res. 2007;17(6):669–81. https://doi.org/10.1101/gr.6339607.
48. Graur D, Zheng Y, Price N, Azevedo RBR, Zufall RA, Elhaik E. On the immortality of television sets: 'function' in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol. 2013;5(3):578–90. https://doi.org/10.1093/gbe/evt028.
49. Keeling DM, Garza P, Nartey CM, Carvunis A-R. The meanings of 'function' in biology and the problematic case of de novo gene emergence. eLife. 2019;8:e47014. https://doi.org/10.7554/eLife.47014.
50. Cosby RL, et al. Recurrent evolution of vertebrate transcription factors by transposase capture. Science. 2021;371(6531):eabc6405. https://doi.org/10.1126/science.abc6405.
51. Lupas AN, Bassler J. Coiled coils - a model system for the 21st century. Trends Biochem Sci. 2017;42(2):130–40. https://doi.org/10.1016/j.tibs.2016.10.007.
52. Palazzo A, Escuder E, D'Addabbo P, Lovero D, Marsano RM. A genomic survey of Tc1-mariner transposons in nematodes suggests extensive horizontal transposon transfer events. Mol Phylogenet Evol. 2021;158:107090. https://doi.org/10.1016/j.ympev.2021.107090.
53. Rödelsperger C, Sommer RJ. Computational archaeology of the Pristionchus pacificus genome reveals evidence of horizontal gene transfers from insects. BMC Evol Biol. 2011;11(1):239. https://doi.org/10.1186/1471-2148-11-239.
54. Wissler L, Gadau J, Simola DF, Helmkampf M, Bornberg-Bauer E. Mechanisms and dynamics of orphan gene emergence in insect genomes. Genome Biol Evol. 2013;5(2):439–55. https://doi.org/10.1093/gbe/evt009.
55. Quinlan AR. BEDTools: the Swiss-Army tool for genome feature analysis. Curr Protoc Bioinformatics. 2014;47(1). https://doi.org/10.1002/0471250953.bi1112s47.
56. Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol. 2011;7(10):e1002195. https://doi.org/10.1371/journal.pcbi.1002195.
57. Mistry J, et al. Pfam: the protein families database in 2021. Nucleic Acids Res. 2021;49(D1):D412–9. https://doi.org/10.1093/nar/gkaa913.
58. Slater G, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005;6(1):31. https://doi.org/10.1186/1471-2105-6-31.
59. Rödelsperger C. The community-curated *Pristionchus pacificus* genome facilitates automated gene annotation improvement in related nematodes. BMC Genomics. 2021;22(1):216. https://doi.org/10.1186/s12864-021-07529-x.
60. Casasa S, Biddle JF, Koutsovoulos GD, Ragsdale EJ. Polyphenism of a novel trait integrated rapidly evolving genes into ancestrally plastic networks. Mol Biol Evol. 2021;38(2):331–43. https://doi.org/10.1093/molbev/msaa235.
61. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7. https://doi.org/10.1093/nar/gkh340.
62. Schliep KP. Phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27(4):592–3. https://doi.org/10.1093/bioinformatics/btq706.
63. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20(1):238. https://doi.org/10.1186/s13059-019-1832-y.
64. Rödelsperger C, Dieterich C. CYNTENATOR: Progressive Gene Order Alignment of 17 Vertebrate Genomes. PLoS One. 2010;5(1):e8861. https://doi.org/10.1371/journal.pone.0008861.

Publisher's Note

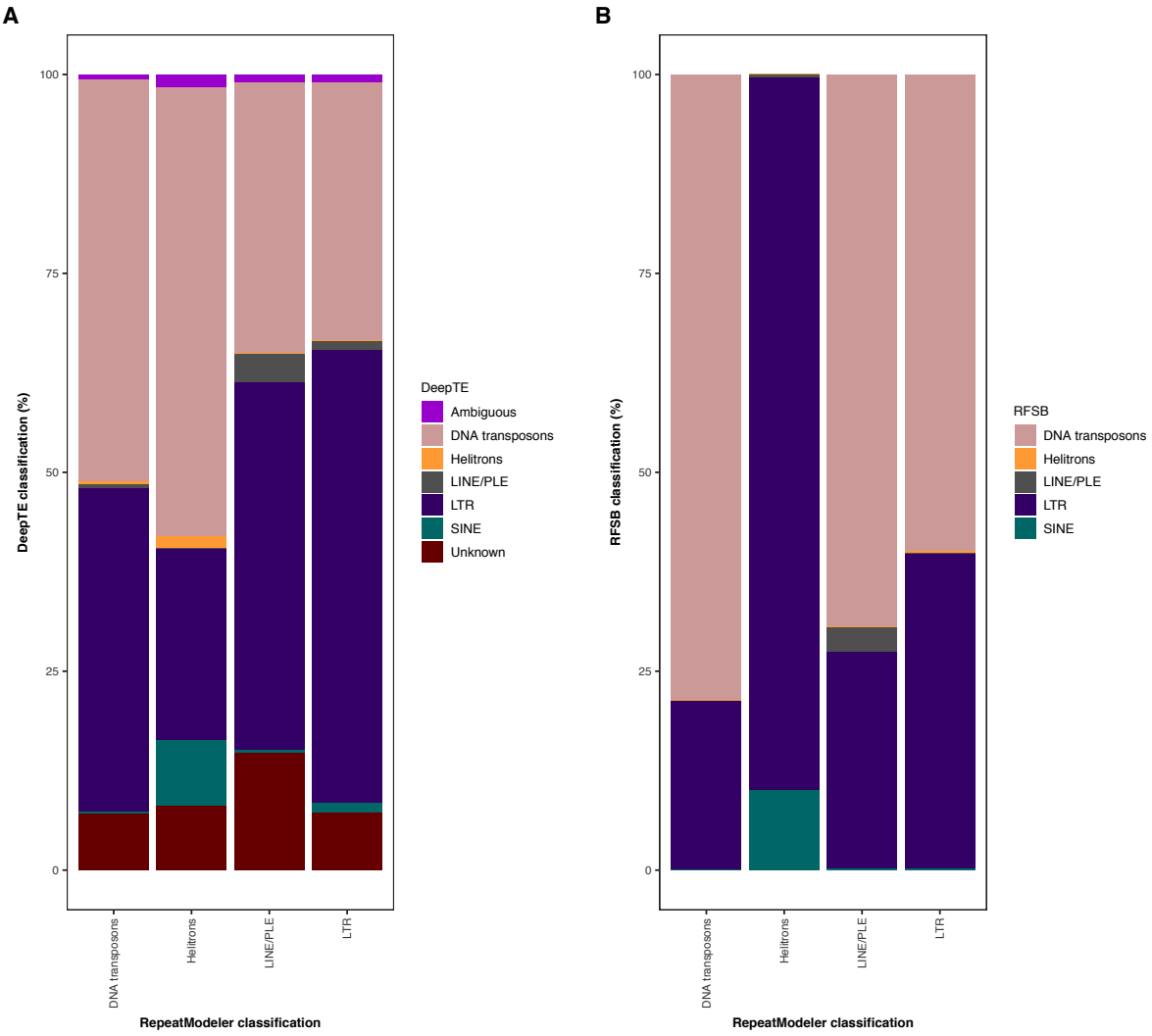


Fig. S1 (A) Reclassification of the RepeatModeler2-annotated TEs using DeepTE resulted in inconsistencies, particularly when distinguishing LTRs from DNA transposons. The same was observed with LINEs which are the most abundant TE order in the RepeatModeler2 dataset. **(B)** RFSB was more accurate when classifying DNA transposons but showed disagreement in more than 50% of LTRs. LINEs were also underrepresented in the RFSB dataset.

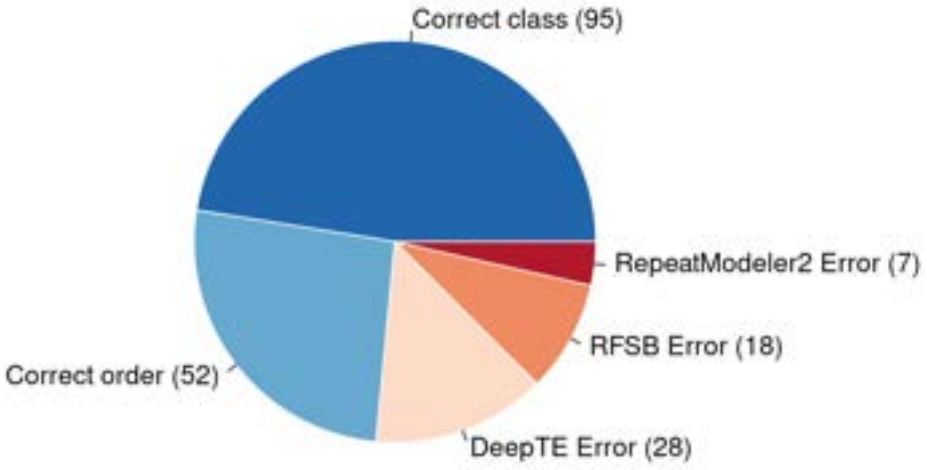


Fig. S2. The manual inspection of 200 transcribed single exon genes that have been classified as transposons shows that all three classification methods agree in 73.5% on the transposon class (DNA transposon or retrotransposon). Correct order indicates correct assignment of order, as defined by Wicker et al. 2007, e.g. LTR, or LINE. RepeatModeler2 appears to have the lowest error rate (3.5%), followed by RFSB (9.0%), and DeepTE (14.0%).

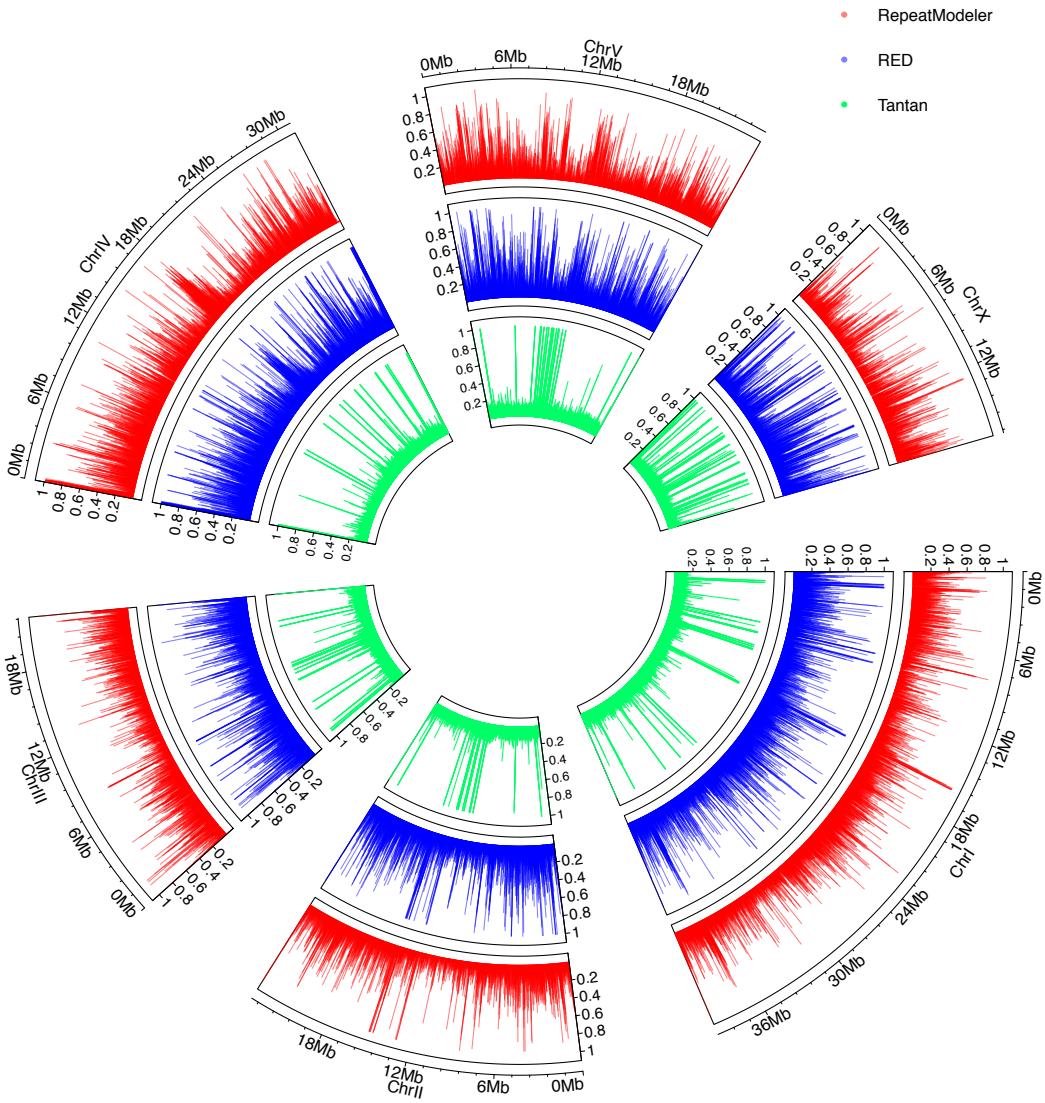


Fig. S3. The distribution of repeat sequences identified by RepeatModeler2 and RED show similar profiles across the *P. pacificus* chromosomes while TANTAN exhibited certain more TE-dense areas in the autosomes.

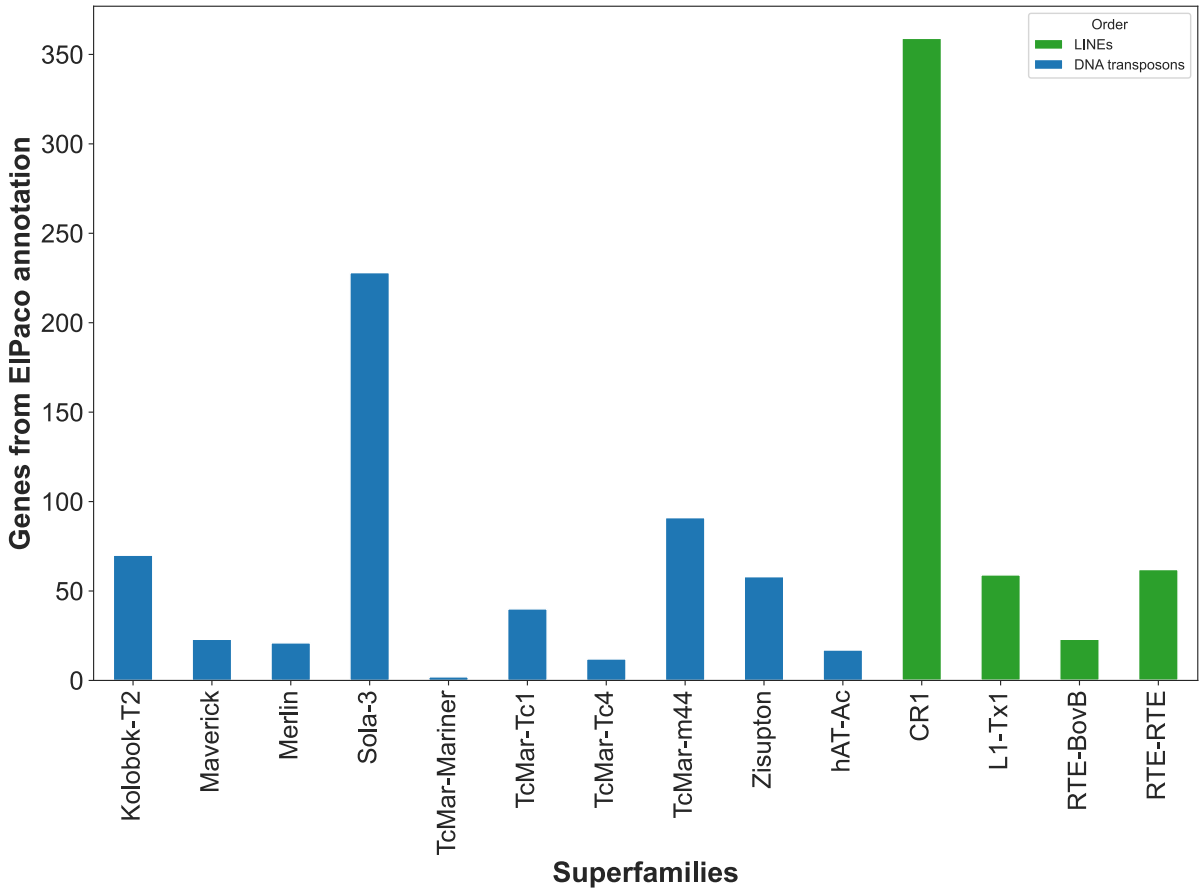


Fig. S4. CR1, Sola-3 and TcMar-m44 were the most overrepresented superfamilies among TEs spanning the exons of protein-coding genes.

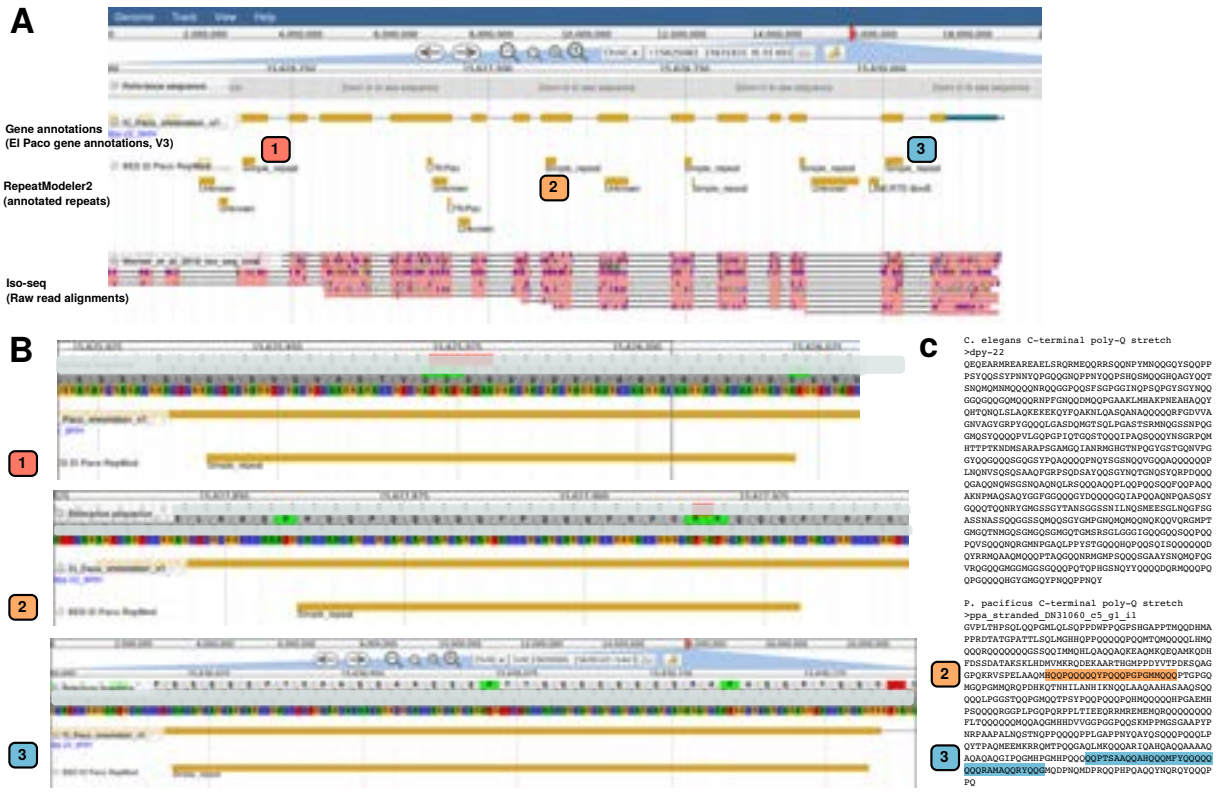


Fig. S5. **(A)** The *P. pacificus* ortholog (ppa_stranded_DN31060_c5_g1_i1) of *C. elegans dpy-22* has multiple simple repeats in protein-coding exons. The screenshots show the genomic locus of the C-terminal region of *P. pacificus dpy-22*. The gene structure is well supported by transcriptomic data (Iso-seq track). **(B)** Selected repeats in the C-terminal region are shown with the repeat sequences and their protein translations. **(C)** The poly-glutamine stretch in the C terminal region is found in *C. elegans* and *P. pacificus*.

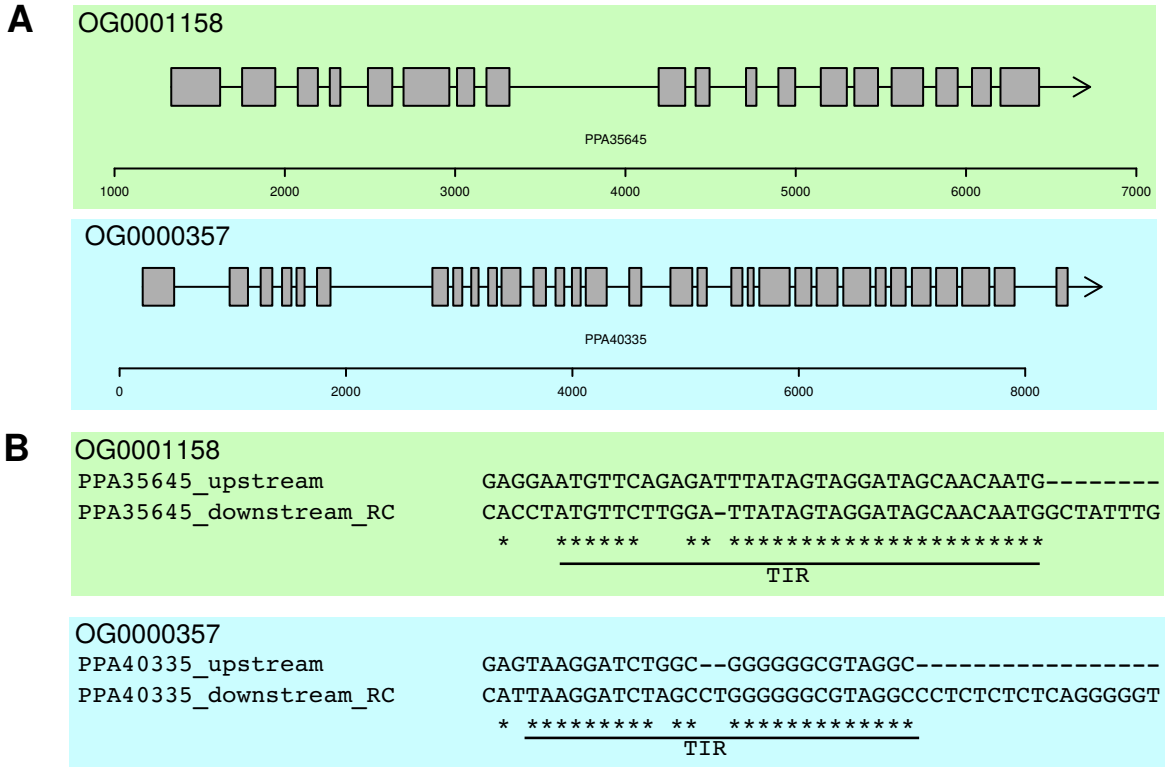


Fig. S6. **(A)** The comparison of gene structures for representative genes of both Zisupton-related orthogroups showed pronounced differences. **(B)** Alignment 5' and reverse complemented 3' non-coding regions showed evidence of terminal inverted repeats (TIRs).

A New Hope: A Hermaphroditic Nematode Enables Analysis of a Recent Whole Genome Duplication Event

Sara S. Wighard, Marina Athanasouli, Hanh Witte, Christian Rödelsperger, and Ralf J. Sommer *

Department for Integrative Evolutionary Biology, Max Planck Institute for Biology Tübingen, Max Planck Ring 9, 72076 Tübingen, Germany

*Corresponding author: E-mail: ralf.sommer@tuebingen.mpg.de.

Accepted: 22 November 2022

Abstract

Whole genome duplication (WGD) is often considered a major driver of evolution that leads to phenotypic novelties. However, the importance of WGD for evolution is still controversial because most documented WGD events occurred anciently and few experimental systems amenable to genetic analysis are available. Here, we report a recent WGD event in the hermaphroditic nematode *Allodiplogaster sudhausi* and present a comparison with a gonochoristic (male/female) sister species that did not undergo WGD. Self-fertilizing reproduction of *A. sudhausi* makes it amenable to functional analysis and an ideal system to study WGD events. We document WGD in *A. sudhausi* through karyotype analysis and whole genome sequencing, the latter of which allowed us to 1) identify functional bias in retention of protein domains and metabolic pathways, 2) show most duplicate genes are under evolutionary constraint, 3) show a link between sequence and expression divergence, and 4) characterize differentially expressed duplicates. We additionally show WGD is associated with increased body size and an abundance of repeat elements (36% of the genome), including a recent expansion of the DNA-hAT/Ac transposon family. Finally, we demonstrate the use of CRISPR/Cas9 to generate mutant knockouts, whereby two WGD-derived duplicate genes display functional redundancy in that they both need to be knocked out to generate a phenotype. Together, we present a novel experimental system that is convenient for examining and characterizing WGD-derived genes both computationally and functionally.

Key words: whole genome duplication, polyploidization, *Allodiplogaster sudhausi*, *Pristionchus pacificus*, Diplogastridae, transposable elements, ohnologs, body size.

Significance

Whole genome duplication (WGD) has been proposed as a major factor for evolution as it results in doubling of the genetic material of an organism. However, its role in evolution is still controversial as all documented cases have occurred in ancient history. Also, no study systems are available for experimental manipulation of WGD in animals. Here, we report that the hermaphroditic nematode *Allodiplogaster sudhausi* has recently undergone WGD. We document WGD by karyotype analysis and whole genome sequencing, which allowed studying several associated features. Finally, we establish CRISPR-mediated gene knockout, which allows functional manipulation in this organism, a useful tool for investigating the consequence of WGD.

Introduction

Whole genome duplication (WGD), also known as polyploidization, is when the full genome, including the chromosomes and regulatory elements, is doubled. The important role of duplication events was hypothesized as far back as the 1930s (Haldane 1932; Bridges 1936), but it is Susumu Ohno's seminal work (Ohno 1970) that popularized the notion of duplication, and WGD in particular,

driving evolution and novelty. WGD-derived duplicate genes are referred to as "ohnologs" in reference to his work. Duplication is believed to be a driver of evolution as it produces additional redundant genetic material that can potentially diverge and gain a new function. WGD is considered a greater evolutionary driver than local gene duplications as the genes are not constrained by dosage compensation (Birchler et al. 2005) and indeed, retention of duplicates derived from WGD is greater than those derived from local gene duplication events (Blomme et al. 2006).

In principle, WGD can come about from one of two ways: 1) Autopolyploidization, where multiple chromosome sets derive from a single taxon—usually due to an error in meiosis or 2) Allopolyploidization where two closely related species hybridize to form a new one. However, it is usually difficult to tease apart the exact origin (Parisod et al. 2010). After duplication, ohnologs are initially redundant with the same expression pattern and role, but after some time the fates of ohnologs tend to diverge. The degeneration or silencing of one ohnolog (nonfunctionalization) is the most common fate, while the addition of a novel function (neofunctionalization)—which could drive evolution—is the rarest (Lynch and Conery 2000; Maere et al. 2005).

It is thought that ancient WGD occurred in most eukaryotic lineages (Wolfe 2015). Indeed, nearly all documented cases of WGD have occurred ancestrally and often the exact timing of the WGD event is poorly understood. For example, WGD is characteristic of most land plants, but its timing remains elusive so that the role of WGD for morphological and functional diversity of land plants is constrained (Clark and Donoghue 2018). At least two rounds of ancient WGD took place in vertebrates (Dehal and Boore 2005), while additional WGD events are also reported in teleost fish (Amores et al. 1998) and *Xenopus laevis* (Session et al. 2016). Aside from multicellular organisms, WGD events have been well-documented in unicellular eukaryotes such as protozoans and yeast (Aury et al. 2006; Marcet-Houben and Gabaldón 2015).

WGD is associated with a number of novelties, including increases in body size (Walsh and Zhang 1992; Otto and Whitton 2000) and an expansion of transposable elements (TEs) after WGD has occurred (Marburger et al. 2018). However, despite the advances in genomics since the effects of WGD were first hypothesized, the importance of WGD in driving evolution is still controversial, with some believing it leads to an evolutionary dead-end. Evaluating the impact of WGD events on genes and evolution is difficult for a number of reasons. For instance, the most reliable indication of WGD driving evolution is the identification of neofunctionalized genes, which have been recorded in some fishes (Zakon et al. 2006; Moriyama et al. 2016). However, neofunctionalization is hard to identify as, after much divergence, the gene may no longer be similar enough to be identified as ohnologs derived from WGD. Thus, the rarity and antiquity of recorded WGD events have prevented full documentation of their impact. Additionally, it is difficult to evaluate WGD in organisms with large complex genomes and limited methods of experimental manipulation.

Nematodes are a useful group of organisms for characterizing genome biology. With their small genome sizes, easy maintenance (in the case of free-living nematodes) and available genetic tools, nematodes are an ideal system to characterize evolutionary and genetic processes. Additionally, there have been multiple evolutionary transitions toward self-fertilizing hermaphroditism, which have created isogenic study systems. Many of these hermaphroditic organisms also produce functional males, which enables genetic crosses (Avise 2011). *Caenorhabditis elegans* is already a well-established model system, while *Pristionchus pacificus* has recently been developed as a model for evolutionary developmental biology (evo-devo) and the evolution of novel traits (Sommer 2015; Schroeder 2021).

Here, we present the nematode *Allodiplogaster sudhausi*, a hermaphrodite in the same family (Diplogastridae) as *P. pacificus* that displays phenotypic plasticity in the form of a mouth-form polyphenism (Fürst von Lieven 2008; Kanzaki et al. 2014; Susoy et al. 2015). It diverged early within its family, is one of the few hermaphroditic species outside of the genus *Pristionchus*, produces functional males enabling genetic crosses, and it has a sister species, the gonochoristic (male/female) *Allodiplogaster seani* (Kanzaki et al. 2015), as another more closely related point of comparison (fig. 1*A*). *Allodiplogaster sudhausi* is strikingly large compared to its relatives (fig. 1*B*), with hermaphrodites and males having body lengths one and a half times those measured in *C. elegans* (Wood 1988) and *P. pacificus* (Sommer et al. 1996) (fig. 1*C* and *D*), and is also much larger than males and females in its sister species *A. seani* (fig. 1*E*).

We show through karyotype analysis and whole-genome sequencing that a WGD event occurred in the lineage leading to *A. sudhausi*, which is absent in *A. seani*. Subsequent analysis uncovered a number of findings. First, we characterized the retained ohnologs and those that likely underwent nonfunctionalization to identify protein domains and metabolic pathways that are more likely to be retained after WGD. Second, we showed a link between sequence and expression divergence. Third, we identified a vast abundance of repeat elements, including the very recent expansion of the DNA transposon hAT-Ac family. Lastly, we demonstrated the use of CRISPR/Cas9 in *A. sudhausi* for the first time. Specifically, we generated mutant knockouts of a common nematode marker gene and show that both ohnologs need to be knocked out to

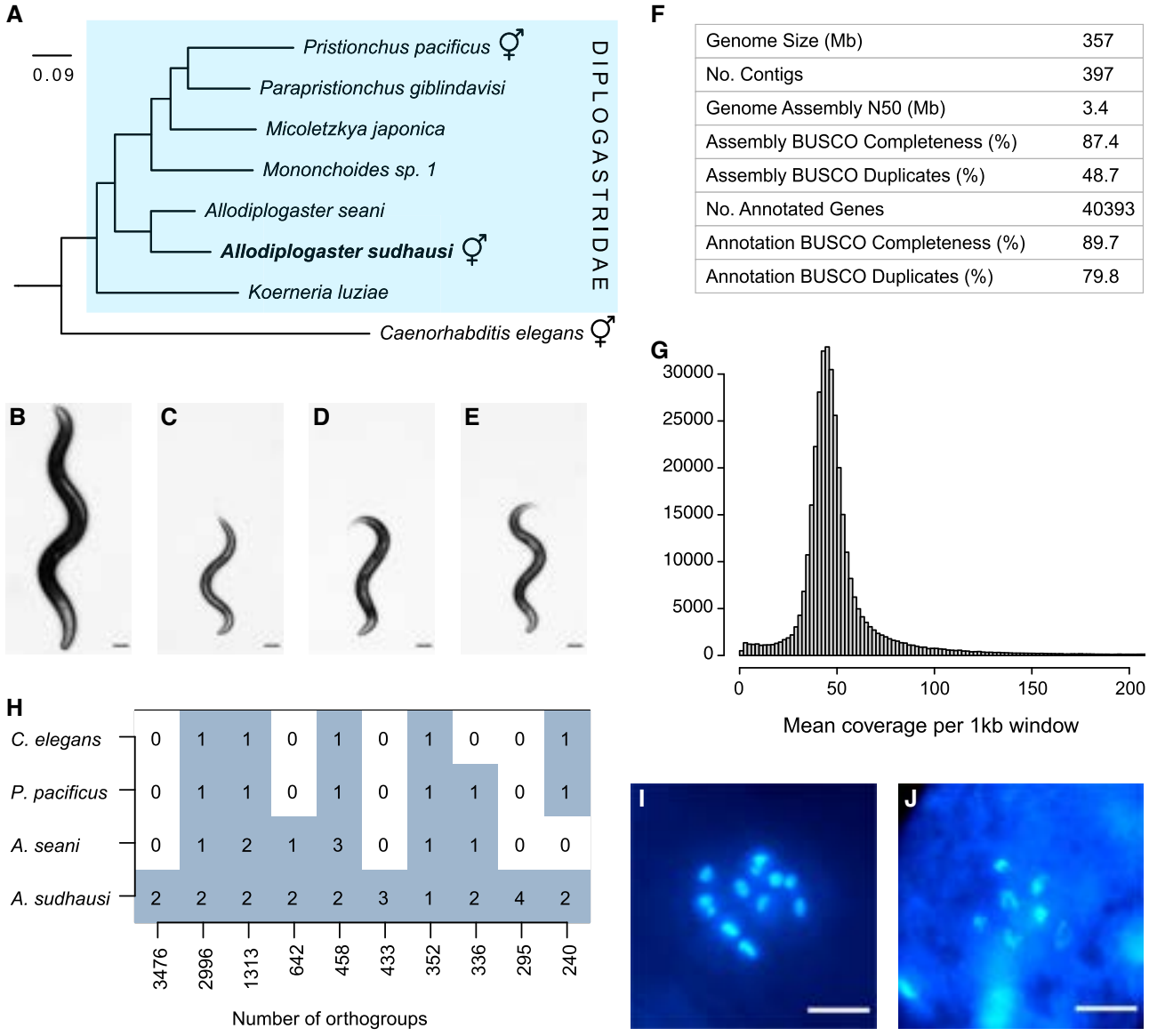


Fig. 1—Whole genome duplication in *A. sudhausi*. (*A*) Phylogenetic relationship of selected Diplogastridae, with *C. elegans* as an outgroup. Hermaphroditic (androdioecious) species are indicated by ♂. The hermaphroditic *A. sudhausi* is an early diverging lineage. The phylogeny represents a subtree from Susoy et al. (2015). (B—E) Young adult hermaphrodites and female (*A. seani*): (*B*) *A. sudhausi*, (*C*) *C. elegans*, (*D*) *P. pacificus*, (*E*) *A. seani*. Scale bar: 100 μM. (*F*) Evaluation of the *A. sudhausi* genome assembly and the resulting size, contig number, N50 as well as the number of predicted gene annotations. BUSCO analysis of the genome assembly and annotations is shown, with duplication rates very high. (*G*) Coverage analysis of the *A. sudhausi* genome (see Methods) shows a single peak, suggesting the high duplication values are not due to allelism. (*H*) Orthology clustering using OrthoFinder based on *A. sudhausi*, *P. pacificus*, and *C. elegans* annotations and the *A. seani* transcriptome. The heatmap shows the 10 most abundant orthogroups that include *A. sudhausi* orthologs. *Allodiplogaster sudhausi* has two copies in most orthogroups, with many orthogroups specific to *A. sudhausi* alone. (*I*) Hoechst staining of the *A. sudhausi* hermaphrodite oocyte shows 12 chromosomes, double the number previously reported. Due to the chromosomes being in different planes, a maximal intensity measurement was used to show all of them at once. Two of the chromosomes are stacked on top of another (shared z axis). Scale bar: 10 μM. (*J*) Hoechst staining of the *A. seani* female oocyte shows seven chromosomes.

generate a phenotype; that is, they display genetic redundancy. Overall, we present a novel nematode system with a relatively small manageable genome that can be functionally evaluated using CRISPR technology. These features allow us to examine a recent WGD duplication and DNA transposon expansion.

Results

The *A. sudhausi* Genome Contains an Exceptionally High Number of Duplicated Genes

We sequenced an inbred line of *A. sudhausi* (SB413B) using PacBio long-read sequencing, resulting in 2 × 11 Gb of

sequencing data from two SMRT cells. We assembled the genome de novo with the Canu assembler (Koren et al. 2017) and obtained approximately 60× coverage. We estimated a genome size of 357 Mb, more than double the 159 Mb genome of *P. pacificus* (Rödelsperger et al. 2017) and larger than the reported genome sizes of other Diplogastridae, which range from 143 to 297 Mb (Prabh et al. 2018). We obtained a BUSCO (Simão et al. 2015) completeness value of 87.4%, in line with previous de novo diplogastrid assemblies (Prabh et al. 2018). However, we observed an extremely high duplication rate of 48.7% (fig. 1*F*), which dwarfs the 1.3% duplicate estimate in *P. pacificus* (Rödelsperger et al. 2019). A high duplication rate is also seen in the Illumina sequencing data generated by Sieriebriennikov et al. (2018) (supplementary Table S1, Supplementary Material online). High duplication values can sometimes be due to heterozygosity; however, this is unlikely in hermaphrodites (Barriere et al. 2009) and the SB413B strain had been extensively inbred to limit such effects. In addition, we examined the coverage of the contigs and found only a single peak (fig. 1*G*), indicating the regions are not allelic.

Next, we annotated the *A. sudhausi* genome and obtained 40,393 evidence-based gene models that are either supported by transcriptomic data or *P. pacificus* protein homology (Rödelsperger 2021). When we applied the OrthoFinder software to group orthologs in *A. sudhausi*, *A. seani*, *P. pacificus*, and *C. elegans* based on sequence similarity, we found that *A. sudhausi* had two gene copies in the majority of orthologous clusters (7 out of 10) (fig. 1*H*). Altogether, these findings provide evidence for a large-scale duplication event in *A. sudhausi*, which is in agreement with previously reported gene duplication events (Sieriebriennikov et al. 2018; Biddle and Ragsdale 2020). We, therefore, wanted to determine if WGD had taken place.

Karyotype Analysis Confirms a Whole Genome Duplication Event in *A. sudhausi*

A WGD (polyploidization) event leads to an instant doubling of the chromosomal number, meaning a higher chromosome number in *A. sudhausi* would reliably indicate that WGD had occurred. Previous work suggested that *A. sudhausi* has six chromosomes (Fürst von Lieven 2008), the same as in *C. elegans* (Wood 1988) and *P. pacificus* (Sommer et al. 1996). However, since our analyses suggested a potential WGD event, we repeated karyotype analysis in *A. sudhausi* by staining the gonads of hermaphrodites using Hoechst 33342 dye. Strikingly, we counted 12 chromosomes in the hermaphrodite oocytes during diakinesis (fig. 1*I*), double the amount seen in *C. elegans* and *P. pacificus*. A recent catalog of chromosome numbers in nematodes revealed that the majority of investigated species have six or seven chromosomes, with some species exhibiting even smaller numbers after chromosome fusions (Gonzalez de la Rosa et al. 2021; Carlton et al. 2022). Thus, the observation of 12 chromosomes in *A. sudhausi* shows that WGD has taken place in the lineage leading to this species.

The Whole Genome Duplication is Specific to *A. sudhausi*

To ascertain when in evolution the WGD occurred, we examined the sister species of *A. sudhausi* that was recently described as *A. seani* (Kanzaki et al. 2015) (fig. 1*A* and *E*). Staining of female oocytes showed there are seven chromosomes in *A. seani* (fig. 1*J*). This chromosome count is similar to the numbers observed in many nematode species (Carlton et al. 2022), thereby suggesting that the WGD event occurred after the divergence of these two *Allodiplogaster* species. Ortholog clustering analyses further support the notion that the WGD occurred after the split of *A. sudhausi* and *A. seani*. Specifically, orthology clustering revealed that for 7 out of 10 orthogroups *A. sudhausi* has two-gene clusters (fig. 1*H*). The two biggest clusters, which have 3,476 and 2,996 orthogroups, are two-gene clusters for *A. sudhausi*, in which *A. seani* has zero and one corresponding orthologs, respectively. Finally, the BUSCO duplication values are also lower in *A. seani* than in *A. sudhausi* (supplementary Table S1, Supplementary Material online). Overall, these results indicate a WGD that is specific to *A. sudhausi*. This finding is compelling as it presents us with a hermaphroditic organism in which to characterize various processes related to WGD, such as the fates of WGD-derived duplicate genes (henceforth referred to as ohnologs). Note that there is unfortunately only one isolate of *A. sudhausi* available to date. Thus, it remains unknown if the WGD is fixed in this species, and if other, more closely related species (if they exist) would share the WGD event.

Comparative Analysis Shows Expansions of GPCR and Ribosomal Domains in *P. pacificus* Relative to *A. sudhausi*

We next examined the predicted Pfam protein domains in *A. sudhausi*, *A. seani*, *P. pacificus*, and *C. elegans* (the latter three have no indication of a recent WGD). The frequency of unique domain predictions for each gene was compared between species. We found an increase in the protein binding domains ankyrin (ANK) and Broad-Complex, Tramtrack and Bric-a-brac (BTB) in *A. seani* compared to *A. sudhausi* (fig. 2*A*). BTB domain-containing proteins are adaptors involved in protein degradation, which show signatures of positive selection in *C. elegans* (Thomas 2006). They have been found to be overrepresented of gonochorists in both *Caenorrhabditis* and *Pristionchus* nematodes due to

gene loss in hermaphroditic species (Rödelsperger et al. 2018; Yin et al. 2018). Thus, the overrepresentation in gonochoristic *A. seani* compared to *A. sudhausi* is consistent with previous results.

Surprisingly, we found domains in seven-transmembrane G-protein-coupled receptor class (7TM GPCRs) are disproportionately under-represented in *A. sudhausi* relative to *P. pacificus* (fig. 2*B*). A comparison with *C. elegans* again shows that 7TM GPCRs are also highly under-represented in *A. sudhausi* (supplementary fig. S1, Supplementary Material online). 7TM GPCRs mediate chemoreception in *C. elegans* where they play important roles in sense and stimuli (Troemel et al. 1997). Due to the abundance of these domains in *C. elegans* and *P. pacificus*, which are separated by large evolutionary distances (fig. 1*A*), we assumed they would be abundant in most free-living nematodes. However, a comparison between the number of domains with 7TM GPCRs across many nematodes shows 7TM GPCRs are only highly overrepresented in *C. elegans* and *Pristionchus* species (supplementary fig. S2, Supplementary Material online). Thus, the high abundance of 7TM GPCRs is not evolutionarily conserved across all free-living nematodes, and rather evolved independently in *Caenorhabditis* and *Pristionchus* species.

Next, we predicted KEGG (Kanehisa and Goto 2000) metabolic pathways and compared the abundance between species. There appears to have been an expansion of genes involved in the Mitogen-activated protein kinase (MAPK) pathway (04010) in *A. seani* (supplementary fig. S3, Supplementary Material online). MAPK is involved in cellular processes such as proliferation, differentiation, and development (Seger and Krebs 1995), as well as oxidative stress response in *C. elegans* (Inoue et al. 2005). Interestingly, we found an expansion of ribosomes (03010) in *P. pacificus* relative to *A. sudhausi* (fig. 2*C*). A greater abundance of ribosomal pathways in *P. pacificus* relative to *C. elegans* had previously been observed (Dieterich et al. 2008). By additionally examining the Pfam domain predictions across many Diplogastridae, we saw that ribosomal domains are highly over-represented in *P. pacificus* relative to others, including fellow *Pristionchus* species (supplementary fig. S4, Supplementary Material online). This finding suggests a recent expansion of ribosomal pathways and domains may have taken place in *P. pacificus*.

Ohnologs With Homeodomains and Those Involved in the mTOR Signaling Pathway are More Likely to be Retained

The young nature of the WGD event allowed us to examine potential functional biases between ohnologs pairs that had been lost or retained in *A. sudhausi*. We identified ohnologs that were retained and those that lost their duplicate based on the orthology clustering analysis described above (fig. 1*H*). For the purposes of this analysis, we deem orthogroups with two copies ohnolog pairs, and orthogroups with one single-copy genes. We hypothesize that these single-copy genes denote orthogroups that lost their ohnologous duplicate; therefore, a case of nonfunctionalization (the loss of one duplicate). We then predicted their respective Pfam domains and determined which domain-containing genes were more likely to either be retained as ohnologs or lose their duplicate. We found only five domains that were more likely to be retained (Fisher's exact test, FDR adjusted $P < 0.05$) (fig. 2*D* and *E* and supplementary Table S2, Supplementary Material online), including zf-C2H2 (PF00096) and Homeodomain (PF00046), potentially suggesting these domains are more dosage sensitive or haploinsufficient. The retention of homeodomains is noteworthy as they are transcribed by *Homeobox* genes, which have remained after WGD in ray-finned fishes and play important roles (Amores et al. 1998; Blomme et al. 2006). Notably, only one homeodomain-containing gene lost its duplicate in *A. sudhausi* (fig. 2*E*). We took this gene, which also has a predicted LIM domain, and identified its *C. elegans* ortholog using the orthology clustering data (fig. 1*H*). The *C. elegans* ortholog is *lim-4*, a LIM homeobox gene that has been found to specify olfactory neurons (Sagasti et al. 1999). This finding might suggest that this process may not be as conserved in *A. sudhausi*, reflecting the previous results that showed surprisingly few genes with GPCR domains in *A. sudhausi* (supplementary fig. S1, Supplementary Material online). In contrast to the small number of significantly retained domains, we found 82 domains that were significantly more likely to lose their ohnologous duplicate (fig. 2*D* and *E* and supplementary Table S3, Supplementary Material online), with an abundance of domains that act as enzymes, specifically ATPases, and those involved in transportation and microtubule binding. We hypothesize there is low dosage sensitivity in genes that contain these domains.

We repeated this analysis using KEGG pathway predictions and found five pathways significantly more likely to be retained as ohnologs (Fisher's exact test, FDR adjusted $P < 0.05$), including endocytosis (04144) and the mTOR signaling pathway (04150) (fig. 2*F* and *G* and supplementary Table S4, Supplementary Material online). The mTOR pathway is a regulator of cell growth and proliferation that is associated with cancers (Sarbassov et al. 2005). There were seven pathways significantly more likely to lose their ohnologous duplicate, with most involved in metabolism, including the oxidative phosphorylation pathway (00190) which is involved in ATP synthesis (Wilson 2017) (fig. 2*F* and *G*, supplementary Table S5, Supplementary Material online). Interestingly, ohnologs more likely to be retained consisted of many pathways with regulatory roles, and previous

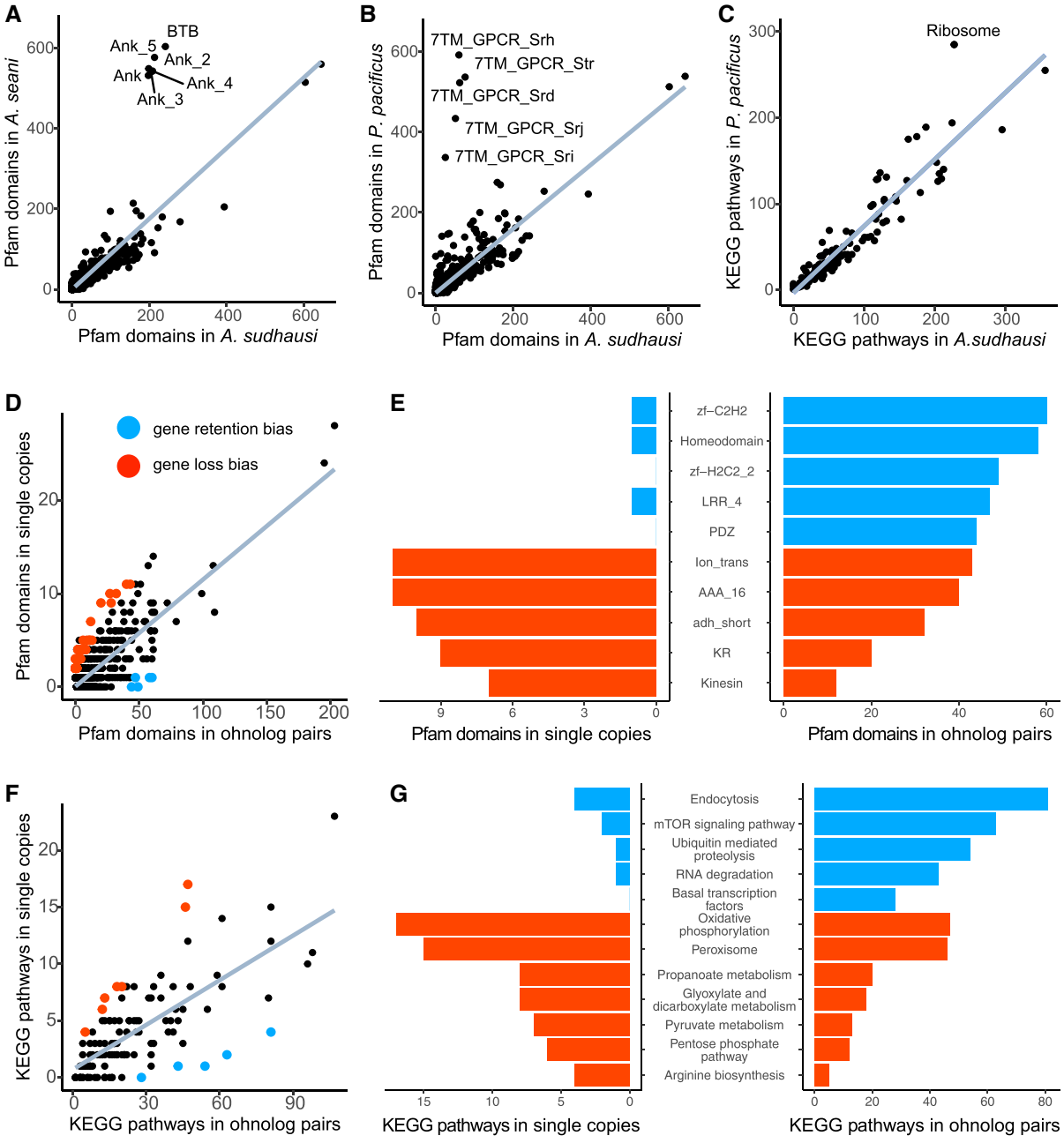


Fig. 2—Functional bias in ohnolog loss and retention. (*A*) A comparison of Pfam protein domain predictions in *A. sudhausi* and *A. seani* indicate a cluster of protein binding domains, ANKs and BTB, are underrepresented in *A. sudhausi* after the species diverged. A linear regression trendline is shown. (*B*) A comparison of Pfam protein domain predictions shows an abundance of 7TM_GPCR chemoreceptors in *P. pacificus* compared to *A. sudhausi*. (*C*) A comparison of KEGG metabolic pathway predictions in *A. sudhausi* and *P. pacificus* indicate there has been a vast increase in ribosomes (pathway 03010) in *P. pacificus*. (*D*) The plot shows the frequency of predicted Pfam domains between orthogroups that either have two copies (putative ohnologs) or 1 copy (candidates that lost their ohnolog copy). Significant domains (Fisher's exact test, FDR adjusted $P < 0.05$) are highlighted. More domains are significantly more likely to lose than retain a copy (counts of 82 and 5, respectively). (*E*) The barplot shows a subset of significant Pfam domains and their abundance in single copy genes (left) and putative ohnologs (right). Domains most likely to lose an ohnolog copy are involved in processes such as transport and binding. The five domains predicted to remain as ohnologs, such as zf-C2H2 and homeodomains, may be dosage sensitive. (*F*) The plot shows the frequency of KEGG pathway predictions between orthogroups that either have two copies (putative ohnologs) or one copy (candidates that lost their ohnolog copy). Significant domains (Fisher's exact test, FDR adjusted $P < 0.05$) are highlighted. (*G*) The bar plot shows the significant KEGG pathways and their abundance in single copy genes (left) and putative ohnologs (right). Genes with pathways involved in metabolism appear more likely to lose an ohnologous copy, while genes with pathways involved in processing are more likely to remain as ohnologs.

studies have suggested regulatory genes are more dosage sensitive (Birchler et al. 2005). Conversely, single copy genes were mostly involved in metabolism. This reflects work done by Shiu et al. (2006) where they showed genes involved in metabolism were less likely to be retained after duplication in mammals. Overall, we show which protein domains and metabolic pathways show functional bias in being lost or retained after WGD.

Ohnologs in *A. sudhausi* are Evolutionarily Constrained and Show Low Sequence Divergence

In order to date the *A. sudhausi* WGD and characterize the evolutionary distance to *A. seani*, we calculated divergence measures based on the orthogroup data. dN and dS estimates were calculated for ohnologs, as well as for orthologous genes that have copies in *A. seani* (as no second strain of *A. sudhausi* is currently available). The median dS for ohnologs is a relatively low value of 0.12 (interquartile range: 0.08 to 0.18), while the median dS for orthologs is 3.17 (fig. 3*A*). These findings imply that *A. sudhausi* and *A. seani* diverged a long time ago. We estimated the timing of the WGD by referring to Cutter (2008), where dS values were used to time events. The interquartile range of 0.08 to 0.18 would roughly correspond to an event that happened somewhere between 1.3 to 3.3 million years ago. Note that this timing of the WGD would only be in the case of autopolyploidization (where the entire chromosome set is duplicated). If allopolyploidization (species hybridization) had taken place, the dS values would not help in timing the event.

Finally, we calculated the dN/dS values for ohnologs and found the vast majority are below 1, with a median dN/dS value of 0.16. This indicates that most ohnologs are evolutionarily constrained (fig. 3*B*). We obtained the ohnolog dN/dS values in the top 5%ile. This subset had values of 1.35 at minimum, suggesting that they might underly positive selection. Interestingly, the majority of these ohnologs did not have associated Pfam predictions for them. Specifically, of the 397 ohnolog pairs, only 9 unique domains were predicted (supplementary Table S6, Supplementary Material online). We then looked at the ohnolog pairs with the highest dN/dS values and found no matching *C. elegans* orthologs (supplementary Table S7, Supplementary Material online), indicating strong sequence divergence in the ohnologs that are positively selected for.

A Third of Ohnologs are Differentially Expressed and Overlap With Sequentially Diverged Ohnologs

Next, we characterized the expression of ohnologs by calculating the fragments per kilobase of transcript per million mapped reads (FPKM) based on transcription analysis. First, we found that for most ohnologs both copies were indeed expressed (fig. 3*C*). Second, when we examined fold change (FC) between ohnolog pairs we found that 30.4% of ohnologs had expression that had a more than two-fold difference in transcript abundance (absolute $\log_2$-(FC) ≥ 1). However, the majority of ohnolog pairs have little to no difference in expression, with a peak at 0 (fig. 3*D*). Thus, there is little expression difference for ohnologs overall, a finding that is consistent with the multiple lines of evidence suggesting that the WGD occurred recently and that there has not been enough time for most ohnologs to diverge. Finally, we predicted the Pfam domains of these differentially expressed (DE) ohnologs, with the 7tm_1 domain showing the highest abundance (fig. 3*E*). We additionally looked at the ohnologs with the highest DE values and found the greatest difference was from ohnologs orthologous to the *C. elegans* ribosomal protein *rps-30* (supplementary Table S8, Supplementary Material online).

Expression divergence between ohnologs could potentially be explained by gene-specific regulatory evolution or by silencing of large chromosomal segments, as seen in dosage compensation (Pala et al. 2008). We, therefore, examined if there was positional bias by comparing the percentage of DE ohnologs for each contig. If there had been no bias, we would expect the distribution of DE ohnologs per contig to fall around 30.4%, which is the DE value calculated genome-wide. Indeed, the actual values were fairly evenly distributed around that figure (fig. 3*F*), with a median of 31.3% calculated. Thus, there is no evidence of ohnolog positional bias, suggesting expression divergence is more likely to be caused by gene-specific evolution.

Finally, we determined if there was overlap between the DE ohnologs and the ohnologs in the top 5 percentile dN/dS. Indeed, we found that the majority of ohnologs with high dN/dS values were also differentially expressed (fig. 3*G*). Specifically, there were more high dN/dS ohnologs that were differentially expressed than not (240 out of 397). This translates to 60.5% of the top dN/dS ohnologs being differentially expressed, which is significantly higher than the overall ohnolog median of 30.4% (Fisher's exact test, FDR adjusted $P < 0.001$). Altogether, this analysis reveals a link between expression and sequence divergence in ohnologs. This could be explained by positive selection on function and gene dosage but could also represent degeneration and silencing of one copy.

Repeat Elements, Particularly DNA Transposons, are Abundant in *A. sudhausi*

WGD is often succeeded by the expansion of TEs (Lien et al. 2016; Blanc-Mathieu et al. 2017; Marburger et al. 2018). We, therefore, analyzed repeat elements in *A. sudhausi* and compared it to those found in *P. pacificus* and *C. elegans*. Libraries were generated for *A. sudhausi,*

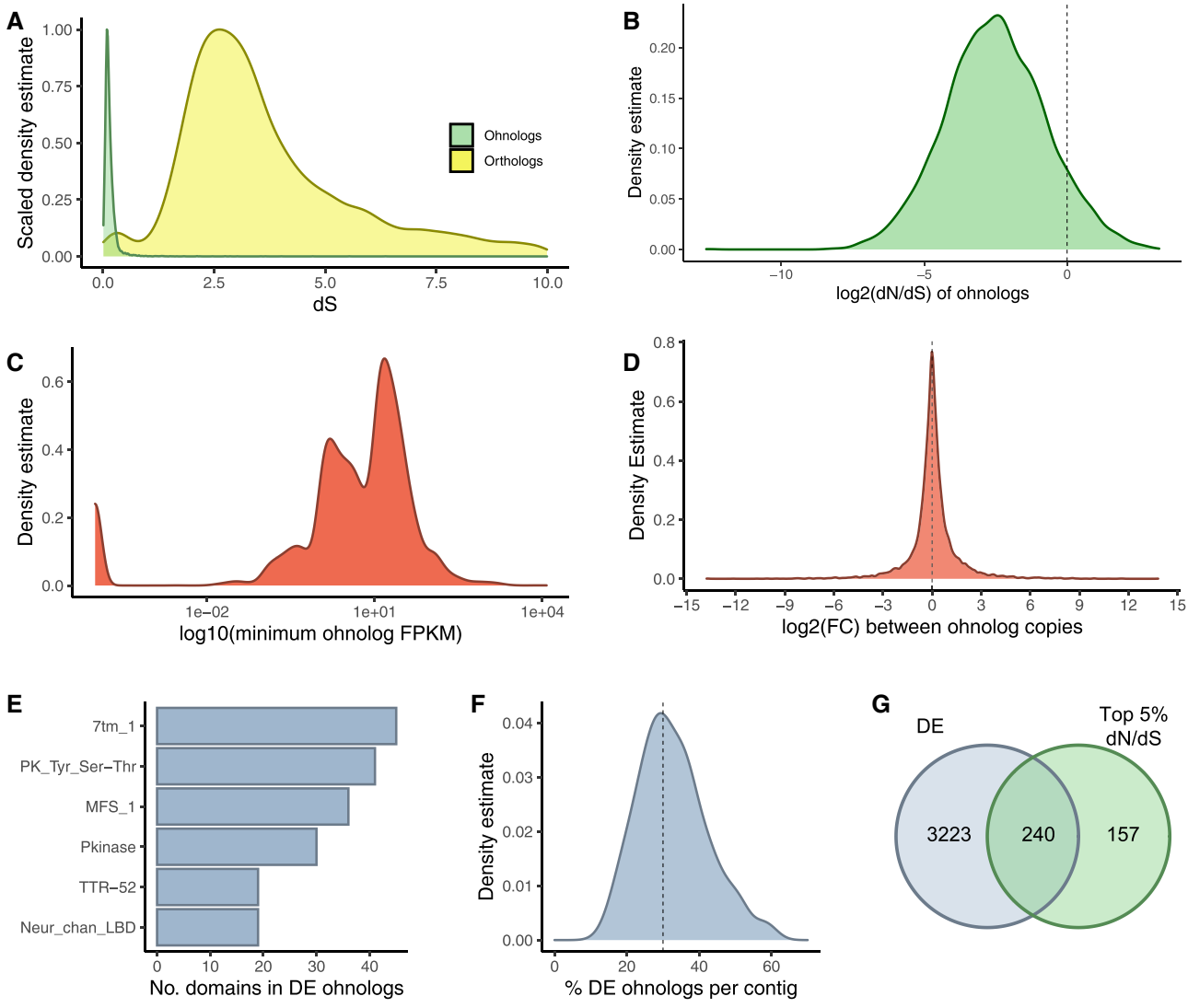


Fig. 3—Sequence and expression divergence in *A. sudhausi*. (*A*) The density plot shows the divergence (dS) of orthologs (between *A. sudhausi* and *A. seani*) and ohnologs. Orthologs exhibit high divergence, with a median value of 3.17 indicating that *A. seani* is distantly related. *Allodiplogaster sudhausi* ohnologs show much lower divergence (median of 0.12), suggesting the WGD is recent. (*B*) The distribution of dN/dS in ohnologs is relatively low, with the majority below 1 (indicated by the dashed line at 0 on the x axis $\log_2$ scale) and a median dN/dS value of 0.16, indicating most ohnologs are evolutionarily constrained. (*C*)The distribution of the lowest FPKM value of each ohnolog pair shows the majority of ohnologs exhibit expression of both copies. Those ohnologs with no expression are shown on the left of the plot, with a pseudocount of $\log_{10} 10^{-4}$. (*D*) Distribution of expression fold changes between ohnolog copies indicates the majority have no difference in expression. Only approximately 30.4% of ohnologs have an absolute $\log_2$(fold change) ≥ 1, which would indicate 1 copy has double or more expression than the other. (*E*) The bars show the most abundant Pfam domain predictions in the DE ohnologs (those with absolute $\log_2$(fold change) ≥ 1). (*F*) The distribution of the DE ohnologs on the contigs (only including contigs containing ≥ 10 ohnologs) shows no evidence of positional bias. The expected peak of 30.4% (indicated by the dashed line), correlates with the overall distribution of the DE ohnologs. (*G*) The Venn diagram displays the overlap between the top 5% highest dN/dS ohnologs and the DE ohnologs (absolute $\log_2$(FC) ≥ 1). The majority of the high dN/dS ohnologs are also differentially expressed (240 out of 397).

P. pacificus, and *C. elegans* using RepeatModeler based on the protocol of Athanasouli and Rödelsperger (2022). Strikingly, we found a far higher proportion of the *A. sudhausi* genome was covered by spans of repeats compared to the other two species. Specifically, 36.3% of the *A. sudhausi* genome is covered by repeat elements, which is far higher than the values of 21.3% in *P. pacificus* and 13.7% in *C. elegans* (fig. 4*A*). The high repeat rate may thus be linked to the WGD.

Further analysis revealed that DNA transposons are by far the most abundant type of repeats in *A. sudhausi* (fig. 4*B*). There are nearly 40Mb of DNA transposons in *A. sudhausi*, which is almost half the total classified repeats (87.5 Mb) (fig. 4*A*). DNA transposons, therefore, drive the

large repeat abundance in *A. sudhausi*. They are also the largest classification of repeats in *C. elegans* (fig. 4*B*); in contrast, the amount of DNA transposons is small in *P. pacificus*, despite them belonging to the same nematode family as *A. sudhausi*. This finding is not altogether unsurprising as there has been shown to be large diversity in transposon superfamilies even between strains in *C. elegans* (Laricchia et al. 2017).

We further broke down the families of the classified transposon types in each species and discovered that DNA/hAT-Ac elements take up the vast majority of repeat sequences in *A. sudhausi* (fig. 4*B*). The extensive amount of DNA transposons can be inferred to largely be due to hAT-Ac. The hAT superfamily, under which the Ac family belongs, is ancient and found in plants, animals, and fungi (Rubin et al. 2001; Wicker et al. 2007). It includes one of the first transposons ever discovered, the Ac element (McClintock 1950). DNA/hAT-Ac covers 3% of the *A. sudhausi* genome alone (fig. 4*D*); however, its numbers are relatively low in *P. pacificus* and *C. elegans*, indicating the expansion of hAT-Ac is specific to *A. sudhausi*. We performed pairwise comparisons of the *A. sudhausi* DNA/hAT-Ac elements to better determine when expansion of this family occurred. We found a peak in the percent identity for the comparisons at 100% (fig. 4*E*), suggesting the expansion is extremely recent. Taken together, our analysis of repeat elements revealed a high abundance in *A. sudhausi*, with a strong overrepresentation of DNA transposons.

Two *A. sudhausi dpy-1* Ohnologs Formed After WGD

WGD as observed for *A. sudhausi* can provide important insight into genome evolution; however, WGD can also limit functional investigation through forward and reverse genetic approaches. CRISPR/Cas9 is now a well-established molecular technique that enables the introduction of mutations into targeted loci (Jinek et al. 2012). We selected the *dpy-1* gene in order to establish CRISPR technology in *A. sudhausi* because the *dpy-1* gene is highly conserved and has an easy-to-score mutant phenotype that is shared in both *P. pacificus* and *C. elegans* (Kenning et al. 2004; Witte et al. 2014). In general, *Dumpy (Dpy)* mutants are shorter than wild type and more than 30 genes of *C. elegans* have been described that result in a *Dpy* phenotype when mutated (Brenner 1974).

BLAST searches revealed the conservation of *dpy-1* in many species, including members of the Rhabditidae (containing *C. elegans*) and Diplogastridae (containing *A. sudhausi* and *P. pacificus*) (fig. 5*A*). Based on the phylogeny of orthologous *dpy-1* sequences, it is clear there is one *dpy-1* copy ancestrally. Notably, and unsurprisingly, *A. sudhausi* has two *dpy-1* genes which diverged very recently, presumably ohnologs that resulted from the WGD. We annotated the gene structure of these ohnologs based on the transcriptome analysis (which was indexed against the genome) and termed them *Asu-dpy-1-A* and *Asu-dpy-1-B* in agreement with the previous nomenclature. The make-up of these ohnologous genes is very similar, with both containing 15 putative exons (fig. 5*B*). Overall, the low phylogenetic evolutionary distance and similar genetic structure suggest a very recent duplication event, consistent with the above analysis.

CRISPR/Cas9-induced *A. sudhausi dpy-1* Knock-out Mutants Demonstrate Genetic Redundancy

CRISPR/Cas9 technology was recently optimized for *P. pacificus* (Han et al. 2020; Hiraga et al. 2021), but nothing has previously been shown in *A. sudhausi*. We attempted to acquire knock-out mutations in *A. sudhausi* following the CRISPR/Cas9 protocol in *P. pacificus* (Witte et al. 2014), and targeted *dpy-1* ohnologs. We chose a gRNA target sequence that was conserved between the *Asu-dpy-1* ohnologs to try and target both genes together (fig. 5*B*). Initially, 30 young adult hermaphrodites were injected with the gRNA. We sequenced the resulting F1 progeny using gene-specific primer pairs to determine what mutations may have resulted. In total, 192 F1 worms were sequenced with two separate primer pairs, resulting in 384 sequences. Of these, 10 resulted in mutations in the targeted genes, with a mutation success rate of 2.6% (fig. 5*C*), where the majority of mutants come from just one injected adult (supplementary Table S9, Supplementary Material online). We also managed to acquire double mutants from a single injection, indicating that both loci can be targeted simultaneously. Together, we demonstrate the first successful example of CRISPR/Cas9 knockouts in a potential new nematode model system.

We found single *dpy-1* mutants displayed no obvious difference in length compared to wild-type worms (figs. 1*B* and 5*D*). However, the two mutant lines with knockouts in both *dpy-1* ohnologs displayed the characteristic dumpy phenotype with a short body length (fig. 5*D*). Interestingly, one double mutant had only in-frame mutations but this was still sufficient to generate a *Dpy* phenotype (supplementary Table S9, Supplementary Material online). Overall, our results shows expression of both *dpy-1* ohnologs needs to be disrupted in order to generate a morphological phenotype, suggesting one gene alone is sufficient for the wild-type phenotype to be produced. This is an example of genetic redundancy, which is common in recent duplication events (Lynch and Conery 2000) and is in line with our expression analysis that showed similar ohnolog expression levels (fig. 3*D*). Indeed, we confirmed there was no significant difference in expression between the *dpy-1* ohnologs (supplementary fig. S5, Supplementary Material online). With this work we provide a model to

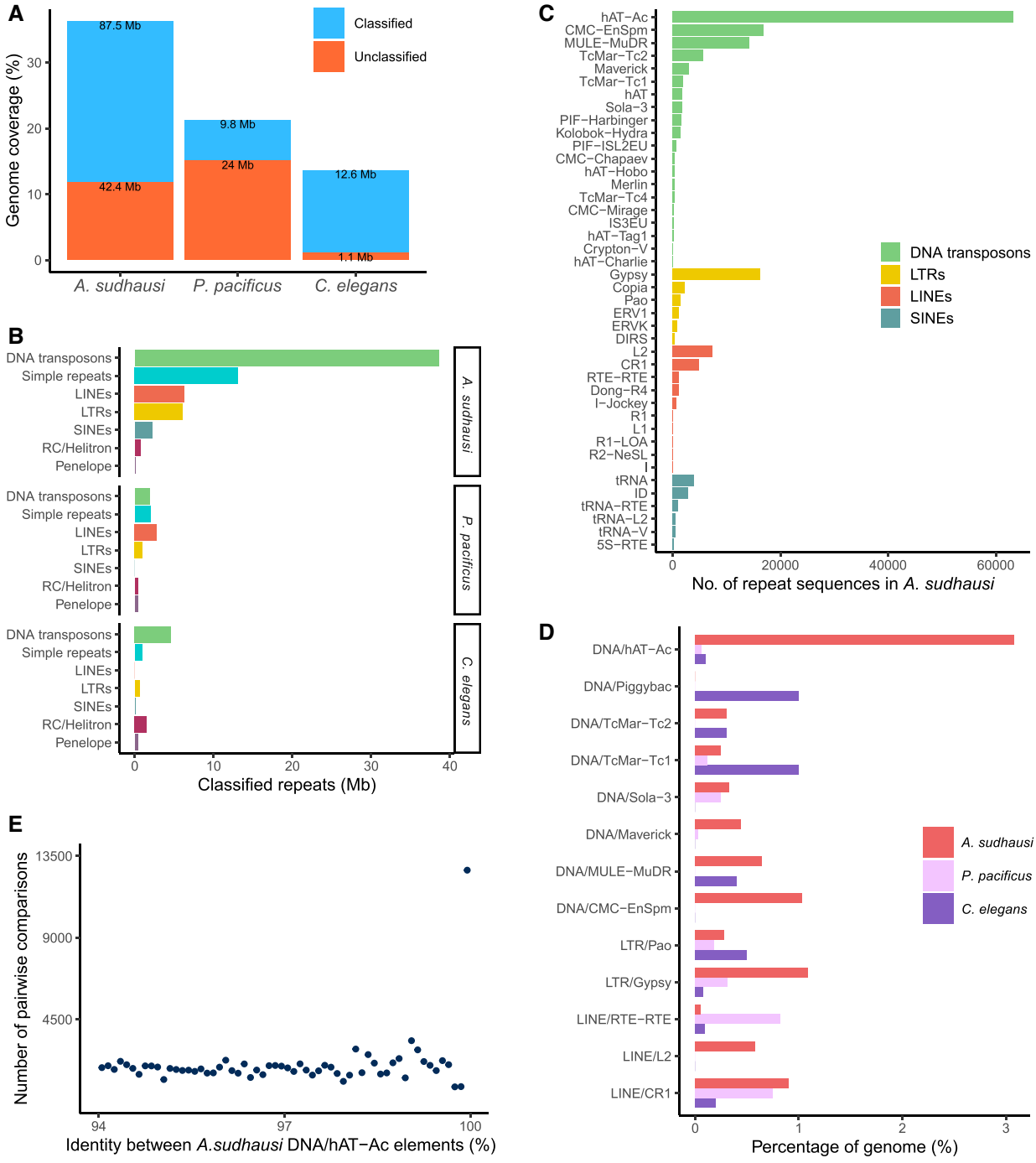
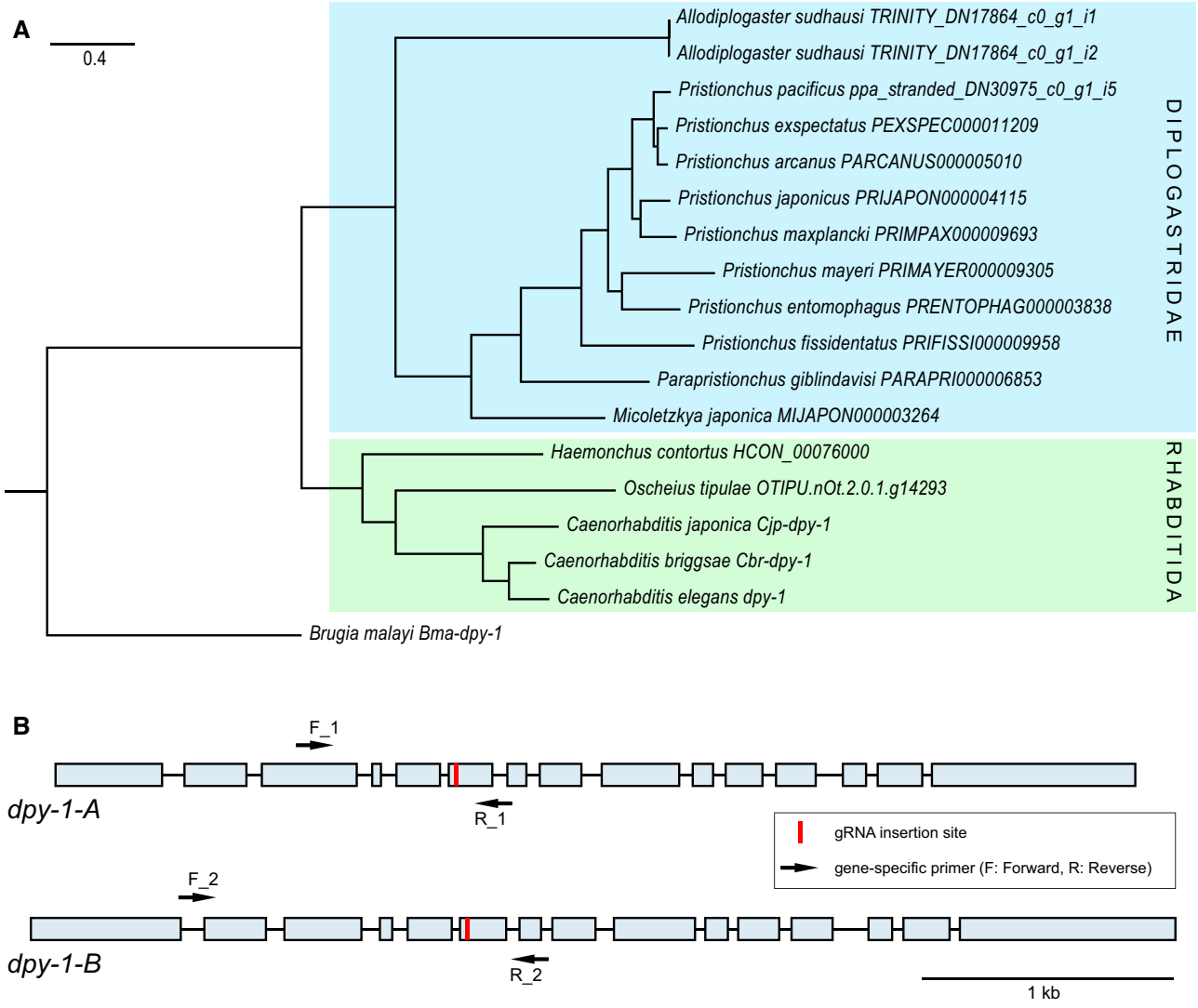


Fig. 4—Increased transposon activity in *A. sudhausi*. (*A*) The comparison of classified and unclassified spans of repeats in *A. sudhausi, P. pacificus*, and *C. elegans* reveals that the percentage of genome coverage is greatest in *A. sudhausi*, which also has the greatest overall repeat content (42.4 Mb unclassified and 87.5 Mb classified repeats). (*B*) The comparison of the types of classified repeat elements between *A. sudhausi*, *P. pacificus*, and *C. elegans* demonstrates a strikingly high amount of DNA transposons in *A. sudhausi*. (*C*) The number of repeat sequences in transposon families (as classified by RepeatModeler) in *A. sudhausi* is dominated by the hAT-Ac family (a DNA transposon element). DNA/CMC-EnSpm, DNA/MULE- MuDR, LINE/L2 and LTR/Gypsy are also overrepresented. (*D*) A comparison of transposon families between *A. sudhausi*, *P. pacificus*, and *C. elegans* (only looking at families that spanned ≥ 1 Mb in one species) shows that LINE/L2 and DNA/CMC-EnSpm appear unique to *A. sudhausi* while Piggybac is not found in either *P. pacificus* or *A. sudhausi*. DNA/hAT-Ac is found in all three species but is highly overrepresented in *A. sudhausi*. (*E*) The plot shows the percentage identity distribution of pairwise comparisons (blastn searches) between DNA/hAT-Ac elements. The peak toward 100% suggests that this transposon burst happened very recently and most likely after the WGD (in comparison with the dS value of 0.12 which translates to an expected identity of 88%).



C

Successful CRISPR mutations	
No. worms	192
No. mutants in *dpy-1-A*	5
No. mutants in *dpy-1-B*	5
No. double *dpy-1* mutants	2
CRISPR success rate (%)	2.6

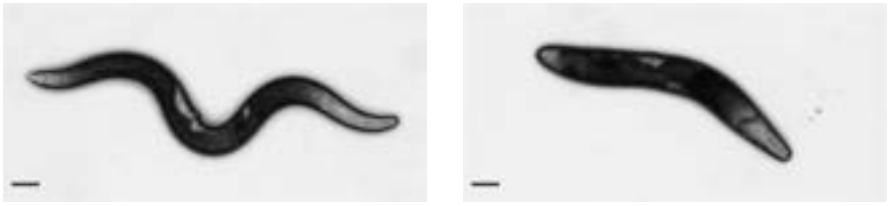


Fig. 5—CRISPR knockouts of *dpy-1* ohnologs demonstrate genetic redundancy in *A. sudhausi*. (*A*) The phylogenetic tree shows DPY-1 orthologues in the order Rhabditida. Protein sequences were obtained by BLASTing *C. elegans* DPY-1 against other nematodes. The majority of species have a single DPY-1 ortholog. *Allodiplogaster sudhausi* has two DPY-1 copies that only very recently duplicated, suggesting they are ohnologs that resulted from the WGD. (*B*) The figure shows the gene structure of *A. sudhausi dpy-1* ohnologs based on the transcriptome. Exons are shown as blocks and introns as connecting lines. The gRNA sequence is shared by both ohnologs and located in the sixth exon. The gene structure and number of exons (15) is similar for both ohnologs, showing the similarities that remain after whole genome duplication. (*C*) Table showing the output of CRISPR knock-out attempts. Out of 192 initial F1 (where P0 has been injected with the CRISPR/Cas9 gRNA), there were five knock-out mutations in *dpy-1-A* and five in *dpy-1-B*. In two cases, there were double mutants wherein both ohnologs had mutations. The overall successful CRISPR rate was 2.6%. (*D*) Single mutant knock-out *dpy-1-A* adult hermaphrodite. The nematode has a wild-type phenotype. Scale bar: 100 μM. (*E*) Double mutant knock-out *dpy-1-A* and *dpy-1-B* adult hermaphrodite. The length of the worm is noticeably smaller and dumpier than the single mutant. Scale bar: 100 μM.

examine the phenotypes of recent duplication events, particularly in WGD, and to further functionally examine the fates of these duplicate genes using CRISPR/Cas9.

Discussion

The role and consequences of WGD for driving evolution are still considered controversial, largely due to two important limitations. First, there are a limited number of organisms available for analysis, and second, most WGD events are of distant age. In the nematode *A. sudhausi*, we discovered WGD took place recently as did an expansion of the DNA transposon hAT-Ac family. With these findings, we were able to make several inferences about the early effects of WGD.

First, *A. sudhausi* differs from its relatives not just by the presence of a WGD, but by its (relatively) gigantic body size (fig. 1*B*–*E*). We hypothesize the large body size is due to the WGD that is specific to *A. sudhausi.* WGD is known to have an effect on body size although this differs depending on the lineage (Otto and Whitton 2000). In plants, it sometimes has an effect, but given the ancestral nature of most WGD events uncertainties remain (Clark and Donoghue 2018). In animals, the consensus is that WGD drives an increased body size in invertebrates, but not in vertebrates. Indeed, the evidence of WGD driving increased invertebrate body size is from studies long ago, where positive correlations between WGD and body size were seen in rotifers (Walsh and Zhang 1992), water fleas (Weider 1987), and even in nematodes (Madl and Herman 1979; Triantaphyllou and Riggs 1979). Flemming et al. (2000) previously showed endoreduplication drives an increased body size in nematodes. We, therefore, theorize the increased body size of *A. sudhausi* (fig. 1*B*–*E*) is due to the WGD.

Second, a benefit of evaluating a recent WGD is that the majority of genes have not yet diverged. This enabled us to determine if there was functional bias in the fate of ohnologs by examining which Pfam domains and KEGG pathways are more likely to be lost. We deemed two-copy orthogroups from *A. sudhausi* to be retained ohnologs and one-copy orthogroups to be nonfunctionalized genes. It can be argued that the lack of an identified duplicate could also be due to neofunctionalization. However, neofunctionalization is far rarer than nonfunctionalization (Moriyama and Koshiba-Takeuchi 2018). Thus, as we examined a large dataset, we feel confident the trend reflects nonfunctionalized genes. We found genes involved in metabolism, transportation, or those that encoded enzymes were significantly more likely to lose their duplicate after WGD (fig. 2*E* and *G*), which is consistent with results from other organisms. For instance, duplicate genes involved in metabolism and transport were also more likely to be lost in both humans and mice (Shiu et al. 2006). Additionally, ohnologs encoding enzymes were more likely to lose their duplicate in *Arabidopsis thaliana* (Seoighe and Gehring 2004). Thus, there appears to be some conserved functional bias across different kingdoms, potentially due to a shared underlying molecular basis. Studies have suggested that housekeeping genes are less likely to be retained while those with regulatory functions are more likely to be retained after duplication (Birchler et al. 2005; Shiu et al. 2006). This is believed to be due to differences in haplosufficiency or dosage sensitivity (Papp et al. 2003; Kondrashov and Koonin 2004). To support this, Kondrashov and Koonin (2004) determined genes encoding enzymes are highly haplosufficient, meaning they are far less dosage sensitive and better able to cope with losing a duplicate. We, therefore, suggest that dosage sensitivity is the main driver in the functional bias of gene fate that appears to be shared across different kingdoms.

Third, WGD is often followed by an expansion of TEs in both plants (Vicient and Casacuberta 2017) and animals (Lien et al. 2016; Blanc-Mathieu et al. 2017). Along these lines, we found a striking abundance of TEs in *A. sudhausi* (fig. 4*A*). TEs are subdivided into two classes: Class 1 retrotransposons and Class II DNA transposons (Makalowski et al. 2019), with the latter being most prevalent in *A. sudhausi* (fig. 1*B*). Studies have shown that TEs can play important roles in regulation, with their ability to produce novel networks and genes via mobility and insertion, or by generating new splice sites (Cosby et al. 2021). They are also involved in regulatory processes (Chénais et al. 2012), including the regulation of duplicate genes (Lisch 2013; Tan et al. 2021), meaning they may play important roles after WGD. The abundance of DNA transposons in *A. sudhausi* can largely be put down to the hAT-Ac family (fig. 4*C* and *D*), which falls under the ancient hAT superfamily (Arensburger et al. 2011). DNA/hAT is found in plants, animals, and fungi (Rubin et al. 2001), and has been shown to be a driver of evolutionary events. For example, DNA/hAT contributed highly to exon shuffling to generate novel genes in tetrapods (Cosby et al. 2021). DNA/hAT is active and recently expanded in the bat genus *Myotis,* which has high plasticity and diversification (Ray et al. 2006). Additionally, a DNA/hAT element is involved in neofunctionalization in *X. laevis* (Hayashi et al. 2022). In *A. sudhausi*, we determined the DNA/hAT-Ac family expanded very recently based on pairwise analysis (fig. 4*E*) and infer that, if autopolyploidization led to WGD, the expansion only happened afterward. As the DNA/hAT superfamily has been shown to contribute to various evolutionary processes, they may also be modulating *A. sudhausi* after the WGD event. Interestingly, a recent study revealed that the Homeobox and zf-C2H2 domains sometimes fuse with transposons to drive novelty and evolution in tetrapods (Cosby et al. 2021). As these domains show high retention in *A. sudhausi* ohnologs (fig. 2*E*), they may potentially work together with DNA/hAT-Ac to drive novelty. However,

further work is necessary to investigate this and the overall role of DNA transposons in regulating WGD and driving evolutionary processes.

Finally, the *A. sudhausi* system has many advantages for genetic analysis. Firstly, it is a self-fertilizing hermaphroditic nematode system. This means it has the benefits of short generation times and easy maintenance in an isogenic study system. Secondly, it behaves like a diploid with two alleles needing to be knocked out in both genes. Diploidization, the reversion of a polyploid system back to a diploid one, is a well-known, although not well understood, phenomenon that commonly occurs following WGD (Wolfe 2001). Lastly, we have CRISPR/Cas9 available to generate mutants. With this tool, it is possible to examine the phenotypic effects of genes. This is particularly useful for examining potential differences in gene fate. For example, neofunctionalized genes could be identified using this approach.

WGD has also been recorded in other nematodes, although it does not seem to be a common phenomenon. It has been well-studied in the plant parasitic genus *Meloidiogyne*, especially in the triploid *M. incognita,* which arose via allopolyploidization (Blanc-Mathieu et al. 2017; Szitenberg et al. 2017). Interestingly there has also been an expansion of TEs following WGD in these species (Blanc-Mathieu et al. 2017). In particular, a high abundance of DNA transposons was found in *M. incognita*, which is thought to drive genomic plasticity (Kozlowski et al. 2021). Additionally, WGD has been identified in the tetraploid plant parasite *Heterodera glycines* (Triantaphyllou and Riggs 1979). Only one other free-living nematode genus has been shown to undergo WGD to our knowledge, the panagrolaimid nematodes that underwent allopolyploidization (Schiffer et al 2019). The study of WGD in all these nematodes has added to our repertoire of knowledge and provides a useful comparative approach. However, of the known nematodes that underwent WGD, *A. sudhausi* is the only diploid and the only one in which CRISPR/Cas9 tools have been optimized, making it ideal for genetic analysis.

One unanswered question is how the WGD took place; whether it was a case of auto- or allopolyploidization. In *Meloidogyne* (Blanc-Mathieu et al. 2017) and panagroilaimid (Schiffer et al 2019) nematodes, the duplication was shown to be due to allopolyploidization. This was determined by comparing the genomes of a number of closely related species. Unfortunately, we only have one strain of *A. sudhausi* available and no other closely related sister species other than *A. seani* at hand, although we and others did multiple sampling trips to the type locality and related regions to find more strains. Additionally, *A. seani* and *A. sudhausi* diverged long ago based on the ortholog analysis (fig. 3*A*), meaning it is also not the best comparison. Although other species of *Allodiplogaster* have been reported, many are hard to keep in laboratory cultures and most are not available as a living material. Without more closely related species we cannot in full confidence make inferences about the origin of the WGD. Thus, the lack of other strains and more closely related species is the biggest limitation in using *A. sudhausi*. Comparative approaches would benefit from having more closely related systems at hand in order to better characterize the effects and origin of WGD. While this is a downside, few other animal systems have such closely related species.

There are a number of interesting avenues of research that could be followed in the future. The generation of reporter lines in *A. sudhausi* would enable us to see if ohnologs show differences in where they are expressed, as sometimes happens in subfunctionalization (Force et al. 1999). Another possibility is to examine the impact of dosage sensitivity, potentially by knocking out only one ohnolog and evaluating fitness consequences. An interesting avenue of research is to determine if there is any novelty in *A. sudhausi* besides the aforementioned increase in body size. WGD potentially drives evolution (Ohno 1970), while TEs drive genomic plasticity and phenotypic change (Faino et al. 2016; Kozlowski et al. 2021). It would, therefore, be interesting to examine if any phenotypic changes have been driven by these processes, particularly as *A. sudhausi* already displays phenotypic plasticity as a mouthform polyphenism (Fürst von Lieven 2008).

In conclusion, we show a recent WGD and DNA transposon expansion occurred in a free-living hermaphroditic nematode. We, therefore, provide another organism to join the small number of recorded animals that underwent WGD. As most recorded WGD events are ancient, the recency of this event allows analysis into genes that are still mostly redundant. With this study, we contributed further evidence to the body of work that examines WGD events and their impacts, particularly in animals.

Materials and Methods

Nematode Maintenance and Inbreeding

The following nematodes and laboratory strains were used in this study: *C. elegans* N2 (*C. elegans* Genetics Center); *P. pacificus* PS312 (Sommer et al. 1996), *A. sudhausi* SB413 (Bar-Eyal et al. 2008; Fürst von Lieven 2008), and *A. seani* RS1982 (Kanzaki et al. 2015). Note that *A. sudhausi* was originally described as *Koerneria sudhausi*. The taxonomic status of the genus *Koerneria* was recently revised, resulting in a split into the genera *Koerneria* and *Allodiplogaster* (Kanzaki et al. 2014). All strains were maintained on nematode growth medium (NGM) agar plates (Sommer et al. 1996). For inbreeding, a single late J4 stage hermaphrodite was moved to a fresh new plate to lay eggs. This step was repeated for 10 generations to eventually

isolate an inbred isogenic line that was used for subsequent downstream applications under the strain designation SB413B.

DNA Extraction and Sequencing

Allodiplogaster sudhausi nematodes were washed off of 100 NGM agar plates using M9 buffer and pelleted by centrifugation at 1,300 × g for 1 min. The pellet was washed twice in M9 before worms were frozen in liquid nitrogen and ground to a fine powder using a mortar and pestle. The powder was directly transferred into the lysis buffer from the QIAGEN genomic DNA extraction kit, which was used in combination with QIAGEN genomic tip columns (500/G) (QIAGEN, Hamburg, Germany). The protocol was performed following the manufacturer's instructions. All steps involving vortexing of the sample were replaced by inversion to limit unwanted DNA shearing. DNA quality and quantity were determined with a NanoDrop ND 1000 spectrometer (PeqLab, Erlangen, Germany), a Qubit 2.0 Fluorometer (Thermo Fisher Scientific, Waltham, USA), and by a Femto pulse system (Agilent, CA, USA). A total of 20 μg *A. sudhausi* genomic DNA was sheared to a target fragment size of 45 kb using a needle. A 45-kb template library was prepared using the BluePippin size-selection system according to the manufacturer's protocol (P/N 100-286-000-07, Pacific Biosciences, California, USA). The final library was sequenced on a Pacific Biosciences Sequel instrument following the Magbead loading protocol and version 1.2.1 sequencing kits. A total of two SMRT cells (version 1.2.1) generated 40 Gb (100-fold coverage).

Genome Assembly and Evaluation

Raw long-read data were assembled using Canu version 1.8 (Koren et al. 2017). The completeness of the genome assembly was evaluated using the BUSCO software (version 3.0.1, with the -m genome option against the nematode odb9 dataset) (Simão et al. 2015). To investigate the sequencing coverage distribution, we downloaded previously generated Illumina sequencing data from the European Nucleotide Archive (Accessions: ERR2208557, ERR2208558, SRR12424054, and SRR12424056) (Sieriebriennikov et al. 2018; Casasa et al. 2021) and aligned these reads against the *A. sudhausi* genome with BWA mem program (version 0.7.17) (Li and Durbin 2009). The coverage profiles were generated from the resulting alignment files by the samtools depth program (version 0.1.18) (Li et al. 2009).

Gene Annotations

We generated a transcriptome assembly of mixed-stage RNA-seq data from *A. sudhausi* with the help of the Trinity program (version 2.2.0, with the –normalize_reads option) (Grabherr et al. 2011). This transcriptome assembly was combined with the community-curated gene annotations of *P. pacificus* (El Paco gene annotations version 3) (Athanasouli et al. 2020) to generate evidence-based gene annotations for *A. sudhausi*. Specifically, both datasets were aligned against the *A. sudhausi* assembly by the exonerate program (version 2.2.0, with –bestn 2 –dnawordlen 20 –maxintron 20,000 options) (Slater and Birney 2005). Subsequently, the alignments were processed by the PPCAC software (version 1.0) to select one representative gene model with the longest open reading frame per locus (Rödelsperger 2021). Final gene annotations were assessed by the BUSCO software (version 3.0.1, with the -m prot option against the nematode odb9 dataset) (Simão et al. 2015; Rödelsperger 2021).

Chromosome Staining

We made a 1:100 dilution of 20 mM Hoechst 33342 dye (Chazotte 2011) in sperm salts (50 mM PIPES, 25 mM KCl, 1 mM $MgSO_4$, 45 mM NaCl, 2 mM $CaCl_2$, pH 7). 10μl of this solution was then pipetted onto a microscope slide. Ten adult hermaphrodites were moved onto the solution. A surgical blade was used to decapitate the heads of the worms (cutting right below the pharynx). This resulted in the gonadal arms being pushed out due to the internal pressure, freeing them from the worm body. A cover slip was then placed on top and the worms were imagined using a Nomarsky DIC microscope and the chromosomes were subsequently counted.

Comparative Genomic Analysis, KEGG, and Pfam Annotations

For comparative genomic analyses, we compiled additional protein data for *A. seani*, *P. pacificus* (El Paco gene annotations version 3) (Athanasouli et al. 2020), and *C. elegans* (WormBase ParaSite version WBPS16) (Howe et al. 2016). The dataset for *A. seani* was generated from mixed-stage RNA-seq data and was assembled with the Trinity program (version 2.2.0, with the –normalize_reads option) (Grabherr et al. 2011). To reduce isoform information, we selected the assembled transcript with the longest open reading frame per Trinity gene and further clustered protein sequences with the cd-hit program (version 4.3) (Li and Godzik 2006). Protein domains were annotated by the hmmsearch program (version 3.3, with -E 0.001 option) using the Pfam-A data (version 3.1b2) as target database (Bateman et al. 1999). Clusters of orthologous genes were generated by the OrthoFinder software (version 2.5.2). From the resulting orthogroups, we extracted orthogroups with ohnologs (orthogroups with two *A. sudhausi* copies and at most one copy in *A. seani*) and orthogroups with *A. seani* orthologs (either one or two copies in *A. sudhausi* and one copy in *A. seani*). dN and dS values were computed from intraspecies pairs of the ohnolog

orthogroups and cross-species pairs of the orthogroups with *A. seani* orthologs. This was done by aligning protein sequences with MUSCLE (version 3.8.31) (Edgar 2004), conversion into codon alignment with PAL2NAL (version 14) (Suyama et al. 2006), and divergence estimation with the codeml program of PAML (version 4.9) (Yang 2007). Metabolic pathway annotations were generated by identification of orthologs in the KEGG database using the blastkoala web application (with the "eukaryotes" and "family_eukaryotes" values for taxonomic group and database level, respectively) (Kanehisa et al. 2016). Genes with orthologs in the KEGG database were then annotated with the corresponding KEGG accessions.

Ohnolog Expression

The FPKM values were calculated from the *A. sudhausi* transcriptome. This was done by 1) Summing up the total number of reads and dividing by 1×10^6 to get the scaling factor (per million), 2) Dividing each read count by the scaling factor, 3) Dividing these values by the length of each gene (in kb). This was calculated separately for four biological replicates. The mean value for each gene was then calculated to get the FPKM. Fold change was calculated by dividing one ohnolog by its duplicate. To examine position bias of ohnologs on contigs, we first filtered for contigs that had ≥ 10 ohnologs. We then examined the distribution of ohnologs on the contigs (by referring to the assembly) and calculated the percentage of DE ohnologs (absolute $\log_2(FC) \geq 1$) for each contig.

Repeat Annotation

We used RepeatModeler2 (version 2.0.1, parameters: -LTRstruct) (Flynn et al. 2020) for de novo repeat detection in *A. sudhausi* and compared the TE content with available TE data for *P. pacificus* (Athanasouli and Rödelsperger 2022) and *C. elegans*. The *C. elegans* TE dataset was created using RepeatMasker's incorporated libraries (parameters: -species worm) (A.F.A. Smit, R. Hubley, and P. Green, http://repeatmasker.org/). Based on the RepeatModeler2 classification, we calculated the percentage of the genome coverage by repeat elements as well as the overall length of the TEs and simple repeats across the three species. We investigated the span of the superfamilies present in *A. sudhausi* for the four major TEs orders (DNA transposons, LINEs, LTRs and SINEs). Furthermore, we identified the superfamilies spanning at least 1 Mb in any of the 3 species and evaluated the percentage of the genome covered by each superfamily in the organisms being compared. To time the DNA/hAT-Ac expansion, we extracted DNA/hAT-Ac sequences from the RepeatModeler/RepeatMasker output files and ran all-against-all blastn searches (version 2.10.1, with -dust no, -evalue 0.001, -qcov _hspperc 80, and -perc_indentity 80 options). Pairwise percentage identities were extracted from blastn hits between different DNA/hAT-Ac elements (exclusion of self-hits) that span at least 100 nucleotides.

CRISPR Injection and *dpy-1* Mutant Identification

CRISPR knockouts were generated following the *P. pacificus* protocol (Witte et al. 2014). CRISPR RNAs (crRNAs) and trans-activating crispr RNA (tracrRNA) (Cat. No. 1072534) were synthesized by Integrated DNA Technologies (IDT), while the Cas9 endonuclease (Cat. No. 1081058) was purchased from IDT. The CRISPR/Cas9 complex was prepared by mixing 0.5 mg/ml Cas9 nuclease, 0.1 mg/ml tracrRNA, and 0.056 mg/ml crRNA in the TE buffer followed by a 10-min incubation at 37° C. Microinjections were performed in late-stage J4 hermaphrodites following standard practice using an Eppendorf microinjection system. The gRNA (CTCAAAGAGAACTCCAGCTG) sequence was designed just before an NGG PAM site and targeted exon six of both *dpy-1* genes. Gene specific primers were designed for both. Transcriptomic reads were mapped against the *A. sudhausi* genome using IGV (Integrative Genomics Viewer, version 2.8.9) to see where the coding regions were. For *dpy-1-A*, the forward primer F_1 (5′-CTTCAGGCACCCCTCTAGGCA-3′) was designed in exon 3 and the reverse R_1 (5′-GCAACATGCTCGGCAAGGCT-3′) in exon 6 (amplicon size: 650 bp). For *dpy-1-B*, the forward primer F_2 (5′-CCCAAACTCATTCGTTGCC-3′) was designed in exon 2 and the reverse R_2 (5′-CACTTAATTCCACGCTCTTC′-3) in the intron between exons 6 and 7 (amplicon size: 1200 bp). Polymerase chain reactions (PCRs) were run using both primer pairs on F1 young adults of injected worms. Heterozygotes were identified via Sanger sequencing and homozygous mutants were then obtained by self-fertilizing heterozygous F1 to eventually obtain homozygous knock-out mutants.

Phylogeny Generation

A subset of species from Susoy et al. (2015) was edited using figTree software v.1.4.4 (tree.bio.ed.ac.uk/software/figtree) to obtain the species tree. For the DPY-1 phylogeny, the protein sequence *C. elegans* DPY-1 isoform a (the longest) was obtained from wormbase.org (version WS284). This sequence was used to BLAST (Altschul et al. 1990) (query type: protein) against selected nematodes on parasite.wormbase.org (version WBPS16). For species belonging to the *Pristionchus* genus, pristionchus.org (version 2.0.0.rc8) was used to BLAST DPY-1 against the protein databases. It should be noted that we obtained two copies each in *P. expectatus* and *P. arcanus*, but believe the copies are due to heterozygosity which is common in out-crossing species. Thus we only took one copy of each for the

phylogeny. The *A. sudhausi* DPY-1 proteins were translated from the best matches in the *A. sudhausi* transcriptome. The DPY-1 protein phylogeny was then generated using RAxML version 8.2.12 (raxmlHPC -f a -m PROTGAMMAAUTO -p 12345 -x 12345 -N 100) (Stamatakis 2014).

Statistical Analysis

Each unique domain or pathway prediction per annotation was counted and compared both within and between species. A Fisher's exact test was run using the 2X2 matrix, whereby each gene with a given domain/pathway was compared against the whole dataset to identify those that were significantly different. The *P* values were adjusted using the false discovery rate (FDR). All analyses were performed using RStudio Statistical Software (v1.4.1717; R Core Team 2021).

Supplementary material

Supplementary data are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We thank Waltraud Röseler for help with the generation of the PacBio library and all members of the Sommer Lab for discussion. This study was funded by the Max-Planck Society.

Author Contributions

S.S.W. and R.J.S. designed the study. S.S.W., M.A., and CR conducted genome analysis and H.W. performed the CRISPR experiments. S.S.W. and R.J.S. wrote the manuscript with input from others.

Data availability

Raw reads, genome, and transcriptome assemblies have been submitted to the European Nucleotide archive under the study accession PRJEB48369.

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410.
Amores A, et al. 1998. Zebrafish hox clusters and vertebrate genome evolution. Science 282:1711–1714.
Arensburger P, et al. 2011. Phylogenetic and functional characterization of the hAT transposon superfamily. Genetics 188:45–57.
Athanasouli M, et al. 2020. Comparative genomics and community curation further improve gene annotations in the nematode *Pristionchus pacificus*. BMC Genomics 21:708.
Athanasouli M, Rödelsperger C. 2022. Analysis of repeat elements in the *Pristionchus pacificus* genome reveals an ancient invasion by horizontally transferred transposons. BMC Genomics 23:523.
Aury J-M, et al. 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature 444:171–178.
Avise JC. 2011. Hermaphroditism: A primer on the biology, ecology, and evolution of dual sexuality. New York: Columbia University Press.
Bar-Eyal M, Sharon E, Spiegel Y, Oka Y. 2008. Laboratory studies on the biocontrol potential of the predatory nematode *Koerneria sudhausi* (Nematoda: Diplogastridae). Nematology 10:633–637.
Barrière A, et al. 2009. Detecting heterozygosity in shotgun genome assemblies: lessons from obligately outcrossing nematodes. Genome Res 19:470–480.
Bateman A, et al. 1999. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. Nucleic Acids Res 27: 260–262.
Biddle JF, Ragsdale EJ. 2020. Regulators of an ancient polyphenism evolved through episodic protein divergence and parallel gene radiations. Proc R Soc B: Biol Sci 287:20192595.
Birchler JA, Riddle NC, Auger DL, Veitia RA. 2005. Dosage balance in gene regulation: biological implications. Trends Genet 21: 219–226.
Blanc-Mathieu R, Perfus-Barbeoch L, Aury JM, Da Rocha M, Gouzy J, Sallet E, Martin-Jimenez C, Bailly-Bechet M, Castagnone-Sereno P, Flot JF, et al. 2017. Hybridization and polyploidy enable genomic plasticity without sex in the most devastating plant-parasitic nematodes. PLoS Genet 13:e1006777.
Blomme T, et al. 2006. The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol 7:R43.
Brenner S. 1974. The genetics of *Caenorhabditis elegans*. Genetics 77: 71–94.
Bridges CB. 1936. The bar "gene" a duplication. Science 83:210–211.
Carlton PM, Davis RE, Ahmed S. 2022. Nematode chromosomes. Genetics 221:1.
Casasa S, Biddle JF, Koutsovoulos GD, Ragsdale EJ. 2021. Polyphenism of a novel trait integrated rapidly evolving genes into ancestrally plastic networks. Mol Biol Evol 38:331–343.
Chazotte B. 2011. Labeling nuclear DNA with hoechst 33342. Cold Spring Harb Protoc 2011:pdb.prot5557.
Chénais B, Caruso A, Hiard S, Casse N. 2012. The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. Gene 509:7–15.
Clark JW, Donoghue PCJ. 2018. Whole-genome duplication and plant macroevolution. Trends Plant Sci 23:933–945.
Cosby RL, et al. 2021. Recurrent evolution of vertebrate transcription factors by transposase capture. Science 371:6531.
Cutter AD. 2008. Divergence times in *Caenorhabditis* and *Drosophila* inferred from direct estimates of the neutral mutation rate. Mol Biol Evol 25:778–786.
Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol 3:e314.
Dieterich C, et al. 2008. The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism. Nat Genet 40:1193–1198.
Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 1–19.
Faino L, et al. 2016. Transposons passively and actively contribute to evolution of the two-speed genome of a fungal pathogen. Genome Res 26:1091–1100.
Flemming AJ, Shen ZZ, Cunha A, Emmons SW, Leroi AM. 2000. Somatic polyploidization and cellular proliferation drive body size evolution in nematodes. Proc Natl Acad Sci USA 97:5285–5290.
Flynn JM, et al. 2020. Repeatmodeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci USA 117: 9451–9457.

Force A, et al. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545.
Fürst von Lieven A. 2008. *Koerneria sudhausi* n. sp. (Nematoda: Diplogastridae); a hermaphroditic diplogastrid with an egg shell formed by zygote and uterine components. Nematology 10: 27–45.
Gonzalez de la Rosa PM, et al. 2021. A telomere-to-telomere assembly of *Oscheius tipulae* and the evolution of rhabditid nematode chromosomes. G3 11(1):jkaa020.
Grabherr MG, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29: 644–652.
Haldane JBS. 1932. The causes of evolution. London: Longmans, Green and Co.
Han Z, et al. 2020. Improving transgenesis efficiency and CRISPR-associated tools through codon optimization and native intron addition in *Pristionchus* nematodes. Genetics 216:947–956.
Hayashi S, et al. 2022. Neofunctionalization of a noncoding portion of a DNA transposon in the coding region of the chimerical sex-determining gene dm-W in *Xenopus* frogs. Mol Biol Evol 39:1–6.
Hiraga H, Ishita Y, Chihara T, Okumura M. 2021. Efficient visual screening of CRISPR/Cas9 genome editing in the nematode *Pristionchus pacificus*. Dev Growth Differ 63:488–500.
Howe KL, et al. 2016. Wormbase 2016: expanding to enable helminth genomic research. Nucleic Acids Res 44:D774–D780.
Inoue H, et al. 2005. The *C. elegans* p38 MAPK pathway regulates nuclear localization of the transcription factor SKN-1 in oxidative stress response. Genes Dev 19:2278–2283.
Jinek M, et al. 2012. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. Science 337:816–821.
Kanehisa M, Goto S. 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30.
Kanehisa M, Sato Y, Morishima K. 2016. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. J Mol Biol 428:726–731.
Kanzaki N, Giblin-Davis RM, Ragsdale EJ. 2015. *Allodiplogaster josephi* n. sp. and *A. seani* n. sp. (Nematoda: Diplogastridae), associates of soil-dwelling bees in the eastern USA. Nematology 17:831–863.
Kanzaki N, Ragsdale E, Giblin-Davis R. 2014. Revision of the paraphyletic genus *Koerneria* Meyl, 1960 and resurrection of two other genera of Diplogastridae (Nematoda). ZooKeys 442:17–30.
Kenning C, Kipping I, Sommer RJ. 2004. . Isolation of mutations with dumpy-like phenotypes and of collagen genes in the nematode *Pristionchus pacificus*. Genesis 40:176–183.
Kondrashov FA, Koonin EV. 2004. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. Trends Genet 20:287–290.
Koren S, et al. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 27:722–736.
Kozlowski DKL, et al. 2021. Movements of transposable elements contribute to the genomic plasticity and species diversification in an asexually reproducing nematode pest. Evol Appl 14:1844–1866.
Laricchia KM, Zdraljevic S, Cook DE, Andersen EC. 2017. Natural variation in the distribution and abundance of transposable elements across the *Caenorhabditis elegans* species. Mol Biol Evol 34: 2187–2202.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.
Li H, et al. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079.
Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659.
Lien S, et al. 2016. The Atlantic salmon genome provides insights into rediploidization. Nature 533:200–205.
Lisch D. 2013. How important are transposons for plant evolution? Nat Rev Genet 14:49–61.
Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155.
Madl JE, Herman RK. 1979. Polyploids and sex determination in *Caenorhabditis elegans*. Genetics 93:393–402.
Maere S, et al. 2005. Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci USA 102:5454–5459.
Makalowski W, Gotea V, Pande A, Makalowska I. 2019. Transposable elements: classification, identification, and their use as a tool for comparative genomics. Methods Mol Biol 1910:177–207.
Marburger S, et al. 2018. Whole genome duplication and transposable element proliferation drive genome expansion in Corydoradinae catfishes. Proc R Soc B Biol Sci 285:20172732.
Marcet-Houben M, Gabaldón T. 2015. Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the Baker's yeast lineage. PLoS Biol 13:e1002220.
McClintock B. 1950. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci USA 36:344–355.
Moriyama Y, et al. 2016. Evolution of the fish heart by sub/neofunctionalization of an elastin gene. Nat Commun 7:10397.
Moriyama Y, Koshiba-Takeuchi K. 2018. Significance of whole-genome duplications on the emergence of evolutionary novelties. Brief Funct Genomics 17:329–338.
Ohno S. 1970. Evolution by gene duplication. London: Springer-Verlag.
Otto SP, Whitton J. 2000. Polyploid incidence and evolution. Annu Rev Genet 34:401–437.
Pala I, Coelho MM, Schartl M. 2008. Dosage compensation by gene-copy silencing in a triploid hybrid fish. Curr Biol 18:1344–1348.
Papp B, Pál C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. Nature 424:194–197.
Parisod C, Holderegger R, Brochmann C. 2010. Evolutionary consequences of autopolyploidy. New Phytol 186:5–17.
Prabh N, et al. 2018. Deep taxon sampling reveals the evolutionary dynamics of novel gene families in *Pristionchus nematodes*. Genome Res 28:1664–1674.
R Core Team. 2021. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.
Ray DA, Pagan HJT, Thompson ML, Stevens RD. 2006. Bats with hATs: evidence for recent DNA transposon activity in genus *Myotis*. Mol Biol Evol 24:632–639.
Rödelsperger C, et al. 2017. Single-molecule sequencing reveals the chromosome-scale genomic architecture of the nematode model organism *Pristionchus pacificus*. Cell Rep 21:834–844.
Rödelsperger C, et al. 2018. Phylotranscriptomics of *Pristionchus nematodes* reveals parallel gene loss in six hermaphroditic lineages. Curr Biol 28:3123–3127.e5.
Rödelsperger C, et al. 2019. Crowdsourcing and the feasibility of manual gene annotation: a pilot study in the nematode *Pristionchus pacificus*. Sci Rep 9:1–9.
Rödelsperger C. 2021. The community-curated *Pristionchus pacificus* genome facilitates automated gene annotation improvement in related nematodes. BMC Genomics 22:216.
Rubin E, Lithwick G, Levy AA. 2001. Structure and evolution of the hAT transposon superfamily. Genetics 158:949–957.
Sagasti A, Hobert O, Troemel ER, Ruvkun G, Bargmann CI. 1999. Alternative olfactory neuron fates are specified by the LIM homeobox gene lim-4. Genes Dev 13(14):1794–1806.
Sarbassov DD, Ali SM, Sabatini DM. 2005. Growing roles for the mTOR pathway. Curr Opin Cell Biol 17:596–603.

Schiffer PH, et al. 2019. Signatures of the evolution of parthenogenesis and cryptobiosis in the genomes of Panagrolaimid nematodes. iScience 21:587–602.
Schroeder NE. 2021. Introduction to *Pristionchus pacificus* anatomy. J Nematol 53:1–9.
Seger R, Krebs EG. 1995. The MAPK signaling cascade. FASEB J 9: 726–735.
Seoighe C, Gehring C. 2004. Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome. Trends Genet 20:461–464.
Session AM, et al. 2016. Genome evolution in the allotetraploid frog *Xenopus laevis*. Nature 538:336–343.
Shiu SH, Byrnes JK, Pan R, Zhang P, Li WH. 2006. Role of positive selection in the retention of duplicate genes in mammalian genomes. Proc Natl Acad Sci USA 103:2232–2236.
Sieriebriennikov B, et al. 2018. A developmental switch generating phenotypic plasticity is part of a conserved multi-gene locus. Cell Rep 23:2835–2843.e4.
Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210–3212.
Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:1–11.
Sommer RJ, Carta LK, Kim SY, Sternberg PW. 1996. Morphological, genetic and molecular description of *Pristionchus pacifiais* sp. n. (Nematoda : Neodiplogastridae). Fundam Appl Nematol 19: 511–521.
Sommer RJ. 2015. Nematoda. In: Evolutionary developmental biology of invertebrates. Vol. 3. Berlin: Springer. p. 15–34.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30: 1312–1313.
Susoy V, Ragsdale EJ, Kanzaki N, Sommer RJ. 2015. Rapid diversification associated with a macroevolutionary pulse of developmental plasticity. eLife 4:1–39.
Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 34:W609–W612.
Szitenberg A, et al. 2017. Comparative genomics of apomictic root-knot nematodes: hybridization, ploidy, and dynamic genome change. Genome Biol Evol 9:2844–2861.
Tan S, et al. 2021. DNA Transposons mediate duplications via transposition-independent and -dependent mechanisms in metazoans. Nat Commun 12:1–14.
Thomas JH. 2006. Adaptive evolution in two large families of ubiquitin-ligase adapters in nematodes and plants. Genome Res 16:1017–1030.
Triantaphyllou AC, Riggs RD. 1979. Polyploidy in an amphimictic population of *Heterodera glycines*. J Nematol 11:371–376.
Troemel ER, Kimmel BE, Bargmann CI. 1997. Reprogramming chemotaxis responses: sensory neurons define olfactory preferences in *C. elegans*. Cell 91:161–169.
Vicient CM, Casacuberta JM. 2017. Impact of transposable elements on polyploid plant genomes. Ann Bot 120:195–207.
Walsh EJ, Zhang L. 1992. Polyploidy and body size variation in a natural population of the rotifer euchlanis dilatata. J Evol Biol 5: 345–353.
Weider LJ. 1987. Life history variation among low-Arctic clones of obligately parthenogenetic *Daphnia pulex*: a diploid-polyploid complex. Oecologia 73:251–256.
Wicker T, et al. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–982.
Wilson DF. 2017. Oxidative phosphorylation: regulation and role in cellular and tissue metabolism. J Physiol (Lond) 595:7023–7038.
Witte H, et al. 2014. Gene inactivation using the CRISPR/Cas9 systemin the nematode *Pristionchus pacificus*. Dev Genes Evol 225:55–62.
Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. Nat Rev Genet 2:333–341.
Wolfe KH. 2015. Origin of the yeast whole-genome duplication. PLoS Biol 13:e1002221.
Wood WB. 1988. The nematode Caenorhabditis elegans. New York: Cold Spring Harbor Laboratory.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591.
Yin D, et al. 2018. Rapid genome shrinkage in a self-fertile nematode reveals sperm competition proteins. Science 359:55–61.
Zakon HH, Lu Y, Zwickl DJ, Hillis DM. 2006. Sodium channel genes and the evolution of diversity in communication signals of electric fishes: convergent molecular evolution. Proc Natl Acad Sci 103: 3675–3680.

Associate editor: Prof. Kenneth Wolfe

A New Hope: A hermaphroditic nematode enables analysis of a recent whole genome duplication event

Sara S. Wighard, Marina Athanasouli, Hanh Witte, Christian Rödelsperger & Ralf J. Sommer*

Supplementary Online Material

Content

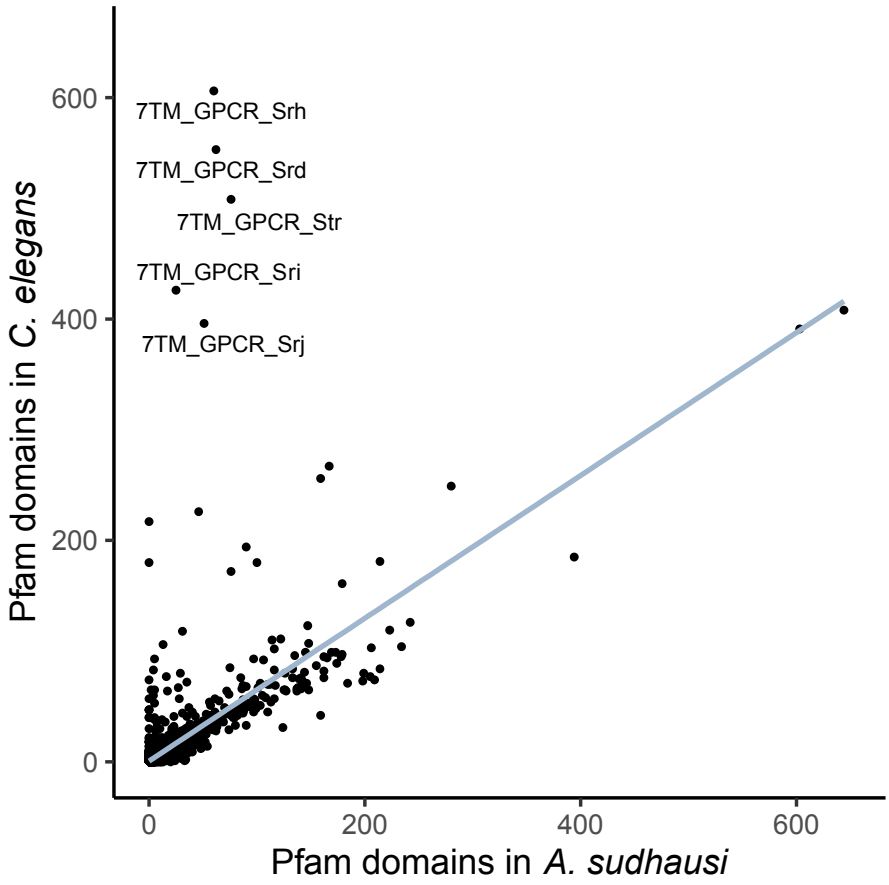


Fig. S1. A comparison between *A. sudhausi* and *C. elegans* predicted Pfam domains

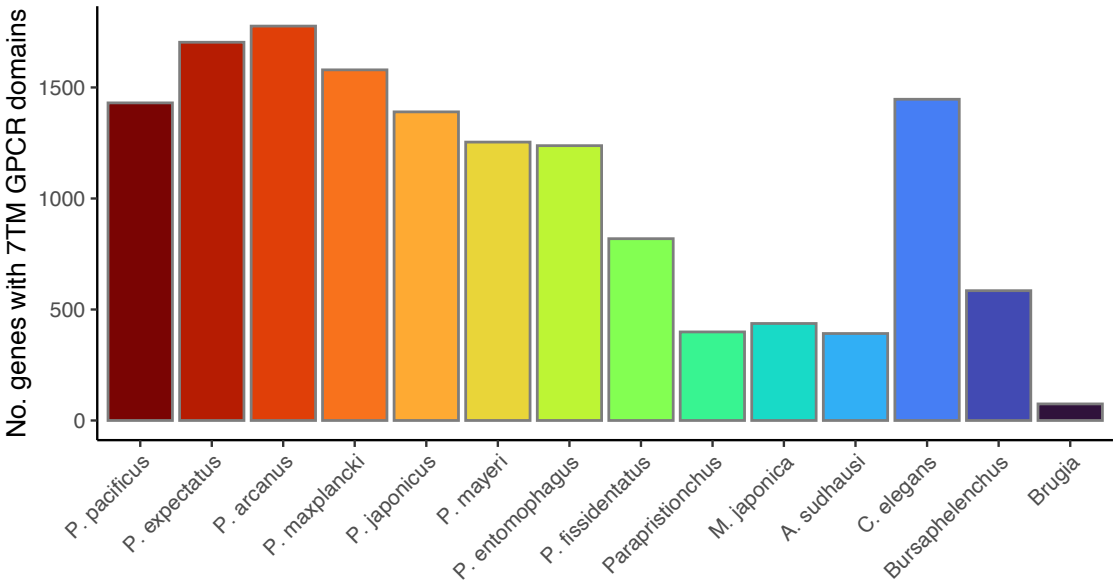


Fig. S2. A comparison between the number of domains with 7TM-GPCRs across various nematode species shows 7TM-GPCRs are only highly overrepresented in *C. elegans* and *Pristionchus* species

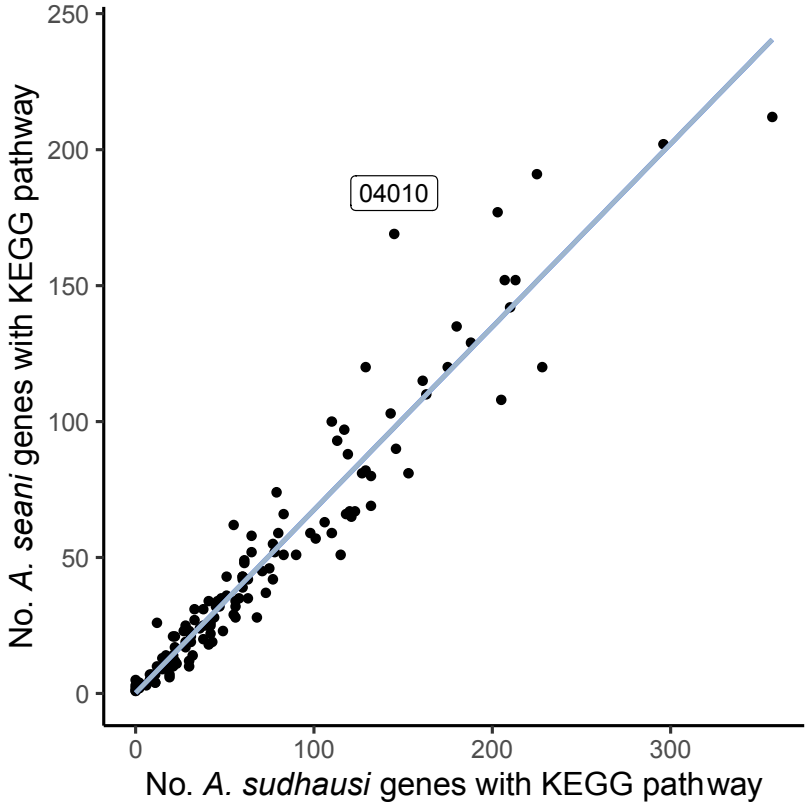


Fig. S3. A KEGG comparison showed an upregulation of the MAPK pathway (04010) in *A. seani* compared to *A. sudhausi*

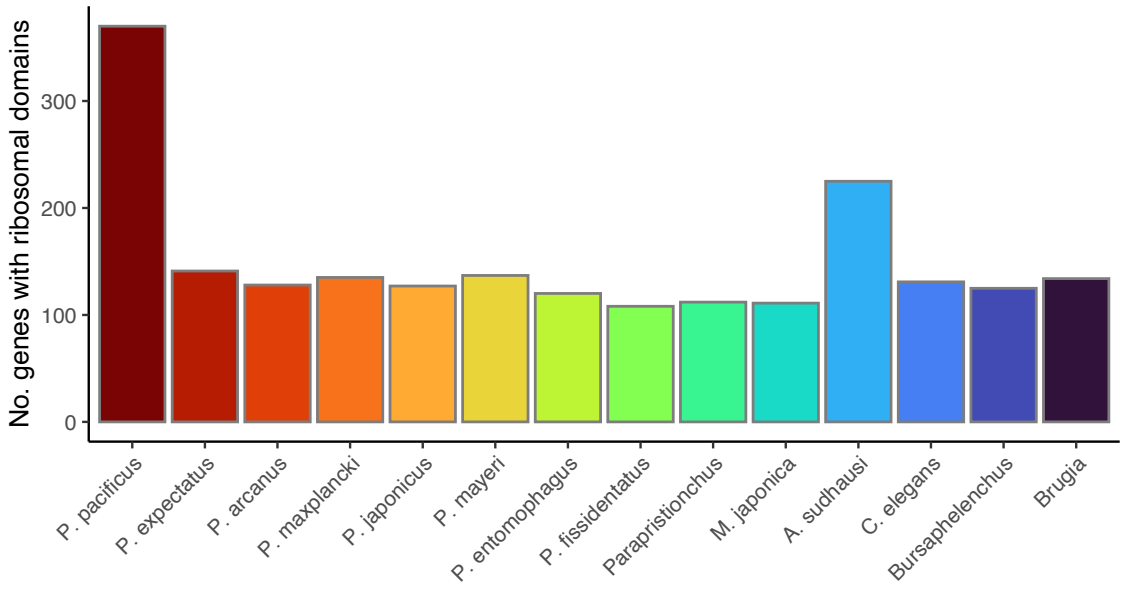


Fig. S4. The frequency of ribosomal Pfam domains across various nematode species, with *P. pacificus* showing particularly high representation.

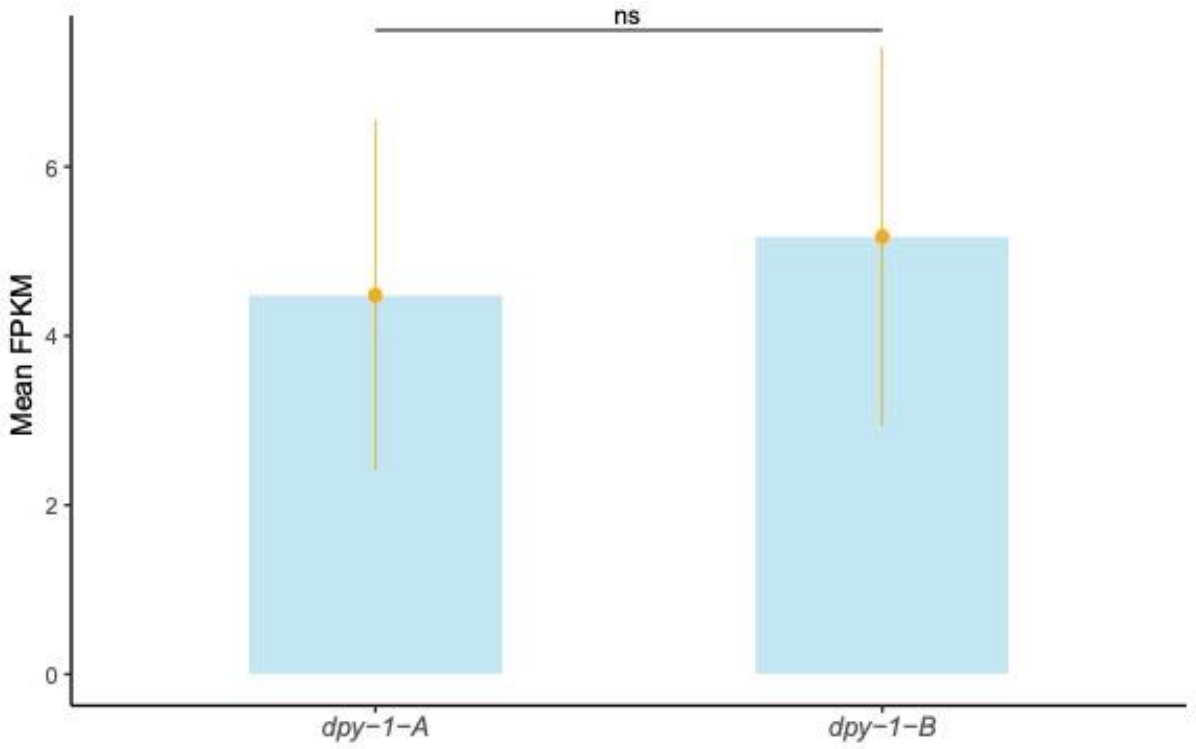


Fig. S5. The expression values of the *A. sudhausi dpy-1* ohnologs display no significant difference from one another (Welch Two Sample t-test, $p < 0.05$). Error bars indicate standard deviation.

Table S1. BUSCO analysis of the *A. sudhausi* and *A. seani* assemblies.

	A sudhausi PacBio Genome	*A.sudhausi* Illumina Genome	*A.sudhausi* Annotations	*A.sudhausi* Transcriptome	*A.seani* Transcriptome
Complete BUSCOs (C)	87.4%	82.7%	89.7%	90.1%	84.2%
Complete and single-copy BUSCOs (S)	38.7%	46.1%	9.9%	29.0%	36.2%
Complete and duplicated BUSCOs (D)	48.7%	36.6%	79.8%	61.1%	48.0%
Fragmented BUSCOs (F)	6.5%	9.0%	5.9%	5.9%	10.0%
Missing BUSCOs (M)	6.1%	8.3%	4.4%	4.0%	5.8%
Total BUSCO groups searched (n)	982	982	982	982	982

Table S2. Pfam Domains significantly more likely to be retained (i.e. found in ohnologs). Determined using the Fisher's 2x2 matrix.

Domain	**Accession**	**p value**	**FDR adj. p value**
Homeodomain	PF00046.31	0.012	0.0263023255813953
LRR_4	PF12799.9	0.039	0.0403928571428571
PDZ	PF00595.26	0.008	0.0263023255813953
zf-C2H2	PF00096.28	0.013	0.0263023255813953
zf-H2C2_2	PF13465.8	0.005	0.0263023255813953

Table S3. Pfam domains significantly more likely to be lost (i.e. found in single copies). Determined using the Fisher's 2x2 matrix

Domain	Accession	p value	FDR adj. p value
AA_permease	PF00324.23	0.038	0.0398313253012048
AAA_16	PF13191.8	0.041	0.0419647058823529
AAA_5	PF07728.16	0.031	0.0396455696202532
ACBP	PF00887.21	0.035	0.0396455696202532
Acyl-CoA_dh_M	PF02770.21	0.038	0.0398313253012048
adh_short	PF00106.27	0.023	0.0396455696202532
adh_short_C2	PF13561.8	0.007	0.0263023255813953
Aminotran_1_2	PF00155.23	0.005	0.0263023255813953
An_peroxidase	PF03098.17	0.028	0.0396455696202532
ATPgrasp_ST	PF14397.8	0.035	0.0396455696202532
ATPgrasp_YheCD	PF14398.8	0.035	0.0396455696202532
Beta_helix	PF13229.8	0.035	0.0396455696202532
BTB_2	PF02214.24	0.045	0.0455232558139535
Calreticulin	PF00262.20	0.035	0.0396455696202532
Calsequestrin	PF01216.19	0.035	0.0396455696202532
Carn_acyltransf	PF00755.22	0.008	0.0263023255813953
Cation_ATPase_N	PF00690.28	0.022	0.0396455696202532
Coatomer_WDAD	PF04053.16	0.036	0.0396455696202532
Diphthamide_syn	PF01866.19	0.035	0.0396455696202532
DO-GTPase2	PF19993.1	0.035	0.0396455696202532
DUF815	PF05673.15	0.022	0.0396455696202532
E1-E2_ATPase	PF00122.22	0.038	0.0398313253012048
EFG_C	PF00679.26	0.012	0.0263023255813953
EFG_III	PF14492.8	0.012	0.0263023255813953
EFG_IV	PF03764.20	0.005	0.0263023255813953
Ephrin_rec_like	PF07699.15	0.035	0.0396455696202532
Fer4_7	PF12838.9	0.035	0.0396455696202532
G-patch	PF01585.25	0.028	0.0396455696202532
GIDA	PF01134.24	0.002	0.0263023255813953
Glyco_transf_7C	PF02709.16	0.038	0.0398313253012048
HSP90	PF00183.20	0.035	0.0396455696202532
Hydrolase	PF00702.28	0.028	0.0396455696202532
Hydrolase_3	PF08282.14	0.035	0.0396455696202532
IMPDH	PF00478.27	0.035	0.0396455696202532
Ion_trans	PF00520.33	0.049	0.049
KIF1B	PF12423.10	0.035	0.0396455696202532
Kinesin	PF00225.25	0.003	0.0263023255813953
Kinesin_assoc	PF16183.7	0.035	0.0396455696202532

KR	PF08659.12	0.004	0.0263023255813953
MCM	PF00493.25	0.013	0.0263023255813953
MCM_lid	PF17855.3	0.002	0.0263023255813953
MCM_N	PF14551.8	0.002	0.0263023255813953
MCM_OB	PF17207.5	0.002	0.0263023255813953
Mg_chelatase	PF01078.23	0.028	0.0396455696202532
Microtub_bd	PF16796.7	0.015	0.0296590909090909
P_proprotein	PF01483.22	0.035	0.0396455696202532
Peptidase_M13	PF01431.23	0.028	0.0396455696202532
Peptidase_S8	PF00082.24	0.012	0.0263023255813953
PKD_channel	PF08016.14	0.036	0.0396455696202532
PRELI	PF04707.16	0.035	0.0396455696202532
Prot_ATP_ID_OB	PF16450.7	0.012	0.0263023255813953
Ricin_B_lectin	PF00652.24	0.036	0.0396455696202532
RVT_1	PF00078.29	0.035	0.0396455696202532
SCP2	PF02036.19	0.012	0.0263023255813953
SLC12	PF03522.17	0.035	0.0396455696202532
TatD_DNase	PF01026.23	0.035	0.0396455696202532
TPP_enzyme_C	PF02775.23	0.005	0.0263023255813953
TPP_enzyme_M	PF00205.24	0.035	0.0396455696202532
XPG_I	PF00867.20	0.035	0.0396455696202532
Zeta_toxin	PF06414.14	0.035	0.0396455696202532
4F5	PF04419.16	0.001	0.0263023255813953
Bravo_FIGEY	PF13882.8	0.013	0.0263023255813953
CERK_C	PF19280.1	0.013	0.0263023255813953
COX6B	PF02297.19	0.013	0.0263023255813953
DHO_dh	PF01180.23	0.013	0.0263023255813953
DUF4139	PF13598.8	0.013	0.0263023255813953
DUF4140	PF13600.8	0.013	0.0263023255813953
EF_TS	PF00889.21	0.013	0.0263023255813953
FAD_binding_4	PF01565.25	0.001	0.0263023255813953
Fer4_21	PF14697.8	0.013	0.0263023255813953
Gln-synt_C	PF00120.26	0.013	0.0263023255813953
Glyco_hydro_15	PF00723.23	0.013	0.0263023255813953
Indigoidine_A	PF04227.14	0.013	0.0263023255813953
KASH_CCD	PF14662.8	0.013	0.0263023255813953
KPBB_C	PF19292.1	0.013	0.0263023255813953
Lyase_1	PF00206.22	0.013	0.0263023255813953
P-mevalo_kinase	PF04275.16	0.013	0.0263023255813953
PTPlike_phytase	PF14566.8	0.013	0.0263023255813953
Ribosomal_L13	PF00572.20	0.013	0.0263023255813953
SRCR	PF00530.20	0.013	0.0263023255813953

TPP_enzyme_N	PF02776.20	0.013	0.0263023255813953
XPG_N	PF00752.19	0.013	0.0263023255813953

Table S4. KEGG pathways significantly more likely to be retained (i.e. found in ohnologs). Determined using the Fisher's 2x2 matrix.

KEGG Pathway	KEGG Description	p value	FDR adj. p value
03018	RNA degradation	0.025	0.0312
03022	Basal transcription factors	0.026	0.0312
04120	Ubiquitin mediated proteolysis	0.005	0.024
04144	Endocytosis	0.010	0.03
04150	mTOR signaling pathway	0.006	0.024

Table S5. KEGG pathways significantly more likely to be lost (i.e. found in single copies). Determined using the Fisher's 2x2 matrix.

KEGG Pathway	KEGG Description	p value	FDR adj. p value
00030	Pentose phosphate pathway	0.029	0.0316363636363636
00190	Oxidative phosphorylation	0.006	0.024
00220	Arginine biosynthesis	0.026	0.0312
00620	Pyruvate metabolism	0.014	0.0312
00630	Glyoxylate and dicarboxylate metabolism	0.020	0.0312
00640	Propanoate metabolism	0.046	0.046
04146	Peroxisome	0.023	0.0312

Table S6. List of Pfam domains predicted for ohnologs with the top 5% highest dN/dS values

Domain	Accession	Frequency
BBIP10	PF14777.8	1
CAP	PF00188.28	1
Chromo	PF00385.26	1
DUF4256	PF14066.8	1
ECH_1	PF00378.22	1
GBP	PF02263.21	1
MLANA	PF14991.8	1
Sod_Cu	PF00080.22	1
TTR-52	PF01060.25	1

Table S7: Table showing the top 20 ohnolog pairs with the highest dN/dS values (minimum dS ≥ 0.01). The corresponding expression values and absolute log2(FC) values are also shown. No *C. elegans* orthologs were found for these ohnologs.

Orthogroup	Ohnolog 1	Ohnolog 2	dN	dS	dN/dS	Mean FPKM Ohnolog 1	Mean FPKM Ohnolog 2	Absolute log2(FC)	*C.elegans* ortholog
OG0017566	ALDISUDHAUS000007547	ALDISUDHAUS000007599	0,10	0,01	9,32	2,14	4,44	1,05	NA
OG0019875	ALDISUDHAUS000031914	ALDISUDHAUS000031976	0,19	0,02	7,88	0,00	0,11	Inf	NA
OG0020086	ALDISUDHAUS000032896	ALDISUDHAUS000039664	0,08	0,01	7,79	1,99	1,96	0,03	NA
OG0019055	ALDISUDHAUS000026520	ALDISUDHAUS000030560	0,11	0,01	7,23	0,52	0,10	2,43	NA
OG0018831	ALDISUDHAUS000024096	ALDISUDHAUS000027749	0,09	0,01	7,02	0,77	1,38	0,85	NA
OG0020112	ALDISUDHAUS000033025	ALDISUDHAUS000033311	0,09	0,01	6,96	2,18	1,57	0,48	NA
OG0020269	ALDISUDHAUS000033887	ALDISUDHAUS000035022	0,15	0,02	6,85	2,33	0,31	2,91	NA
OG0018255	ALDISUDHAUS000019364	ALDISUDHAUS000036621	0,07	0,01	6,71	19,63	19,26	0,03	NA
OG0018894	ALDISUDHAUS000024662	ALDISUDHAUS000024950	0,09	0,01	6,34	2,39	1,04	1,20	NA
OG0019818	ALDISUDHAUS000031678	ALDISUDHAUS000039109	0,26	0,04	6,23	7,27	0,26	4,81	NA
OG0020291	ALDISUDHAUS000034016	ALDISUDHAUS000039894	0,08	0,01	6,22	0,13	0,46	1,78	NA
OG0014260	ALDISUDHAUS000022716	ALDISUDHAUS000027390	0,06	0,01	6,08	3562,42	3119,11	0,19	NA
OG0014326	ALDISUDHAUS000024525	ALDISUDHAUS000025343	0,08	0,01	5,95	105,12	119,19	0,18	NA
OG0019209	ALDISUDHAUS000028500	ALDISUDHAUS000028726	0,06	0,01	5,81	0,99	0,89	0,16	NA
OG0020125	ALDISUDHAUS000033105	ALDISUDHAUS000037123	0,07	0,01	5,76	0,50	0,59	0,25	NA
OG0017584	ALDISUDHAUS000007902	ALDISUDHAUS000010320	0,07	0,01	5,71	7,58	5,80	0,38	NA
OG0014673	ALDISUDHAUS000034782	ALDISUDHAUS000038967	0,12	0,02	5,54	1,00	0,58	0,79	NA
OG0019969	ALDISUDHAUS000032319	ALDISUDHAUS000033823	0,07	0,01	5,44	3,50	1,12	1,64	NA
OG0020076	ALDISUDHAUS000032839	ALDISUDHAUS000036366	0,06	0,01	5,38	0,00	0,82	Inf	NA
OG0018895	ALDISUDHAUS000024683	ALDISUDHAUS000028406	0,06	0,01	5,31	1,57	1,58	0,01	NA

Table S8: Table showing the top 20 ohnolog pairs with the highest differential expression differences. The top log2(FC) values are all infinite due to one ohnolog copy having zero expression. The dN, dS and dN/dS values are also shown, as well as the *C. elegans* orthologs.

Orthogroup	**Ohnolog 1**	**Ohnolog 2**	**dN**	**dS**	**dN/dS**	**Mean FPKM Ohnolog 1**	**Mean FPKM Ohnolog 2**	**Absolute log2(FC)**	***C. elegans* ortholog**
OG0010977	ALDISUDHAUS000025105	ALDISUDHAUS000027375	0,00	0,10	0,03	1003,90	0	Inf	*C26F1,4 (rps-30)*
OG0014240	ALDISUDHAUS000022219	ALDISUDHAUS000026032	0,04	0,01	3,99	308,97	0	Inf	
OG0018098	ALDISUDHAUS000017139	ALDISUDHAUS000018536	0,01	0,25	0,06	298,47	0	Inf	
OG0020253	ALDISUDHAUS000033770	ALDISUDHAUS000037970	0,03	0,17	0,18	164,75	0	Inf	
OG0017709	ALDISUDHAUS000010158	ALDISUDHAUS000022172	0,02	0,12	0,19	66,55	0	Inf	
OG0006834	ALDISUDHAUS000021463	ALDISUDHAUS000021814	0,01	0,12	0,05	59,86	0	Inf	*K10C3,2 (ensa-1)*
OG0004688	ALDISUDHAUS000006992	ALDISUDHAUS000008228	0,00	0,09	0,00	57,63	0	Inf	*R13A5,8 (rpl-9)*
OG0017689	ALDISUDHAUS000009667	ALDISUDHAUS000023377	0,06	0,07	0,79	43,59	0	Inf	
OG0009384	ALDISUDHAUS000029847	ALDISUDHAUS000031626	0,00	0,17	0,00	38,64	0	Inf	Y41E3,8a
OG0007269	ALDISUDHAUS000033948	ALDISUDHAUS000039075	0,00	0,14	0,03	32,29	0	Inf	*Y37D8A,14 (cox-5A)*
OG0017562	ALDISUDHAUS000007512	ALDISUDHAUS000019487	0,59	1,29	0,46	29,01	0	Inf	
OG0019745	ALDISUDHAUS000031316	ALDISUDHAUS000033468	0,03	0,08	0,35	27,88	0	Inf	
OG0019571	ALDISUDHAUS000030628	ALDISUDHAUS000033858	0,21	0,55	0,39	26,00	0	Inf	
OG0020406	ALDISUDHAUS000034697	ALDISUDHAUS000036548	0,02	0,16	0,11	23,04	0	Inf	
OG0007582	ALDISUDHAUS000032431	ALDISUDHAUS000037466	0,00	0,23	0,02	20,99	0	Inf	*C32E8,3 (tppp-1)*
OG0020192	ALDISUDHAUS000033511	ALDISUDHAUS000040026	0,02	0,03	0,68	20,75	0	Inf	
OG0020603	ALDISUDHAUS000036095	ALDISUDHAUS000038387	0,07	0,04	1,66	20,30	0	Inf	
OG0019197	ALDISUDHAUS000028356	ALDISUDHAUS000038200	0,02	0,10	0,25	19,51	0	Inf	
OG0006166	ALDISUDHAUS000007646	ALDISUDHAUS000008647	0,00	0,16	0,00	18,34	0	Inf	*F39B2,2 (uev-1)*
OG0017753	ALDISUDHAUS000011072	ALDISUDHAUS000025136	0,32	0,15	2,17	18,17	0	Inf	

Table S9. Successful CRISPR mutants with their exact mutations listed. Most mutants came from the same injected P0 worm A. Two double mutants were obtained from one injection. Both these mutants had a dumpy phenotype, despite being in-frame. The mutations in their sequence are listed, with the parts containing the gRNA sequence underlined. The majority of mutations were deletions. The strain and allele names of the lines that were frozen is listed.

No.	Injected P0	Mutation		Dumpy phenotype?	Mutation in Sequence		Strain	Allele
		dpy-1-A	*dpy-1-B*		*dpy-1-A*	*dpy-1-B*		
1	A		6bp del	×		GAGAACT------GTGG		
2		6bp indel (14bp ins+8bp del)		×	AAGAG--------**GGAATCATCAGATT**CTGTGG		RS3704	tu1470
3			3bp del	✓		GAGAACTC---CTGTGG		
4		**5bp del**	**7bp ins**	✓	GAACTCCAG-----GCTCATCCG	GAACTCCAG**AAGAGAA**CTGTGG	RS3702	tu1468
5			**7bp ins**	×			RS3705	tu1471
6		**5bp del**		×	GAACTCCAG-----GCTCATCCG		RS3703	tu1469
7			12bp del	×		GAGAACTC----------ATCCGAGT		
8	B	6bp del		×	GAGAACT------GTGG			
9	C	6bp del		×	GAGAA------CTGTGG			

PLOS GENETICS

RESEARCH ARTICLE

Thousands of *Pristionchus pacificus* orphan genes were integrated into developmental networks that respond to diverse environmental microbiota

Marina Athanasouli, Nermin Akduman¤, Waltraud Röseler, Penghieng Theam, Christian Rödelsperger*

Department for Integrative Evolutionary Biology, Max Planck Institute for Biology, Tübingen, Germany

¤ Current address: Interfaculty Institute of Microbiology and Infection Medicine, University of Tübingen, Tübingen, Germany
* christian.roedelsperger@tuebingen.mpg.de

Citation: Athanasouli M, Akduman N, Röseler W, Theam P, Rödelsperger C (2023) Thousands of *Pristionchus pacificus* orphan genes were integrated into developmental networks that respond to diverse environmental microbiota. PLoS Genet 19(7): e1010832. https://doi.org/10.1371/journal.pgen.1010832

Editor: Kaveh Ashrafi, University of California San Francisco, UNITED STATES

Received: March 2, 2023

Accepted: June 15, 2023

Published: July 3, 2023

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pgen.1010832

Data Availability Statement: Raw reads were deposited at the European Nucleotide archive under the study accession PRJEB60166.

Abstract

Adaptation of organisms to environmental change may be facilitated by the creation of new genes. New genes without homologs in other lineages are known as taxonomically-restricted orphan genes and may result from divergence or *de novo* formation. Previously, we have extensively characterized the evolution and origin of such orphan genes in the nematode model organism *Pristionchus pacificus*. Here, we employ large-scale transcriptomics to establish potential functional associations and to measure the degree of transcriptional plasticity among orphan genes. Specifically, we analyzed 24 RNA-seq samples from adult *P. pacificus* worms raised on 24 different monoxenic bacterial cultures. Based on coexpression analysis, we identified 28 large modules that harbor 3,727 diplogastrid-specific orphan genes and that respond dynamically to different bacteria. These coexpression modules have distinct regulatory architecture and also exhibit differential expression patterns across development suggesting a link between bacterial response networks and development. Phylostratigraphy revealed a considerably high number of family- and even species-specific orphan genes in certain coexpression modules. This suggests that new genes are not attached randomly to existing cellular networks and that integration can happen very fast. Integrative analysis of protein domains, gene expression and ortholog data facilitated the assignments of biological labels for 22 coexpression modules with one of the largest, fast-evolving module being associated with spermatogenesis. In summary, this work presents the first functional annotation for thousands of *P. pacificus* orphan genes and reveals insights into their integration into environmentally responsive gene networks.

Author summary

The inference of biological function for genes in newly sequenced genomes heavily relies on sequence conservation with classical model organisms. Consequently, newly evolved

Funding: This work was funded by the Max Planck Society. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

orphan genes that do not have homologs will not be associated with any function. Here, we use coexpression with known genes in order to assign potential functions to orphan genes in the nematode *Pristionchus pacificus*. To this end, we generated transcriptome profiles of *P. pacificus* worms that were grown on 24 different bacteria and clustered the genes into 28 large coexpression modules which contain thousands of orphan genes. Integrative analysis could associate most coexpression modules with biological processes or tissues, which results in the first functional annotation for thousands of orphan genes. Complementary analysis of gene ages shows that modules associated with female reproduction are highly constrained whereas a male-reproductive module is much more likely to integrate new genes, which could possibly be explained by sperm competition. This links sexual conflict with gene network evolution and environmental regulation.

Introduction

The evolution of phenotypic diversity across all domains of life has been accompanied by the formation of new genes. As a result, up to one third of the gene content in extant genomes consists of orphan genes that lack homologs in other taxonomic lineages [1]. Although the definition of these taxonomically-restricted orphan genes is context-dependent and the numbers of identified orphan genes will vary with the phylogenetic resolution and methods for homology detection, orphan genes exist in any given genome [2]. Orphan genes may reflect ancient genes that evolve so fast that homology cannot be detected [3]. Alternatively, orphan genes may arise *de novo* from previously non-coding sequences [4,5]. This has been demonstrated in multiple taxonomic groups including vertebrates [6,7], insects [8,9], nematodes [10,11], yeast [12,13], and plants [14,15]. Horizontal gene transfer can be considered as a third mechanism that would lead to the generation of orphan genes [16,17]. The origin of orphan genes can most conclusively be studied for relatively recent gene births because potentially neutrally evolving ancestral sequences may degenerate very fast in sister taxa rendering homology detection impossible [18]. Thus, to what extent more ancient orphan genes arose from divergence or *de novo* formation is currently not known. Another important question concerns the biological functions of orphan genes. Given that large-scale experimental screens for possible functions are only possible in a limited number of model species [19,20] and that inference of function based on sequence conservation is not possible, most orphan genes have no known function. It has been hypothesized that new genes play a role in adapting to changing environments [21], but this is only supported by limited data [22,23]. In this study, we want to test this hypothesis in the nematode model organism *Pristionchus pacificus* by exploring the transcriptomic changes after exposure to different bacteria. *P. pacificus* is a free-living nematode that was initially established as a model system for comparative studies with the classical model organism *Caenorhabditis elegans* [24]. *C. elegans* and *P. pacificus* were estimated to have shared the last common ancestor around 130–310 million years ago [25]. Numerous studies have revealed conserved and divergent patterns across development [26], neurobiology [27], and behavior [28]. When the genome of *P. pacificus* was sequenced, around one third of its genes were classified as orphan genes that do not have any homolog outside the diplogastrid family [29]. By sequencing nine additional diplogastrid genomes, we established a phylogenomic framework to assign these orphan genes into phylostrata and to study their evolutionary dynamics and origin at the inter- and intra-species level [30,31]. These studies confirmed trends such as little expression evidence of orphan genes, rapid evolution, and high turnover which were also found in genomic studies of other animals and plants [32,33]. In contrast to

the broad understanding of new gene evolution in *Pristionchus* nematodes, only two orphan genes have been experimentally characterized. The first orphan gene, *dauerless*, was identified in a screen to dissect the genetic basis of natural variation in dauer formation [34]. The dauer stage represents an alternative developmental stage in most nematodes that allows the worms to survive unfavorable environmental conditions for extended time periods. Overexpression of *dauerless* results in the suppression of dauer formation and it has been speculated that the ability to regulate *dauerless* dosage might provide individual strains an advantage during intraspecific competition [34]. The second orphan gene, *self-1*, encodes a micropeptide that was identified in a screen for killing behavior. Knockouts or modifications of *self-1* resulted in a loss of self-recognition and led to killing by relatives. These two cases demonstrate that *Pristionchus* orphan genes are involved in important developmental decisions and in the evolution of novel behaviors. To complement this detailed functional knowledge of a very limited number of orphan genes with broader functional data on a genome-wide scale, we employ expression data as a proxy for function. This is under the assumption that in order to carry out a certain function, a gene must be expressed at a given condition. Furthermore, large-scale expression data can be used to group genes into functional modules and to transfer functional annotations based on coexpression [35–37]. Specifically, we aim to investigate the *P. pacificus* transcriptome in response to different microbiota which denote the assemblage of microorganisms present in a defined environment [38]. Environmental bacteria can interact with nematodes in a variety of different ways such as serving as food source [39], constituents of the gut microbiome [40], or pathogens [41]. Here, we focus on the transcriptomic response to 24 different bacteria, most of which were isolated previously from *Pristionchus*-associated environments [42]. These bacteria are non-pathogenic in the sense that worm populations can survive for at least several days [42]. This is in contrast to highly pathogenic strains that can kill complete worm populations within a few hours [41]. Our main goals are to group *P. pacificus* genes into functional modules that respond differently to environmental microbiota, to characterize these modules, and finally to test whether some of these modules are enriched in orphan genes. This will allow us to better understand the plastic response to diverse environmental microbiota and to further elucidate the regulation and evolution of the associated gene networks.

Results

The transcriptomic response of nematodes to various microbiota does not strictly reflect bacterial phylogeny

To investigate the transcriptomic response of *P. pacificus* to different environmental microbiota, we grew worms on monoxenic cultures of 24 bacterial strains that included commonly used food bacteria such as *Escherichicha coli* OP50 and HB101 as well as 22 bacterial strains that were previously isolated from *Pristionchus*-associated environments [42]. The selected bacteria represent Alpha-, Beta-, Gammaproteobacteria as well as Flavobacteriia. A single RNA-seq data was generated per bacterial strain from 50 young adult worms that were manually picked from mixed-stage cultures (see *Methods*). Most transcriptome profiles of worms grown on different bacteria appeared to be highly similar (Pearson $r > 0.9$, Fig 1A). One sample, *Wautersiella* LRB104, showed a quite distinct transcriptome with correlation coefficients around 0.8 (Fig 1A and 1B). However, even this outlier shows much higher correlations than transcriptomes from different developmental stages with correlation coefficients of 0.6 [43,44]. The observed correlation coefficients translate into hundreds to thousands of genes with an absolute fold change > 2 between the most similar and most dissimilar pair of samples (S1 Fig). Next, we wanted to test whether the transcriptome profiles follow a phylogenetic pattern,

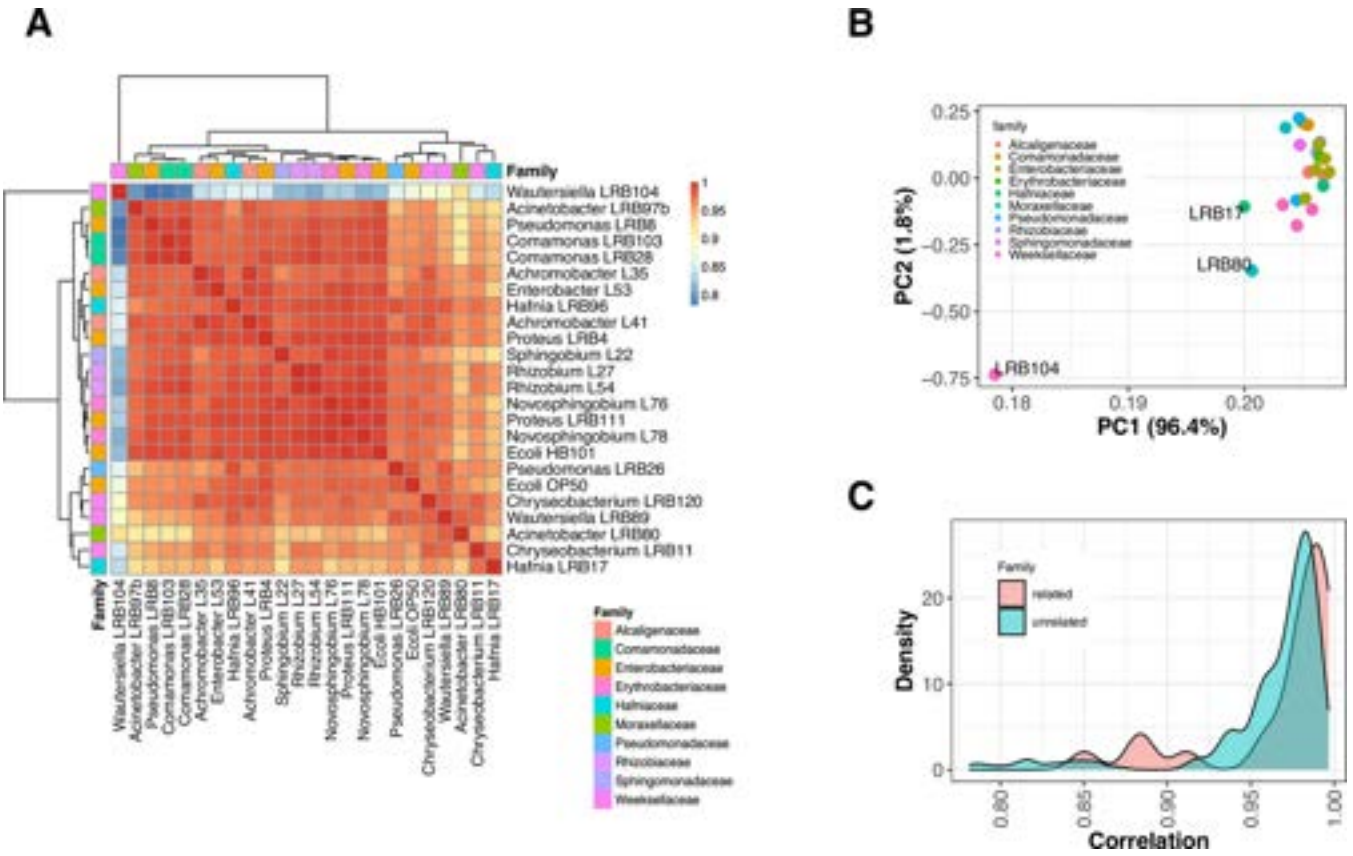


Fig 1. Transcriptional response to 24 bacterial environments. (A) The heatmap shows the correlation between transcriptomes. Apart from *Wautersiella* LRB104 all transcriptomes are highly similar. (B) Complementary analysis using principal component analysis identifies *Wautersiella* LRB104 as the most distinct environment. C) The histograms show the distribution of correlation values for comparisons within and across bacterial families. The expression profiles of *P. pacificus* worms on bacteria of different families can be more similar than the profile on bacteria of the same family. This suggests that the transcriptional response does not strictly reflect bacterial phylogeny.

https://doi.org/10.1371/journal.pgen.1010832.g001

i.e. are transcriptome profiles of worms grown on bacteria of the same family more similar than nematode transcriptomes from more distantly related bacteria (Fig 1C). This did not show significantly higher correlation for transcriptomes from more closely related bacteria (P = 0.35, t-test) suggesting that the transcriptomic response of nematodes to various environmental microbiota do not strictly reflect phylogeny. This observation could indicate that the observed differences are driven by factors that are difficult to control (e.g. bacterial concentration), or alternatively, that strain-specific changes are obscuring family-specific signals. The latter would be the case if critical metabolic pathways are plasmid encoded and could easily be horizontally transferred. In summary, *P. pacificus* nematodes exhibit substantial transcriptomic variation in response to environmental microbiota and these responses do not strictly reflect the bacterial phylogeny.

Almost half of all genes respond to diverse environmental microbiota

Nematodes interact with bacteria in many different ways involving processes like chemical attraction and repulsion [45], digestion [39], and detoxification [46]. Moreover, those processes can trigger secondary effects on worm development and physiology [47]. Thus, the response of worms might involve multiple regulatory and metabolic pathways and might impact many other biological processes. To characterize gene networks that respond to diverse

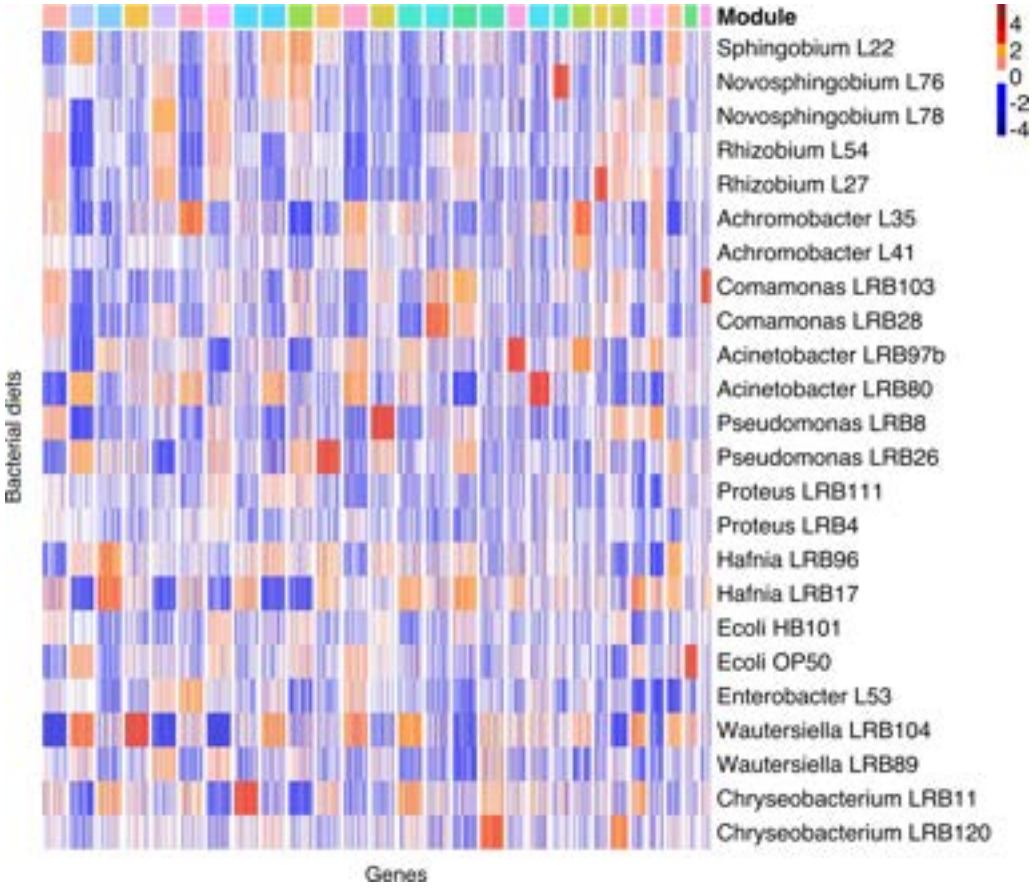


Fig 2. Expression level of environmentally responsive coexpression modules. We visualized the z-score normalized expression levels for the 28 largest modules across the bacterial environments. Modules with more than 100 genes were randomly downsampled to 100 genes. While genes of module 4 are most strongly expressed on *Wautersiella* LRB104, module 3 shows the highest expression in response to the two *Hafnia* strains. This demonstrates that environmental microbiota can modulate specific coexpression modules.

https://doi.org/10.1371/journal.pgen.1010832.g002

environmental microbiota, we computed coexpression modules from the RNA-seq data using a widely used graph clustering approach [48]. Based on correlation coefficient of 0.7 (S2 Fig), we identified 28 large coexpression modules with more than 50 genes (Fig 2 and S1 Table). The largest modules exhibit relatively drastic changes on only one or few bacteria (mostly *Wautersiella* LRB104 and *Hafnia* LRB17), but also more subtle differences on other bacteria. For example, genes of module 1 show strongly reduced expression on *Wautersiella* LRB104, but also mildly lower expression on some other bacteria. These bacteria yield largely opposite trends for genes in module 2. Module 3 shows highest expression on the *Hafnia* strains and the strongest pattern of module 4 results again from high expression on *Wautersiella* LRB104. Also, module 5 shows weak expression on *Wautersiella* LRB104, *Pseudomonas* LRB26, and *Hafnia* LRB17, whereas module 6 stands out by having high expression on *Achromobacter* L35. Given that the most extreme expression differences are frequently observed in the transcriptomic response to *Wautersiella* LRB104, we compared developmental timing between *E. coli* OP50 and both *Wautersiella* strains (LRB89 and LRB104). This demonstrated that worms exhibit a developmental delay on both *Wautersiella* bacteria relative to *E. coli* OP50 (S3 Fig). Given that both *Wautersiella* strains change the developmental timing of *P. pacificus* worms, we cannot fully explain why the transcriptomic response to *Wautersiella LRB104* differs so

starkly from all the other transcriptomes. To test, how strongly this outlier influences the structure of the coexpression network and the subsequent analysis, we recomputed coexpression modules after removing the *Wautersiella* LRB104 data and compared coexpression modules across both data sets (S4 Fig). This showed a clear one-to-one correspondence between most coexpression modules indicating that the structure of the coexpression network is largely robust. In addition, systematic analysis of coexpression networks that were computed from subsampled data revealed that with fewer RNA-seq samples, network modules tend to get larger (S5 Fig). This may be due to two reasons. First, with fewer samples, it is easier to exceed a given correlation threshold just by chance. Second, with higher numbers of RNA-seq samples, expression profiles can only become more complex. This will cause splits of larger modules. Thus, the full data set with all 24 RNA-seq samples yields the most conservative lower estimate of 14,275 for the number of environmentally responsive genes in large coexpression modules. That means that almost half of the 28,896 annotated genes in *P. pacificus* respond to environmental microbiota.

Coexpression modules exhibit developmental signatures

Bacterial diets can alter the developmental rate in *C. elegans* and *P. pacificus* [39,47]. Therefore, we wanted to test whether the coexpression modules also show a developmental signature. To this end, we visualized the normalized expression across *P. pacificus* development [43] for all genes in the largest coexpression modules (Fig 3). This showed that the coexpression modules that were obtained from adult worms on different bacteria also exhibit distinct expression profiles during development. Furthermore, this developmental signature is very consistent between most genes of the same module. For example, genes in module 1 are consistently activated only late in development (>48h, Fig 3). Their expression is preceded by genes in module 2 which consistently increase expression starting from the 40h timepoint (Fig 3). This could mean that the same network modules control development as well as the response to environmental microbiota. Alternatively, this developmental signature could be an indirect effect of altered developmental timing. Even if only adult worms were manually picked for RNA extraction, there might still be differences in the age of these worms. This is because worms may delay or accelerate their development in response to different bacteria [39,47] and consequently, adult worms that were grown on different bacteria may not be of the same chronological age. If the expression of certain modules rather follows the chronological age than the morphological stage, such modules could potentially cause a similar developmental signature. However, no matter if the expression profiles represent immediate response to different bacteria or variation in the chronological age of worms, both scenarios would reflect either direct or indirect consequences of the exposure to different microbial environments.

Coexpression modules have distinct regulatory architecture

The observation of distinct expression profiles among environmentally responsive genes suggests that the coexpression modules might be coregulated by diverse sets of transcription factors. To test this, we searched for overrepresented motifs in the promoter sequences of the largest coexpression modules using a *de novo* motif discovery approach as implemented in the HOMER software [49]. This identified a diverse set of DNA motifs that are highly enriched in specific modules (S6 Fig). While modules 1,5,7, and 23 show enrichments of the same motifs (ZBTB32, CUX1, and LIN54), many other coexpression modules have a unique regulatory architecture with very specific motifs that are only enriched in the given module (e.g. POU5F1 in module 2 and *Foxd3* in module 24, S6 Fig). Unfortunately, it is not straightforward to infer the regulator for a given motif as families of transcription factors might be large and may bind

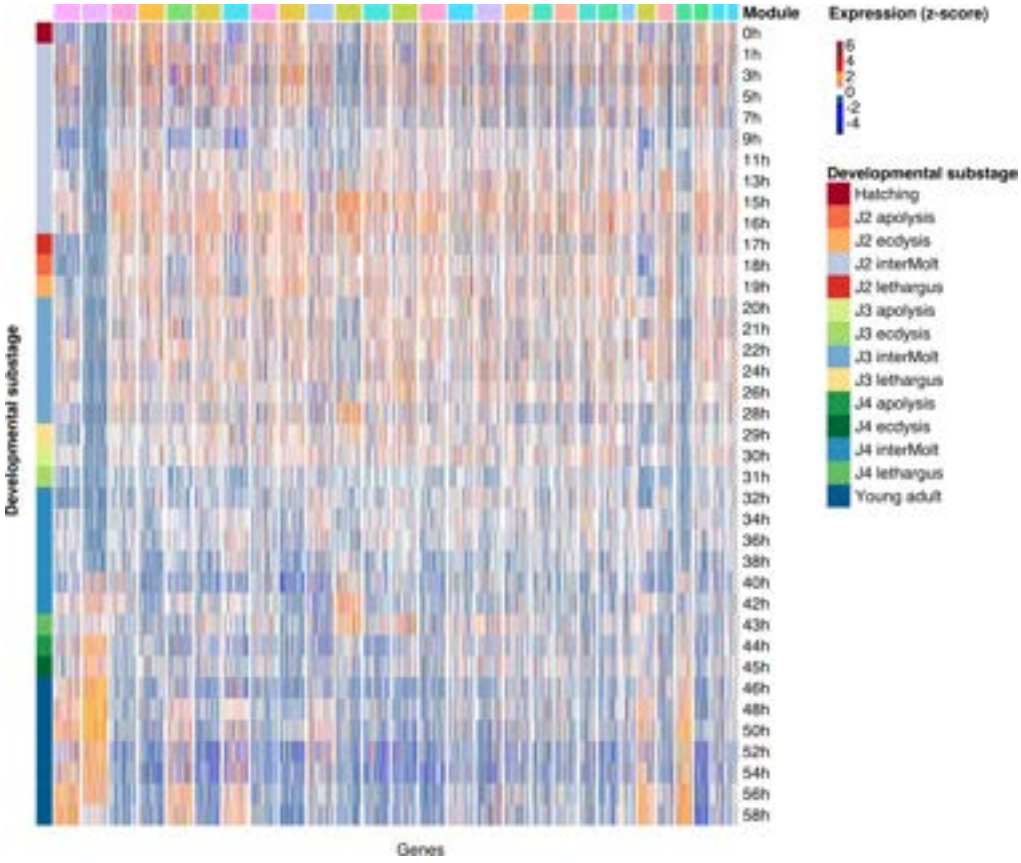


Fig 3. Developmental signature of coexpression modules. We visualized the z-score normalized expression levels of genes in the 28 largest coexpression modules throughout postembryonic development after hatching (0h) on *E. coli* OP50 [43]. Modules with more than 100 genes were randomly downsampled to 100 genes. The coexpression modules show distinct expression profiles suggesting a link between environmental response and developmental regulation.

https://doi.org/10.1371/journal.pgen.1010832.g003

very similar motifs [50]. For example, the human zinc finger and BTB domain containing protein ZBTB32 which binds a motif that is highly enriched in module 1 (S6 Fig) has dozens of orthologs in the genomes of *C. elegans* and *P. pacificus*. However, the most significantly enriched motif of module 2 has the highest similarity with the motif of POU domain homeobox transcription factor POU5F1. This family has only three members (*unc-86*, *ceh-6* and *ceh-18*.) in *C. elegans* [51] and four in *P. pacificus*. Similarly, *C. elegans ces-1* has a one-to-one ortholog in *P. pacificus*. Thus, for individual modules future experimental analysis could be used to dissect the regulatory relationships at a mechanistic level. Nevertheless, already the diversity of motifs that are found across different coexpression modules supports that they have distinct regulatory architecture. In addition, this analysis supports that the genes in a given module are not only coexpressed, but also coregulated.

3,727 *Pristionchus pacificus* orphan genes respond to diverse environmental microbiota

To test to what extent new genes could contribute to the response to diverse environmental microbiota, we performed a phylostratigraphic analysis to determine the distribution of gene ages across coexpression modules. Phylostratigraphy is a commonly used method to map gene birth events to a branch within a species tree by searching for the most distant homolog [52].

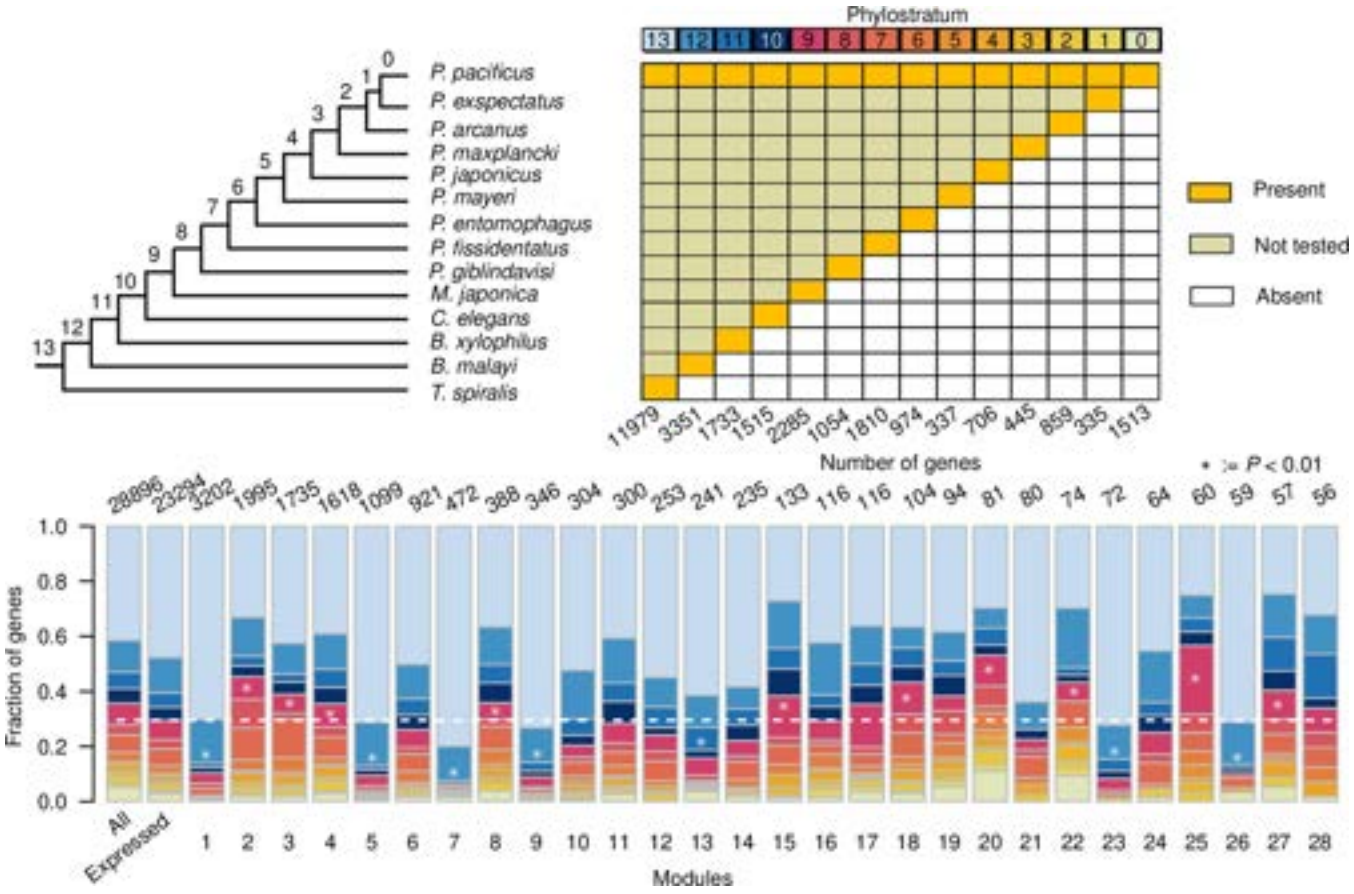


Fig 4. Phylostratigraphic analysis across coexpression modules. Phylostrata were defined based on the presence of homologs for a given gene in the most distantly related species. Each phylostratum defines a branch in the phylogeny where a gene likely originated. The barplot shows the distribution of phylostrata across coexpression modules with the dashed line marking the fraction of diplogastrid-specific orphan genes across all genes. The stars indicate significant enrichment and depletion of orphan genes based on simulating the integration of new genes to existing network modules ($P < 0.01$, S7 Fig).

https://doi.org/10.1371/journal.pgen.1010832.g004

Here, we selected nine additional diplogastrid genomes [30] and the genomes of *C. elegans*, *Bursaphelenchelus xylophilus* [53], *Brugia malayi* [54], and *Trichinella spiralis* [55] to assign genes to phylostrata (S2 Table). Altogether, 10,318 (35.7%) of all genes were defined as diplogastrid family-specific orphan genes (BLASTP e-value $< 10^{-3}$). Visualization of the distribution of phylostrata across the coexpression modules shows that almost half of all the genes in module 2 are diplogastrid-specific orphan genes (Fig 4). Furthermore, among the smaller modules 20 and 22, we observe relatively high ratios of species-specific orphan genes. This finding suggests that the integration of new genes into regulatory networks can happen very rapidly. Alternatively, such genes might represent ancient but rapidly evolving sequences such as antimicrobial peptides where we underestimate the gene age due to the failure to detect homologs [3]. In total, 3,727 diplogastrid-specific orphan genes are found among the 28 largest coexpression modules. As indicated above, this represents a lower estimate for the total number of environmentally-responsive orphan genes as additional orphan genes are found in smaller modules (S4 Fig). Relative to the genome-wide fraction of orphan genes (35.7%), this represents an underrepresentation as only 26.1% of genes in large coexpression modules are orphan genes ($P < 2.2 \times 10^{-16}$, Fisher's exact test). This may be largely explained by the fact that orphan genes are generally lowly expressed [31]. In our data set, only 66.4% of orphan genes are expressed whereas this number increases to 88.5% for non-orphan genes.

Nevertheless, despite their lower level of expression, our study could demonstrate that thousands of orphan genes are indeed embedded into developmental networks that plastically respond to diverse microbiota. Further, it identified which of the orphan genes respond to different environments and it established associations between these orphan genes and specific network modules where they have been embedded.

The major fast-evolving module is associated with spermatogenesis

The distribution of phylostrata across coexpression modules suggests that new genes are not attached randomly into regulatory networks, i.e. not every module is equally likely to acquire a new gene. On the contrary, individual coexpression modules are more likely to integrate new genes than others. To assess the non-randomness in the distribution of diplogastrid-specific orphan genes, we simulated the integration of new genes into existing networks. Specifically, we tested a model where all network modules are equally likely to acquire a new gene and a model where the probability of attachment is proportional to the size of the module (S7 Fig). Basically, the second model predicts much more accurately the number of orphan genes per module and it allows to define fast evolving modules with significant enrichments of orphan genes (Modules 2–4, 8, 15, 18, 20, 22, 25, and 27, $P < 0.01$, S7 Fig and Fig 4). Constrained modules (Modules 1, 5, 7, 9, 13, 23, and 26) were defined analogously as modules that are significantly depleted in orphan genes. To better characterize the biological differences between such fast evolving and more constrained modules, we performed overrepresentation analysis of protein domains (S3 Table), metabolic pathways (S4 Table), and other expression data sets [43,56–58]. The expression data includes 11 sets of regionally expressed genes that were identified from RNA tomography along the anterior-posterior axis of individual worms [58]. These analyses were complemented by tissue enrichment analysis (TEA, S5 Table) using one-to-one orthologs in *C. elegans* [59]. The overrepresentation of germline associated regions P6-P8 and TEA support that the constrained module 1 is associated with oogenesis (Fig 5A and S5 Table). In contrast, the enrichment of Motile sperm proteins, TEA and expression in sperm-related regions (P5, P9) associate the fast evolving module 2 with spermatogenesis (Figs S8 and 5A and S5 Table). Thus, the two largest environmentally responsive coexpression modules represent a constrained module associated with oogenesis and a fast evolving module that is associated with spermatogenesis. In *C. elegans*, the reproductive system is known to plastically respond to changes of environmental microbiota [60,61]. Thus, it is unsurprising to see the corresponding modules to respond dynamically to diverse bacteria in *P. pacificus*. These results also recapitulate previous findings of signatures of rapid evolution in spermatogenesis-associated genes in *C. elegans* and *P. pacificus* [58,62] This out-of-testis trend of new gene formations was also observed in insects and mammals and likely reflects a combination between rapid evolution of spermatogenesis-associated genes and a more permissive chromatin state [63,64].

Thousands of orphan genes can be associated with biological processes or tissues

The strong support for the association of the two largest coexpression modules with oogenesis and spermatogenesis motivated us to test if more coexpression modules could be annotated with biologically meaningful labels. Therefore, we complemented protein domain information (S8 Fig), spatial transcriptomics (Fig 5A), and TEA (S5 Table) with the overrepresentation of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Fig 5B) and other *P. pacificus* expression gene sets (Fig 5C) [43,47,56,57,65–67]. This allowed us to assign labels for 22 of the 28 largest coexpression modules (Table 1). Please note that these labels are not exclusive but rather describe the most strongly enriched biological terms. Notably, nine of the coexpression

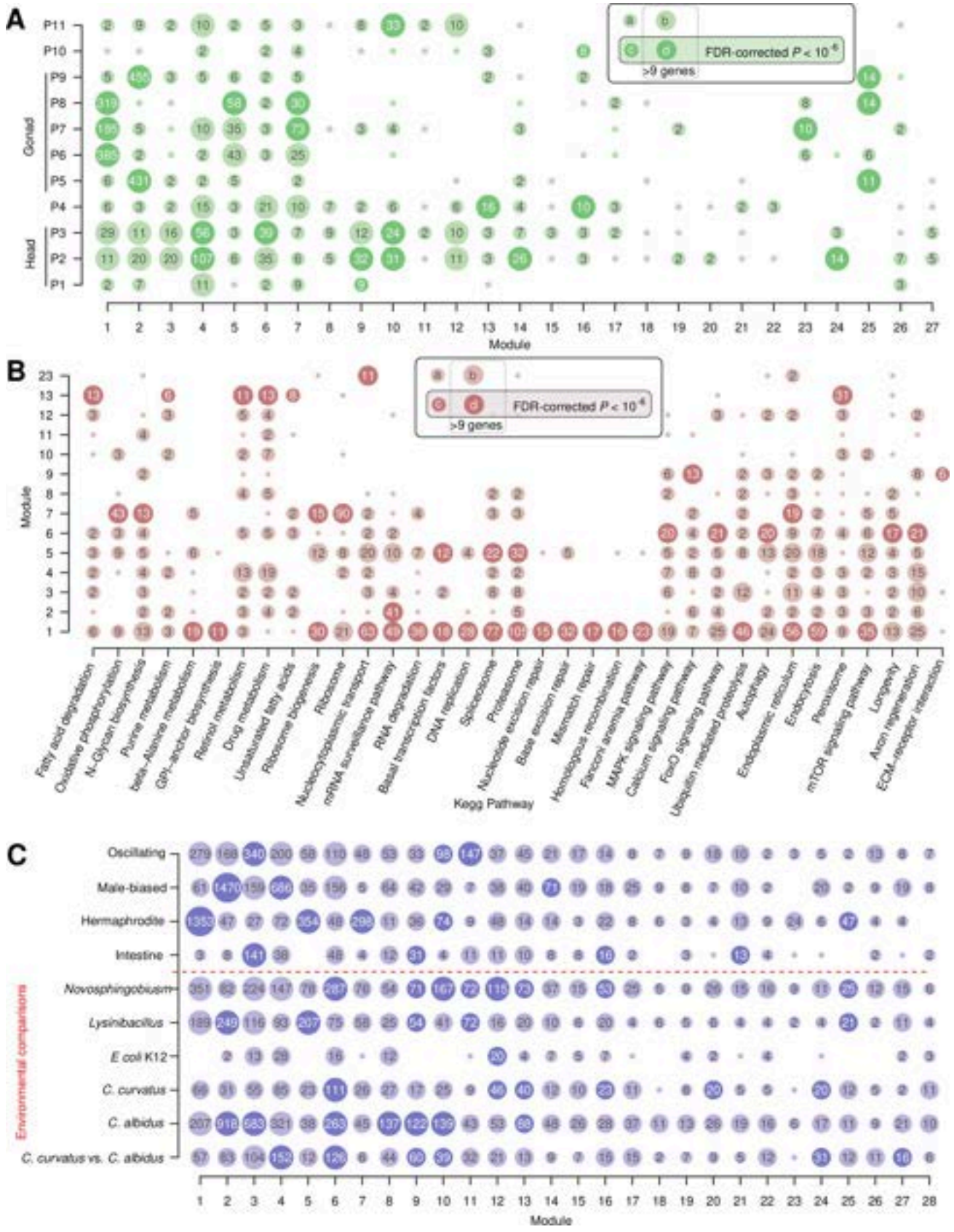

A
Gonad
Head
P11
P10
P9
P8
P7
P6
P5
P4
P3
P2
P1
FDR-corrected P < 10⁻⁶
>9 genes
Module
B
Module
FDR-corrected P < 10⁻⁶
>9 genes
Fatty acid degradation
Oxidative phosphorylation
N-Glycan biosynthesis
Purine metabolism
beta-Alanine metabolism
GPI-anchor biosynthesis
Retinol metabolism
Drug metabolism
Unsaturated fatty acids
Ribosome biogenesis
Ribosome
Nucleocytoplasmic transport
mRNA surveillance pathway
RNA degradation
Basal transcription factors
DNA replication
Spliceosome
Proteasome
Nucleotide excision repair
Base excision repair
Mismatch repair
Homologous recombination
Fanconi anemia pathway
MAPK signaling pathway
Calcium signaling pathway
FoxO signaling pathway
Ubiquitin mediated proteolysis
Autophagy
Endoplasmic reticulum
Endocytosis
Peroxisome
mTOR signaling pathway
Longevity
Axon regeneration
ECM–receptor interaction
Kegg Pathway
C
Oscillating
Male-biased
Hermaphrodite
Intestine
Novosphingobium
Lysinibacillus
E coli K12
C. curvatus
C. albidus
C. curvatus vs. C. albidus
Environmental comparisons
Module

Fig 5. Transcriptomic and metabolic enrichment of coexpression modules. (A) Coexpression modules show distinct overlaps with regional genes (P1-P11) that were identified from spatial transcriptomics. (B) Multiple metabolic pathways are enriched in individual coexpression modules. (C) Coexpression modules were compared with multiple expression gene sets. 16 out of 28 coexpression modules show significant overlap with differentially expressed genes in previous comparisons of environmental microbiota.

https://doi.org/10.1371/journal.pgen.1010832.g005

modules are associated with the nervous system which is also supported by the patterns of G protein coupled receptors (GPCRs), neuropeptides, and nuclear hormone receptors (NHRs) (S9 Fig). All these gene classes are thought to be enriched in neurons [68,69]. In addition, we identified a gland cell related module which captures 17 out of 24 target genes of the regulators of the mouth-form polyphenism in *P. pacificus* [70] (S10 Fig). Altogether, this analysis predicts associations for 13,923 (48%) *P. pacificus* genes with biological processes and anatomical structures. This includes predicted associations for 3,556 (35%) of diplogastrid-specific orphan genes.

Comparisons with previous RNA-seq studies reveal interactions between environmental microbiota and the nervous system

In the current study, we wanted to get a very broad overview of expression changes across a wide range of bacterial environments. Therefore, we decided to sequence only a single

Table 1. Biological annotation for coexpression modules. Most coexpression modules could be labeled with biological processes or tissues based on manual inspection of the results of different overrepresentation analyses.

Module	Number of genes	Name	Evidence
1	3,202	Oogenesis 1	P6-P8, sex-bias, TEA
2	1,995	Spermatogenesis	P5+P9, motile sperm proteins, sex-bias, TEA
3	1,735	Intestine 1	Intestine RNA-seq
4	1,618	Nervous system 1	GPCRs, neuropeptides, NHRs, P1-P3, TEA
5	1,099	Oogenesis 2	P6-P8, sex-bias, TEA
6	921	Nervous system 2	GPCRs, Axon regeneration, P2-P3, TEA
7	472	Oogenesis 3	P6-P8, sex-bias
8	388	Nervous system 3	GPCRs
9	346	Muscle / intestine 2	Intestine RNA-seq, TEA
10	304	Cuticle 1	Collagens, oscillation
11	300	Cuticle 2	Collagens, oscillation, TEA
12	253	Nervous system 4	GPCRs, NHRs
13	241	Intestine 3	Fatty acid degradation, drug metabolism, P4
14	235	Nervous system 5	GPCRs, P2, TEA
15	133	Unknown	
16	116	Intestine 4	Intestine RNA-seq, P4
17	116	Unknown	
18	104	Nervous system 6	TEA
19	94	Nervous system 7	TEA
20	81	Orphan 1	
21	80	Intestine 4	Intestine RNA-seq
22	74	Orphan 2	
23	72	Oogenesis4	P6-P8, TEA
24	64	Gland cell / mouthform	P2, Astacins
25	60	Orphan 3	
26	59	unknown	
27	57	Nervous system 8	TEA
28	56	Nervous system 9	GPCRs

https://doi.org/10.1371/journal.pgen.1010832.t001

transcriptome from many different bacterial environments. This is in contrast to previous studies that focussed on specific interactions between hosts and microbes and defined more robust transcriptomic changes with regard to the standard diet *E. coli* OP50 by sequencing multiple biological replicates [47,65–67]. In order to gain additional support for the environmental regulation of the identified coexpression modules, we quantified how many of the coexpression modules exhibit associations with sets of significantly differentially expressed genes in six transcriptomic comparisons of distinct microbial environments including *Novsphingobium*, *Lysinobacillus*, *E. coli* K12, and two *Cryptococcus* yeast strains [47,65–67]. This showed significant enrichments for 16 modules with differentially expressed genes in at least one of the previous studies (Fig 5C). Moreover, we find the intestinal module 9 and the nervous system related module 6 to be most frequently overrepresented. This further supports the impact of different bacteria on the nematode nervous system with potential consequences on behavior. Such an effect has been experimentally characterized in *P. pacificus* for the case of *Novosphingobium* bacteria. Feeding on these bacteria accelerates development and makes the worms more efficient predators of *C. elegans* [47]. Another example of behavioral modulation constitutes the effect of a neurotransmitter release by bacteria on *C. elegans* perception [45]. Notably, six of the nine nervous system related modules are significantly affected in at least one environmental comparison. This highlights the potential of microbial environments to modulate the nervous system and behavior.

Discussion

Which genes respond to changing environments? Do orphan genes play a role in adapting to diverse microbiota? Can we infer potential functions for orphan genes based on coexpression with known genes? How are new genes integrated into existing biological networks? To study these questions, we characterized the transcriptomic response of *P. pacificus* nematodes to 24 environmental microbiota. This analysis let to the identification of 28 large coexpression modules that contain almost half of all genes in *P. pacificus*. Further integrative analysis associated 22 of these modules with biological processes or tissues. These modules capture previously characterized gene sets such as collagens with oscillating expression [43] or target genes of mouth-form regulators that are expressed in the gland cell [70]. Moreover, these functionally annotated modules contain 3,556 (35%) of diplogastrid-specific orphan genes, of which a large fraction is found in the spermatogenesis associated module 2. While sperm associated genes are known to evolve rapidly [62,71], our study adds evidence that the expression of spermatogenesis-associated genes can be modulated by different microbiota. One major limitation of the current study consists in the fact that we only sequenced a single biological replicate per environment. This strongly limits the usefulness of our data set to dissect the effect of individual bacterial strains on nematode gene expression. However, we would argue that this does not undermine our main findings as coexpression signals can still be inferred from such a data set even if single expression estimates are not robust. In addition, we would like to point out that since the worms have been grown on selected bacteria, the effect of the bacteria likely accumulates in the nematodes throughout development and might be different from an immediate transcriptomic response after short time exposure. Thus, the observed differences could partially represent an indirect consequence of altered developmental rates. Even if only young adults were picked for RNA-seq experiments, we cannot exclude the possibility that the expression of certain gene modules reflects a chronological age rather than a morphological age. Future work could elucidate to what extent the transcriptomic responses represent immediate response to different bacteria or developmental variation. In addition, we currently do not know to what extent different bacteria constitute a diet or could also interact with

nematodes by colonizing their gut [72]. Nevertheless, despite all these uncertainties, we would still argue that most of the observed transcriptomic variation represents either direct or indirect responses to different microbial environments.

One of the main achievements of our work constitutes the association of thousands of *P. pacificus* genes with biological processes and tissues. This includes thousands of diplogastrid-specific orphan genes. Of course, this type of functional annotation represents a completely different level than the knowledge for the two experimentally characterized orphan genes, where knockout and transgenic lines are available that show organism level phenotypes [34,73]. Our functional associations rely on the assumption that coexpression implies cofunctionality. Although this is a frequently employed method for functional assignment [35,36], the example of the out of testis pattern shows that this assumption might not always be true. Specifically, it has been shown that testis-biased expression of many new genes might be an effect of an overall permissive chromatin state leading to higher transcriptional complexity [64]. Thus, our data should rather be considered as a source to generate new hypotheses that have to be tested experimentally. As such, our functional annotations may be helpful to interpret future transcriptomic studies in *P. pacificus*. Some of these modules might also be relevant for studying the polyphenism of the mouth morphology in *P. pacificus*. This phenomenon has developed into one of the best studied animal systems for developmental plasticity [74,75]. Our observation of environmentally dependent expression variation in a gland cell related module that includes many target genes of mouth-form regulators [70] suggests an additional control layer that is independent of the mouth-form. This is because no bacteria are currently known that alter the mouth-form ratio in the highly predatory *P. pacificus* reference strain PS312 [47]. However, it could well be that the same bacteria can induce the predatory morph in a different wild isolate with a low or intermediate frequency of the predatory morph.

Another major finding concerns the integration of new genes into existing gene regulatory networks. Our analysis clearly shows that new genes are not randomly attached to existing modules, but rather specific modules have much higher propensity to acquire new genes. The likelihood to integrate new genes may be associated with the biological function of a given module. For example, while the oogenesis associated modules 1, 5, 7, and 23 are composed of the oldest gene sets, the spermatogenesis-associated module 2 has one of the youngest gene contents. The strong difference in evolvability between these two reproductive processes could be due to strong sperm competition which results from the difference in the number of gametes between both sexes [76]. Another factor could consist in the differential control of transposons between spermatocytes and oocytes resulting in higher potential for molecular innovation during spermatogenesis [77,78]. Another aspect of network evolution concerns the timing of events. The presence of many species-specific orphan genes in some coexpression modules suggests that integration can happen relatively fast. Together with a recent study on the evolution of the polyphenism network [79], our work reveals first insights into the evolution of environmentally responsive networks in *Pristionchus* nematodes and similar studies in other taxonomic groups could be done to test whether these observations reflect general patterns of gene-regulatory network evolution.

Methods

Bacterial Culture Conditions

All bacterial strains were recovered from glycerol stocks [42] and then plated on nematode growth medium (NGM) and incubated overnight at 37˚C. From these plates single colonies were seeded in lysogeny broth (LB) medium and grown overnight at 37˚C in a shaking incubator.

Nematode Culture Conditions

The wild type strain of *P. pacificus* (PS312) was maintained at 20˚C on nematode growth medium (NGM) seeded with *E. coli* OP50 before use in experiments [80]. From every generation, five young adults were transferred to fresh plates with a wormpick.

RNA sequencing

Bleaching was used to synchronize *P. pacificus* nematodes and to remove *E. coli* OP50 before transferring bleached eggs to NGM plates with bacterial strains [80]. These bacterial plates were obtained by spotting 50 μL of overnight bacterial cultures onto each 6 cm NGM plate with an L-spreader followed by incubation for 2 days. The concentration of the overnight bacterial cultures was quantified by measuring the optical density at 600 nm (OD600). To make the OD600 of the initial cultures identical, the OD600 of the overnight cultures were measured and diluted to OD600 = 1 with fresh LB. Worms were grown on these plates at 20˚C and every generation young adult worms were passaged to fresh bacterial plates. This was done for at least two generations. From mixed-stage cultures, 50 young adults with only one egg were manually selected with a wormpick under a binocular microscope into an Eppendorf tube containing 100 μl M9 buffer and immediately frozen at -80˚C. Total RNA was extracted using Direct-Zol RNA Mini prep kit from Zymo Research according to the manufacturer's guidelines. RNA libraries were prepared using the Illumina Truseq RNA library prep kit according to the manufacturer's guidelines. The libraries were quantified using a combination of Qubit and Bioanalyzer (Agilent Technologies) and normalized to 2.5 nM. Samples were sequenced as 150 bp single end reads on multiplexed lanes of an Illumina HiSeq3000 in our inhouse sequencing facility. Raw reads were depositted at the European Nucleotide archive under the study accession PRJEB60166.

Developmental timing

Bacterial strains were grown overnight from a single colony in LB medium. The medium was shaken at 180 rpm and incubated at 30˚C for LRB89 and LRB104 and 37˚C for OP50. These cultures were spotted on 6-cm NGM plates and were grown for 2 days at RT. Synchronized J2 worms were obtained by bleaching and were put on these NGM plates [80]. Nematode cultures were grown at 20˚C. 56h later, the distribution of developmental stages was scored based on the vulval development under a ZEISS SteREO Discovery microscope [47].

Identification of coexpression modules

The raw reads were aligned to the reference *P. pacificus* genome (version El Paco) with STAR (version 2.7.3a) and quantified with featureCounts from the Subread R package (version 2.0.1) using the current gene annotations (El Paco gene annotations version 3) [81–83]. After prefiltering the count matrix by removing genes (rows) that have less than 10 reads total, we retained 23,294 (80.6%) of all *P. pacificus* genes with some evidence of expression. The read counts across the different conditions were normalized with the DESeq2 counts function (with normalized = TRUE option) [84]. The normalized read counts were used to create a coexpression network with MCL [48]. Different correlation and inflation thresholds were assessed for the coexpression network, ranging from 0.6 to 0.8 and from 1.5 to 4 respectively. The network was constructed using a correlation of 0.7 and an inflation threshold of 2. The network modules containing more than 50 genes were selected for further analysis. To assess the robustness of module assignments, we recomputed coexpression networks after either removing the outlier LRB104 or systematically downsampling the data set. Taking the full coexpression network

calculated from 24 RNA-seq samples as reference, we evaluated the classification of 10,000 randomly chosen gene pairs from a subsampled RNA-seq data set. For the full and the subsampled data set, a gene pair was either classified to be part of the same module or not. If a gene pair was classified as being part of the same module in both data sets, such a pair was classified as true positive. If a pair was only part of a module in the subsampled data set, but not in the full data set, this was scored as false positive. True negative and false negative cases were defined analogously. Subsequently, the positive predictive value (PPV) and negative predictive value (NPV) for 10 randomly subsampled data sets of a fixed size (S5 Fig). For comparison, the data set without LRB104 yielded a PPV of 84% and an NPV of 98%.

Motif analysis and phylostratigraphy

To test for overrepresented motifs in the promoter regions, we focused on the 500-bp nucleotide sequence upstream of each annotated gene in a given module. These sequences were taken as input for the findMotif.pl script of the Hypergeometric Optimization of Motif EnRichment (HOMER) suite (version v4.10.1) [49]. As a background set, we used the 500-bp upstream sequences of all the other large coexpression modules. For visualization, we selected among the significant motifs only one representative motif per known motif match and applied a cutoff that a motif needs to be present in at least 20% of the promoter regions of a given module. For analyzing the distribution of phylostrata across coexpression modules, we assigned a gene to a phylostratum based on the presence of the most distant homolog. For this purpose, we performed BLASTP searches (e-value < 0.001) of all 28,896 *P. pacificus* proteins against protein data of nine diplogastrid genomes (version PPCAC) [30], *C. elegans*, *B. xylophilus*, *B. malayi*, and *T. spiralis* (WormBase ParaSite version WBPS14) [85]. This identified 10,318 diplogastrid specific orphan genes. To simulate the integration of new genes into existing network modules, we first defined the ancestral module sizes based on the number of ancient genes (non-orphan genes) for the 28 large modules. We then simulated the integration of new genes by either assigning equal probabilities for attachment to all network modules or by scaling the attachment probability as a linear function of the module size. The second model predicts much more accurately the number of orphan genes per module (Pearson's r = 0.62, S7 Fig). Therefore, we used 100 iterations of the second model to calculate empirical P-values for the overrepresentation of orphan genes ($P < 0.01$, Figs S7 and 4).

Compilation of *P. pacificus* expression gene sets

Sets of 3,502 *P. pacificus* genes with regional expression were extracted from supplementary table S3 in [58]. These genes exhibit enriched expression in at least one out of eleven regions (P1-P11) that were identified by RNA tomography of adult *P. pacificus* animals. A conversion table between assembled transcripts (version trinity 2016) and current gene annotations (El Paco gene annotations V3) were obtained from the supplementary table S10 in [43]. We extracted the set of 2,964 developmentally oscillating genes from the supplementary table S3 in [43]. In addition, we reanalyzed *P. pacificus* RNA-seq data sets to compile candidate genes with intestinal expression [56], sex-biased expression [57], and differentially expressed genes in response to altered microbial environments [47,65–67]. For the intestinal data and exposure to *Lysinibacillus*, we identified candidate genes by aligning RNA-seq reads to the *P. pacificus* reference assembly (version El Paco) with the help of the tophat alignment program (version v2.0.14, default settings) and testing for differential expression using the cuffdiff program (version v2.2.1, $P<0.1$) [86]. This yielded 723 intestine enriched genes, and 1,959 candidate genes with differential expression after one hour exposure to *Lysinibacillus*. For the data sets with multiple biological replicates [47,57,65,67], we realigned RNA-seq data with STAR (version

2.7.1a) [81], generated count matrices with the featureCounts function of the subread package in R (version 3.6.3) [82] and for significant differential expression with the DESeq2 package (FDR-corrected $P<0.05$) [84].

Overrepresentation of gene families, metabolic pathways, and expression sets

To test for overrepresentation of specific gene sets among the coexpression modules, we first generated protein domain annotations by the hmmsearch program (version 3.3, -E 0.001 option) against the Pfam-A.hmm database (version 3.1b2). Similarly, KEGG annotations were obtained by identification of orthologs in the KEGG database using the blastkoala web application (with the 'eukaryotes' and 'family_eukaryotes' values for taxonomic group and database level, respectively) [87]. Genes with orthologs in the KEGG database were then annotated with the corresponding *C. elegans* KEGG accessions. From the protein domain predictions, we identified 1,434 genes with protein domains that were termed as GPCRs and 275 putative NHR genes based on the presence of the Hormone_recep domain (PF00104). Potential neuropeptides were identified by BLASTP search (-evalue 0.001) of 107 *C. elegans* neuropeptides (*flp* and *nlp* genes) against the current *P. pacificus* proteins. For 59 *C. elegans* neuropeptides, we could identify 85 homologs in *P. pacificus* of which we only used 38 single copy candidates as neuropeptide candidates. These annotations were combined with various expression data sets to perform overrepresentation analysis using the Fisher's exact-test with multiple testing correction (FDR corrected $P < 10^6$) in R (version 3.6.3).

Annotation of coexpression modules

In order to associate the coexpression modules with biological processes or tissues, we complemented the results of the overrepresentation analysis in *P. pacificus* with tissue enrichment analysis (TEA) of *C. elegans* orthologs [59]. Specifically, *C. elegans* orthologs of *P. pacificus* genes were extracted from the best reciprocal BLASTP hit data set from Athanasouli *et al.* (2020) [83] and submitted to the WormBase web interface (https://wormbase.org/tools/enrichment/tea/tea.cgi, version WS283) [83]. We manually inspected the results of all different analyses and subjectively assigned biological processes or tissues for a given module. Functional assignments with multiple types of evidence (e.g. Protein domains, spatial transcriptomics, sex-biased expression) should be considered as high confidence associations whereas associations that are supported by only a single analysis are more uncertain.

Supporting information

S1 Fig. Analysis of the most similar and dissimilar RNA-seq data sets. The scatter plots show the normalized expression (TPM) for different pairs of samples. As lowly expressed genes tend to be more variable, we visualized the number of genes with at least two-fold expression difference across multiple mean expression levels.
(PDF)

S2 Fig. Parameter combinations for MCL clustering. We evaluated the total number of modules and singleton modules as a function of the inflation factor (I) and correlation coefficient (r). High r and I values generally increase the number of modules, whereas low r and I parameters generate fewer but larger modules. We decided to use r = 0.7 and I = 2 for the final analysis, because it gave a moderate number of modules with a relatively low number of singletons.
(PDF)

S3 Fig. Developmental timing on different bacteria. The distribution of developmental stages 56h after J2 synchronization is shown for *P. pacificus* worms on *E. coli* OP50 and two *Wautersiella* bacteria (5 biological replicates). Nematodes grow slower on both *Wautersiella* strains. The significance level was computed by a χ^2 -test (mean *P*-value from all pairwise comparisons).
(PDF)

S4 Fig. Comparison of MCL networks constructed with or without LRB104. The heatmap shows fraction genes from modules of the complete coexpression network (including LRB104) that overlap within a given module from the coexpression network without LRB104. For most modules there exists a 1–1 correspondence between both networks, indicating that the network structure is robust with regard to the *Wautersiella* LRB104 data set.
(PDF)

S5 Fig. Comparison of Coexpression networks on subsampled RNA-seq data. Taking the full coexpression network calculated from 24 RNA-seq samples as reference, we evaluated the classification of 10,000 randomly chosen gene pairs using subsampled RNA-seq data. For the full and the subsampled data set, a gene pair was either classified to be part of the same module or not. This allowed us to calculate the positive predictive value (PPV, panel A) and negative predictive value (NPV, panel B) for 10 randomly subsampled data sets of a fixed size. While the NPV is always close to 1, the PPV shows drastic differences between the full and subsampled data. This suggests that with fewer RNA-seq samples, additional gene pairs are assigned to the same coexpression module. However, with additional samples, such modules may be split into smaller components. Consistently, the number of genes in large modules ($N>50$) is much higher for smaller sample sizes (panel C). Panel D shows the number of diplogastrid-specific orphan genes in large modules.
(PDF)

S6 Fig. Regulatory architecture of coexpression modules. Significantly overrepresented motifs were identified for each module by the HOMER software. We arbitrarily selected motifs that occurred in at least 20% of promoters of a given module and visualized their distribution across all modules. Note that the complementary regulation by less frequent motifs and other regulatory mechanisms such as microRNAs are not considered here. Sequence logos for each motif are shown at the right and the labels to the left indicate the best motif match among known motifs.
(PDF)

S7 Fig. Enrichment analysis for simulated network evolution. We simulated the integration of novel genes into existing networks by estimating ancestral module sizes based on the number of ancient genes (non-orphan genes) and then assigning an equivalent number of orphan genes to existing modules with equal probabilities (panel A and C) and with probabilities that were proportional to the module size (panel B and D). The scatterplots show the observed and simulated number of orphan genes per module. The barplots show the median enrichment of the observed relative to the simulated number of orphan genes (error bars indicate the minimal and maximal values from 100 simulations).
(PDF)

S8 Fig. Overrepresented protein domains in coexpression modules. Specific protein domains are strongly overrepresented in twelve coexpression modules. The left barplot shows the number of genes with a given protein domain for each module and the the right bar plot shows the negative logarithm of the FDR corrected P-value (Fisher's exact test). The most

significant association is between Motile sperm proteins (PF00635) and coexpression module 2.
(PDF)

S9 Fig. Distribution of GPCRs, NHRs, and neuropeptides across the coexpression modules.
(PDF)

S10 Fig. Expression of selected module 24 genes. The heatmap shows the expression of genes that are shared between module 24 and the target genes of mouth form regulators and additional module 24 genes with 1–1 orthologs in *C. elegans*. The *C. elegans* ortholog of PPA25527 (D1044.3) is reported to be expressed in the gland cell and reporter lines of multiple candidate genes show expression in the gland of *P. pacificus* (Sieribriennikov et al. 2020) [70].
(PDF)

S1 Table. Coexpression modules.
(XLSX)

S2 Table. Phylostrata.
(XLSX)

S3 Table. Protein domain information.
(XLSX)

S4 Table. *Pristionchus pacificus* KEGG annotation.
(XLSX)

S5 Table. Tissue enrichment analysis (TEA).
(XLSX)

Acknowledgments

We would like to thank all members of the Sommer lab for helpful discussions.

Author Contributions

Conceptualization: Christian Rödelsperger.

Data curation: Marina Athanasouli, Christian Rödelsperger.

Formal analysis: Marina Athanasouli, Christian Rödelsperger.

Investigation: Nermin Akduman, Waltraud Röseler, Penghieng Theam.

Methodology: Marina Athanasouli, Christian Rödelsperger.

Project administration: Marina Athanasouli, Christian Rödelsperger.

Supervision: Christian Rödelsperger.

Visualization: Marina Athanasouli, Christian Rödelsperger.

Writing – original draft: Marina Athanasouli, Christian Rödelsperger.

Writing – review & editing: Christian Rödelsperger.

References

1. Tautz D, Domazet-Lošo T. The evolutionary origin of orphan genes. Nat Rev Genet. 2011; 12: 692–702. https://doi.org/10.1038/nrg3053 PMID: 21878963

2. Arendsee ZW, Li L, Wurtele ES. Coming of age: orphan genes in plants. Trends Plant Sci. 2014; 19: 698–708. https://doi.org/10.1016/j.tplants.2014.07.003 PMID: 25151064
3. Weisman CM, Murray AW, Eddy SR. Many but not all lineage-specific genes can be explained by homology detection failure. PLoS Biol. 2020; 18: e3000862. https://doi.org/10.1371/journal.pbio.3000862 PMID: 33137085
4. Schlötterer C. Genes from scratch—the evolutionary fate of de novo genes. Trends Genet. 2015; 31: 215–219. https://doi.org/10.1016/j.tig.2015.02.007 PMID: 25773713
5. Van Oss SB, Carvunis A-R. De novo gene birth. PLoS Genet. 2019; 15: e1008160. https://doi.org/10.1371/journal.pgen.1008160 PMID: 31120894
6. Knowles DG, McLysaght A. Recent de novo origin of human protein-coding genes. Genome Res. 2009; 19: 1752–1759. https://doi.org/10.1101/gr.095026.109 PMID: 19726446
7. Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabidó E, Kondova I, Bontrop R, et al. Origins of De Novo Genes in Human and Chimpanzee. PLoS Genet. 2015; 11: e1005721. https://doi.org/10.1371/journal.pgen.1005721 PMID: 26720152
8. Zhao L, Saelao P, Jones CD, Begun DJ. Origin and spread of de novo genes in Drosophila melanogaster populations. Science. 2014; 343: 769–772. https://doi.org/10.1126/science.1248286 PMID: 24457212
9. Gubala AM, Schmitz JF, Kearns MJ, Vinh TT, Bornberg-Bauer E, Wolfner MF, et al. The Goddard and Saturn Genes Are Essential for Drosophila Male Fertility and May Have Arisen De Novo. Mol Biol Evol. 2017; 34: 1066–1082. https://doi.org/10.1093/molbev/msx057 PMID: 28104747
10. Zhang W, Gao Y, Long M, Shen B. Origination and evolution of orphan genes and de novo genes in the genome of Caenorhabditis elegans. Sci China Life Sci. 2019; 62: 579–593. https://doi.org/10.1007/s11427-019-9482-0 PMID: 30919281
11. Prabh N, Rödelsperger C. De Novo, divergence, and Mixed Origin Contribute to the Emergence of Orphan Genes in Pristionchus Nematodes. G3. 2019; 9: 2277–2286. https://doi.org/10.1534/g3.119.400326 PMID: 31088903
12. Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, et al. Proto-genes and de novo gene birth. Nature. 2012; 487: 370–374. https://doi.org/10.1038/nature11184 PMID: 22722833
13. Vakirlis N, Hebert AS, Opulente DA, Achaz G, Hittinger CT, Fischer G, et al. A Molecular Portrait of De Novo Genes in Yeasts. Mol Biol Evol. 2018; 35: 631–645. https://doi.org/10.1093/molbev/msx315 PMID: 29220506
14. Zhang L, Ren Y, Yang T, Li G, Chen J, Gschwend AR, et al. Rapid evolution of protein diversity by de novo origination in Oryza. Nature Ecology & Evolution. 2019; 3: 679–690.
15. Yates TB, Feng K, Zhang J, Singan V, Jawdy SS, Ranjan P, et al. The Ancient Salicoid Genome Duplication Event: A Platform for Reconstruction of De Novo Gene Evolution in *Populus trichocarpa*. Genome Biology and Evolution. 2021; 13: evab198.
16. Wissler L, Gadau J, Simola DF, Helmkampf M, Bornberg-Bauer E. Mechanisms and dynamics of orphan gene emergence in insect genomes. Genome Biol Evol. 2013; 5: 439–455. https://doi.org/10.1093/gbe/evt009 PMID: 23348040
17. Athanasouli M, Rödelsperger C. Analysis of repeat elements in the Pristionchus pacificus genome reveals an ancient invasion by horizontally transferred transposons. BMC Genomics. 2022; 23: 523. https://doi.org/10.1186/s12864-022-08731-1 PMID: 35854227
18. Rödelsperger C, Prabh N, Sommer RJ. New Gene Origin and Deep Taxon Phylogenomics: Opportunities and Challenges. Trends Genet. 2019; 35: 914–922. https://doi.org/10.1016/j.tig.2019.08.007 PMID: 31610892
19. Vakirlis N, Acar O, Hsu B, Castilho Coelho N, Van Oss SB, Wacholder A, et al. De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. Nat Commun. 2020; 11: 781. https://doi.org/10.1038/s41467-020-14500-z PMID: 32034123
20. Chen S, Zhang YE, Long M. New Genes in *Drosophila* Quickly Become Essential. Science. 2010; 330: 1682–1685.
21. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG. More than just orphans: are taxonomically-restricted genes important in evolution? Trends Genet. 2009; 25: 404–413. https://doi.org/10.1016/j.tig.2009.07.006 PMID: 19716618
22. Cardoso-Silva CB, Aono AH, Mancini MC, Sforça DA, da Silva CC, Pinto LR, et al. Taxonomically Restricted Genes Are Associated With Responses to Biotic and Abiotic Stresses in Sugarcane (spp.). Front Plant Sci. 2022; 13: 923069.
23. Cheng CH. Evolution of the diverse antifreeze proteins. Curr Opin Genet Dev. 1998; 8: 715–20. https://doi.org/10.1016/s0959-437x(98)80042-7 PMID: 9914209

24. Sommer RJ. Pristionchus pacificus. WormBook. 2006; 1–8.

25. Howard RS, Giacomelli M, Lozano-Fernandez J, Edgecombe GD, Fleming JF, Kristensen RM, et al. The Ediacaran origin of Ecdysozoa: integrating fossil and phylogenomic data. J Geol Soc London. 2022; 179: gs2021–107.

26. Wang X, Sommer RJ. Antagonism of LIN-17/Frizzled and LIN-18/Ryk in nematode vulva induction reveals evolutionary alterations in core developmental pathways. PLoS Biol. 2011; 9: e1001110. https://doi.org/10.1371/journal.pbio.1001110 PMID: 21814488

27. Ishita Y, Chihara T, Okumura M. Serotonergic modulation of feeding behavior in Caenorhabditis elegans and other related nematodes. Neurosci Res. 2020; 154: 9–19. https://doi.org/10.1016/j.neures.2019.04.006 PMID: 31028772

28. Lo W-S, Roca M, Dardiry M, Mackie M, Eberhardt G, Witte H, et al. Evolution and Diversity of TGF-β Pathways are Linked with Novel Developmental and Behavioral Traits. Molecular Biology and Evolution. 2022;39: msac252.

29. Borchert N, Dieterich C, Krug K, Schütz W, Jung S, Nordheim A, et al. Proteogenomics of *Pristionchus pacificus* reveals distinct proteome structure of nematode models. Genome Research. 2010; 20: 837–846.

30. Prabh N, Roeseler W, Witte H, Eberhardt G, Sommer RJ, Rödelsperger C. Deep taxon sampling reveals the evolutionary dynamics of novel gene families in *Pristionchus* nematodes. Genome Research. 2018; 28: 1664–1674.

31. Prabh N, Rödelsperger C. Multiple Pristionchus pacificus genomes reveal distinct evolutionary dynamics between de novo candidates and duplicated genes. Genome Res. 2022; 32: 1315–1327. https://doi.org/10.1101/gr.276431.121 PMID: 35618417

32. Stein JC, Yu Y, Copetti D, Zwickl DJ, Zhang L, Zhang C, et al. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. Nat Genet. 2018; 50: 285–296. https://doi.org/10.1038/s41588-018-0040-0 PMID: 29358651

33. Dowling D, Schmitz JF, Bornberg-Bauer E. Stochastic Gain and Loss of Novel Transcribed Open Reading Frames in the Human Lineage. Genome Biol Evol. 2020; 12: 2183–2195. https://doi.org/10.1093/gbe/evaa194 PMID: 33210146

34. Mayer MG, Rödelsperger C, Witte H, Riebesell M, Sommer RJ. The Orphan Gene dauerless Regulates Dauer Development and Intraspecific Competition in Nematodes by Copy Number Variation. PLoS Genet. 2015; 11: e1005146. https://doi.org/10.1371/journal.pgen.1005146 PMID: 26087034

35. Li J, Singh U, Arendsee Z, Wurtele ES. Landscape of the Dark Transcriptome Revealed Through Remining Massive RNA-Seq Data. Front Genet. 2021; 12: 722981. https://doi.org/10.3389/fgene.2021.722981 PMID: 34484307

36. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Research. 2017; 45: D362–D368. https://doi.org/10.1093/nar/gkw937 PMID: 27924014

37. Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, et al. A gene expression map for Caenorhabditis elegans. Science. 2001; 293: 2087–2092. https://doi.org/10.1126/science.1061603 PMID: 11557892

38. Marchesi JR, Ravel J. The vocabulary of microbiome research: a proposal. Microbiome. 2015; 3: 31. https://doi.org/10.1186/s40168-015-0094-5 PMID: 26229597

39. Watson E, MacNeil LT, Ritter AD, Safak Yilmaz L, Rosebrock AP, Caudy AA, et al. Interspecies Systems Biology Uncovers Metabolites Affecting C. elegans Gene Expression and Life History Traits. Cell. 2014; 156: 759–770. https://doi.org/10.1016/j.cell.2014.01.047 PMID: 24529378

40. Zhang F, Berg M, Dierking K, Félix M-A, Shapira M, Samuel BS, et al. Caenorhabditis elegans as a Model for Microbiome Research. Frontiers in Microbiology. 2017; 8: 485. https://doi.org/10.3389/fmicb.2017.00485 PMID: 28386252

41. Sinha A, Rae R, Iatsenko I, Sommer RJ. System wide analysis of the evolution of innate immunity in the nematode model species Caenorhabditis elegans and Pristionchus pacificus. PLoS One. 2012; 7: e44255. https://doi.org/10.1371/journal.pone.0044255 PMID: 23028509

42. Akduman N, Rödelsperger C, Sommer RJ. Culture-based analysis of Pristionchus-associated microbiota from beetles and figs for studying nematode-bacterial interactions. PLoS One. 2018; 13: e0198018. https://doi.org/10.1371/journal.pone.0198018 PMID: 29864131

43. Sun S, Rödelsperger C, Sommer RJ. Single worm transcriptomics identifies a developmental core network of oscillating genes with deep conservation across nematodes. Genome Res. 2021; 31: 1590–1601. https://doi.org/10.1101/gr.275303.121 PMID: 34301622

44. Baskaran P, Rödelsperger C, Prabh N, Serobyan V, Markov GV, Hirsekorn A, et al. Ancient gene duplications have shaped developmental stage-specific expression in Pristionchus pacificus. BMC Evol Biol. 2015; 15: 185. https://doi.org/10.1186/s12862-015-0466-2 PMID: 26370559

45. O'Donnell MP, Fox BW, Chao P-H, Schroeder FC, Sengupta P. A neurotransmitter produced by gut bacteria modulates host sensory behaviour. Nature. 2020; 583: 415–420. https://doi.org/10.1038/s41586-020-2395-5 PMID: 32555456

46. Lansdon P, Carlson M, Ackley BD. Wild-type Caenorhabditis elegans isolates exhibit distinct gene expression profiles in response to microbial infection. BMC Genomics. 2022; 23: 229. https://doi.org/10.1186/s12864-022-08455-2 PMID: 35321659

47. Akduman N, Lightfoot JW, Röseler W, Witte H, Lo W-S, et al. Bacterial vitamin B_{12} production enhances predatory behaviors in nematodes. ISME J. 2020; 14:1494–1507.

48. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002; 30: 1575–1584. https://doi.org/10.1093/nar/30.7.1575 PMID: 11917018

49. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010; 38: 576–589. https://doi.org/10.1016/j.molcel.2010.05.004 PMID: 20513432

50. Narasimhan K, Lambert SA, Yang AWH, Riddell J, Mnaimneh S, Zheng H, et al. Mapping and analysis of Caenorhabditis elegans transcription factor sequence specificities. Elife. 2015; 44: e06967. https://doi.org/10.7554/eLife.06967 PMID: 25905672

51. Burglin TR, Ruvkun G. Regulation of ectodermal and excretory function by the C. elegans POU homeobox gene ceh-6. Development. 2001; 128: 779–790. https://doi.org/10.1242/dev.128.5.779 PMID: 11171402

52. Domazet-Lošo T, Brajković J, Tautz D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends in Genetics. 2007; 23: 533–539. https://doi.org/10.1016/j.tig.2007.08.014 PMID: 18029048

53. Kikuchi T, Cotton JA, Dalzell JJ, Hasegawa K, Kanzaki N, McVeigh P, et al. Genomic insights into the origin of parasitism in the emerging plant pathogen Bursaphelenchus xylophilus. PLoS Pathog. 2011; 7: e1002219. https://doi.org/10.1371/journal.ppat.1002219 PMID: 21909270

54. Foster JM, Grote A, Mattick J, Tracey A, Tsai Y-C, Chung M, et al. Sex chromosome evolution in parasitic nematodes of humans. Nat Commun. 2020; 11: 1964. https://doi.org/10.1038/s41467-020-15654-6 PMID: 32327641

55. Mitreva M, Jasmer DP, Zarlenga DS, Wang Z, Abubucker S, Martin J, et al. The draft genome of the parasitic nematode Trichinella spiralis. Nat Genet. 2011; 43: 228–235. https://doi.org/10.1038/ng.769 PMID: 21336279

56. Lightfoot JW, Chauhan VM, Aylott JW, Rödelsperger C. Comparative transcriptomics of the nematode gut identifies global shifts in feeding mode and pathogen susceptibility. BMC Res Notes. 2016; 9: 142. https://doi.org/10.1186/s13104-016-1886-9 PMID: 26944260

57. Rödelsperger C, Röseler W, Prabh N, Yoshida K, Weiler C, Herrmann M, et al. Phylotranscriptomics of Pristionchus Nematodes Reveals Parallel Gene Loss in Six Hermaphroditic Lineages. Curr Biol. 2018; 28: 3123–3127.e5. https://doi.org/10.1016/j.cub.2018.07.041 PMID: 30245109

58. Rödelsperger C, Ebbing A, Sharma DR, Okumura M, Sommer RJ, Korswagen HC. Spatial Transcriptomics of Nematodes Identifies Sperm Cells as a Source of Genomic Novelty and Rapid Evolution. Mol Biol Evol. 2021; 38: 229–243. https://doi.org/10.1093/molbev/msaa207 PMID: 32785688

59. Angeles-Albores D N Lee RY, Chan J, Sternberg PW. Tissue enrichment analysis for C. elegans genomics. BMC Bioinformatics. 2016; 17: 366.

60. Le TS, Nguyen THG, Ha BH, Huong BTM, Nguyen TTH, Vu KD, et al. Reproductive Span of is Extended by Sp. J Nematol. 2022; 54: 20220010.

61. Kissoyan KAB, Peters L, Giez C, Michels J, Pees B, Hamerich IK, et al. Exploring Effects of Protective Natural Microbiota on Host Physiology. Front Cell Infect Microbiol. 2022; 12: 775728.

62. Cutter AD, Ward S. Sexual and Temporal Dynamics of Molecular Evolution in C. elegans Development. Molecular Biology and Evolution. 2005; 22: 178–188. https://doi.org/10.1093/molbev/msh267 PMID: 15371532

63. Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. Proc Natl Acad Sci U S A. 2006; 103: 9935–9939. https://doi.org/10.1073/pnas.0509809103 PMID: 16777968

64. Soumillon M, Necsulea A, Weier M, Brawand D, Zhang X, Gu H, et al. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. Cell Rep. 2013; 3: 2179–2190. https://doi.org/10.1016/j.celrep.2013.05.031 PMID: 23791531

65. Han Z, Sieriebriennikov B, Susoy V, Lo W-S, Igreja C, Dong C, et al. Horizontally Acquired Cellulases Assist the Expansion of Dietary Range in Pristionchus Nematodes. Mol Biol Evol. 2022;39: msab370. https://doi.org/10.1093/molbev/msab370 PMID: 34978575

66. Lo W-S, Han Z, Witte H, Röseler W, Sommer RJ. Synergistic interaction of gut microbiota enhances the growth of nematode through neuroendocrine signaling. Curr Biol. 2022; 32: 2037–2050.e4. https://doi.org/10.1016/j.cub.2022.03.056 PMID: 35397201

67. Sanghvi GV, Baskaran P, Röseler W, Sieriebriennikov B, Rödelsperger C, Sommer RJ. Life History Responses and Gene Expression Profiles of the Nematode Pristionchus pacificus Cultured on Cryptococcus Yeasts. PLoS One. 2016; 11: e0164881. https://doi.org/10.1371/journal.pone.0164881 PMID: 27741297

68. Sural S, Hobert O. Nematode nuclear receptors as integrators of sensory information. Curr Biol. 2021; 31: 4361–4366.e2. https://doi.org/10.1016/j.cub.2021.07.019 PMID: 34348120

69. Hobert O. The neuronal genome of Caenorhabditis elegans. WormBook. 2013. 1–106. https://doi.org/10.1895/wormbook.1.161.1 PMID: 24081909

70. Sieriebriennikov B, Sun S, Lightfoot JW, Witte H, Moreno E, et al. Conserved nuclear hormone receptors controlling a novel plastic trait target fast-evolving genes expressed in a single cell. PLoS Genet. 2020; 16: e1008687. https://doi.org/10.1371/journal.pgen.1008687 PMID: 32282814

71. Verster AJ, Styles EB, Mateo A, Derry WB, Andrews BJ, Fraser AG. Taxonomically Restricted Genes with Essential Functions Frequently Play Roles in Chromosome Segregation in and. G3. 2017; 7: 3337–3347.

72. Dirksen P, Marsh SA, Braker I, Heitland N, Wagner S, Nakad R, et al. The native microbiome of the nematode Caenorhabditis elegans: gateway to a new host-microbiome model. BMC Biol. 2016; 14: 38. https://doi.org/10.1186/s12915-016-0258-1 PMID: 27160191

73. Lightfoot JW, Wilecki M, Rödelsperger C, Moreno E, Susoy V, Witte H, et al. Small peptide-mediated self-recognition prevents cannibalism in predatory nematodes. Science. 2019; 364: 86–89. https://doi.org/10.1126/science.aav9856 PMID: 30948551

74. Sommer RJ, Dardiry M, Lenuzzi M, Namdeo S, Renahan T, Sieriebriennikov B, et al. The genetics of phenotypic plasticity in nematode feeding structures. Open Biol. 2017;7. https://doi.org/10.1098/rsob.160332 PMID: 28298309

75. Namdeo S, Moreno E, Rödelsperger C, Baskaran P, Witte H, Sommer RJ. Two independent sulfation processes regulate mouth-form plasticity in the nematode. Development. 2018; 145: dev166272.

76. Parker GA. Conceptual developments in sperm competition: a very brief synopsis. Philosophical Transactions of the Royal Society B: Biological Sciences. 2020; 375:20200061.

77. Kurhanewicz NA, Dinwiddie D, Bush ZD, Libuda DE. Elevated Temperatures Cause Transposon-Associated DNA Damage in C. elegans Spermatocytes. Curr Biol. 2020; 30: 5007–5017.e4.

78. Bhalla N. Meiosis: Is Spermatogenesis Stress an Opportunity for Evolutionary Innovation? Current biology: CB. 2020. pp. R1471–R1473. https://doi.org/10.1016/j.cub.2020.10.042 PMID: 33352126

79. Casasa S, Biddle JF, Koutsovoulos GD, Ragsdale EJ. Polyphenism of a Novel Trait Integrated Rapidly Evolving Genes into Ancestrally Plastic Networks. Mol Biol Evol. 2021; 38: 331–343. https://doi.org/10.1093/molbev/msaa235 PMID: 32931588

80. Pires-daSilva A. Pristionchus pacificus protocols. WormBook. 2013; 1–20. https://doi.org/10.1895/wormbook.1.114.2 PMID: 23505072

81. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29: 15–21. https://doi.org/10.1093/bioinformatics/bts635 PMID: 23104886

82. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014; 30: 923–30. https://doi.org/10.1093/bioinformatics/btt656 PMID: 24227677

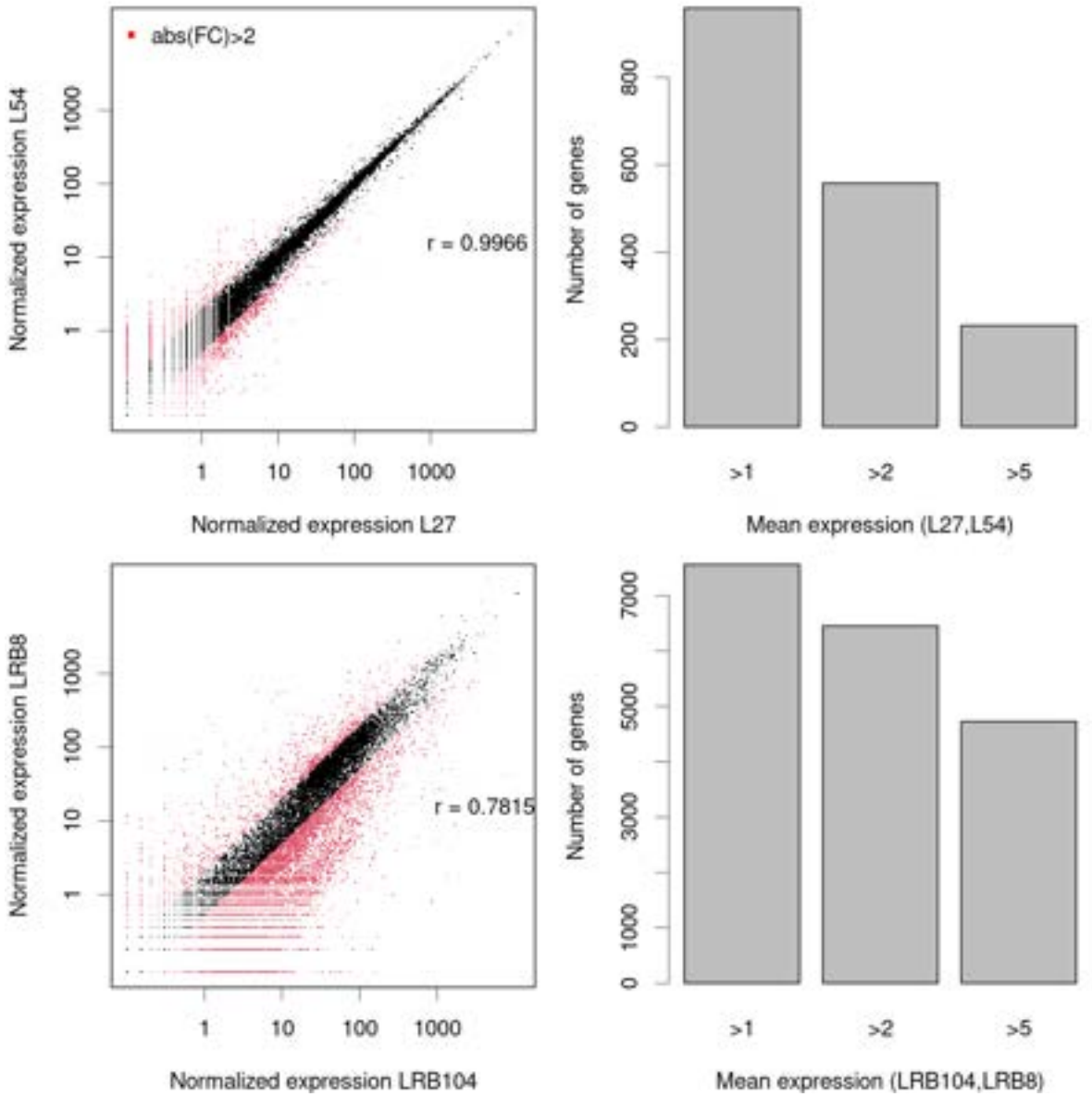
83. Athanasouli M, Witte H, Weiler C, Loschko T, Eberhardt G, Sommer RJ, et al. Comparative genomics and community curation further improve gene annotations in the nematode Pristionchus pacificus. BMC Genomics. 2020; 21: 708. https://doi.org/10.1186/s12864-020-07100-0 PMID: 33045985

84. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15: 550. https://doi.org/10.1186/s13059-014-0550-8 PMID: 25516281

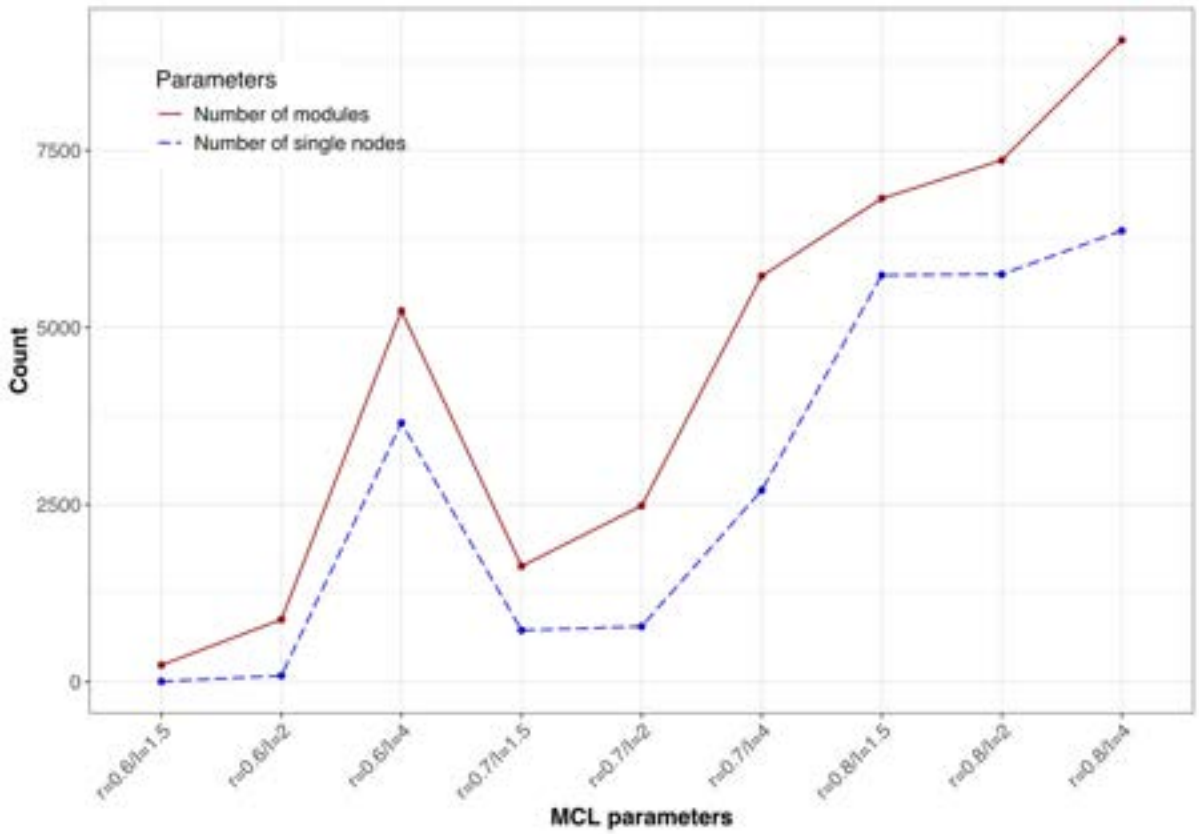
85. Bolt BJ, Rodgers FH, Shafie M, Kersey PJ, Berriman M, Howe KL. Using WormBase ParaSite: An Integrated Platform for Exploring Helminth Genomic Data. Methods Mol Biol. 2018; 1757: 471–491.

86. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012; 7: 562–578. https://doi.org/10.1038/nprot.2012.016 PMID: 22383036

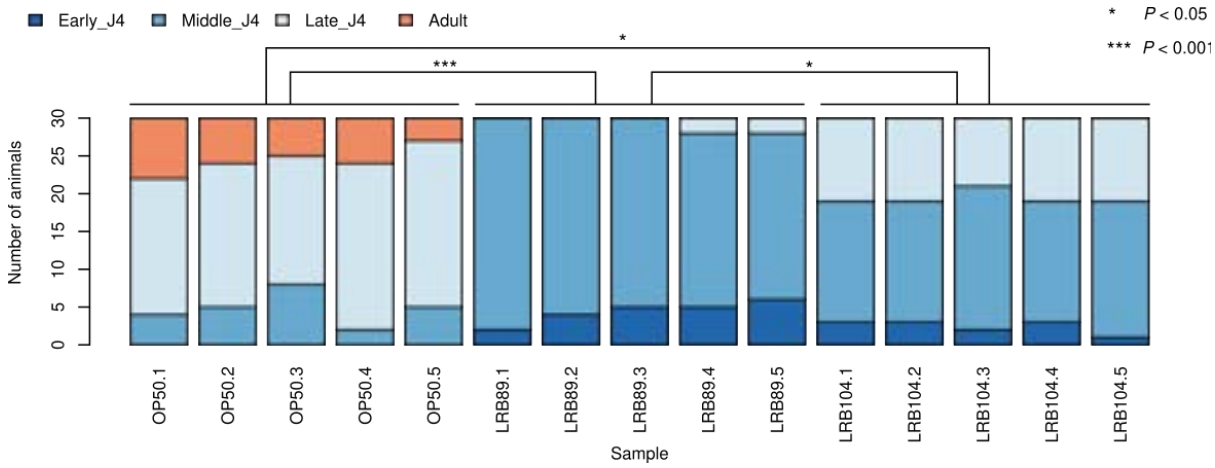
87. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. J Mol Biol. 2016; 428: 726–731. https://doi.org/10.1016/j.jmb.2015.11.006 PMID: 26585406



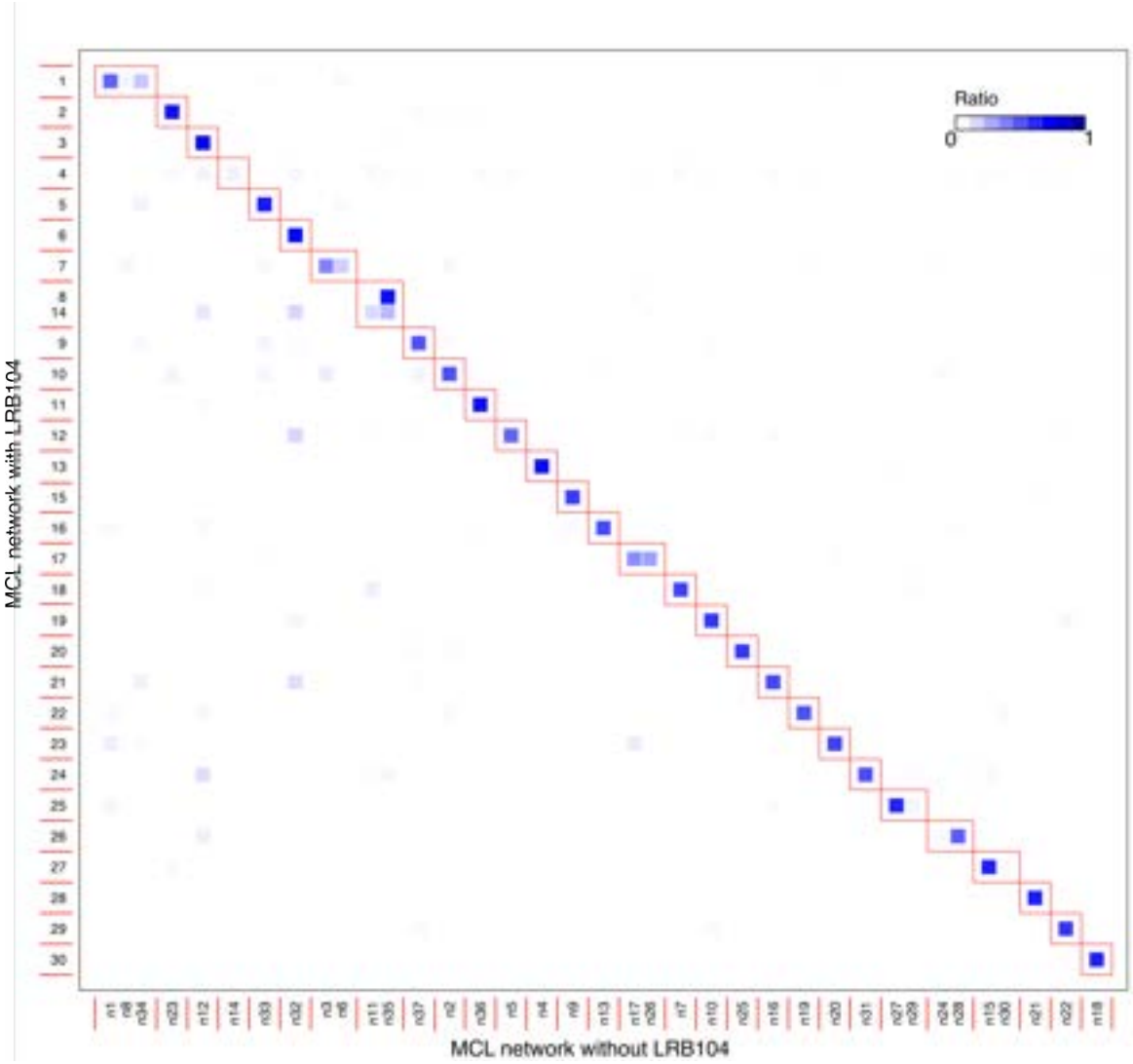
S1 Fig. Analysis of the most similar and dissimilar RNA-seq data sets. The scatter plots show the normalized expression (TPM) for different pairs of samples. As lowly expressed genes tend to be more variable, we visualized the number of genes with at least two-fold expression difference across multiple mean expression levels.



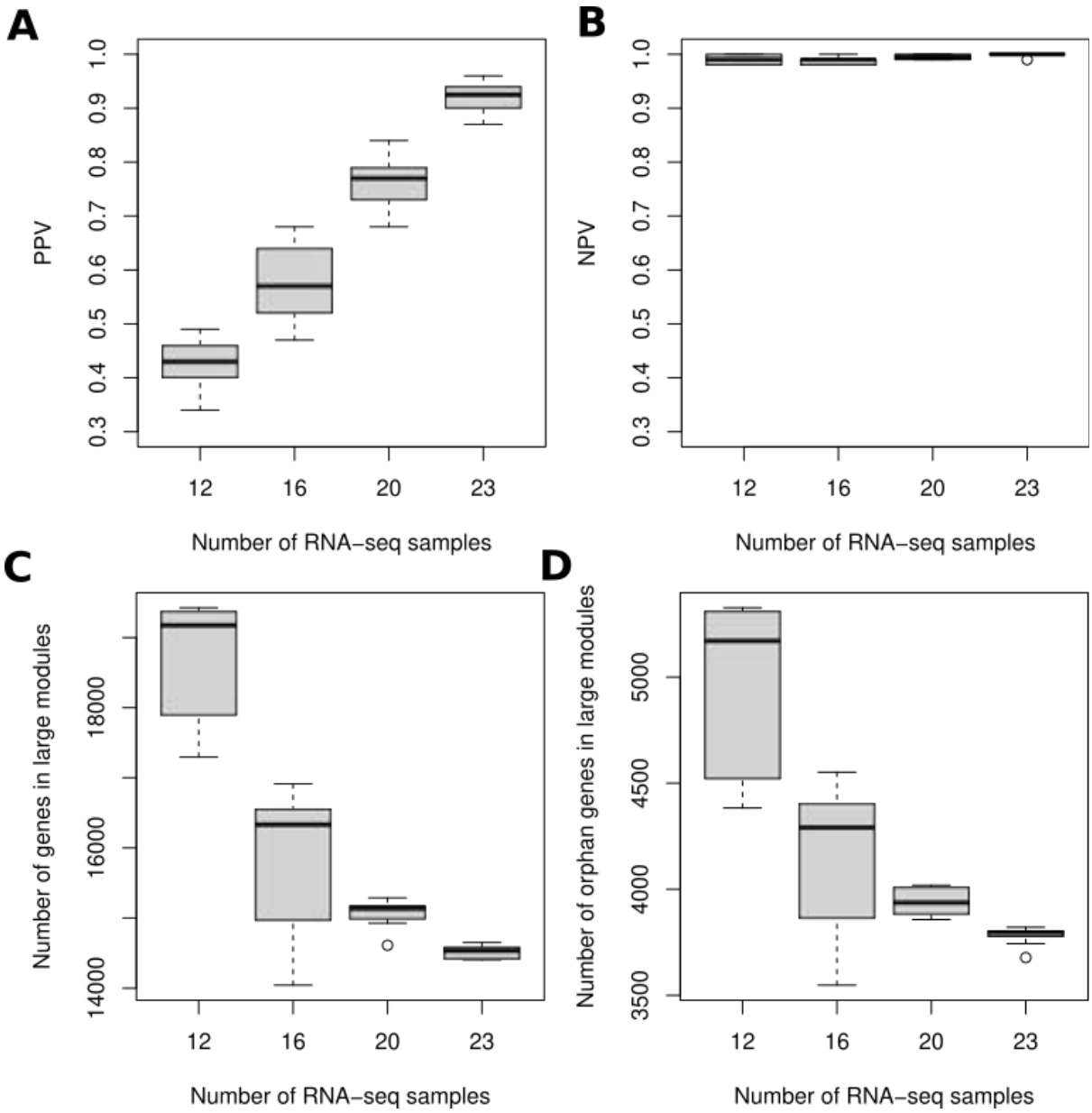
S2 Fig. Parameter combinations for MCL clustering. We evaluated the total number of modules and singleton modules as a function of the inflation factor (I) and correlation coefficient (r). High r and I values generally increase the number of modules, whereas low r and I parameters generate fewer but larger modules. We decided to use r=0.7 and I=2 for the final analysis, because it gave a moderate number of modules with a relatively low number of singletons.



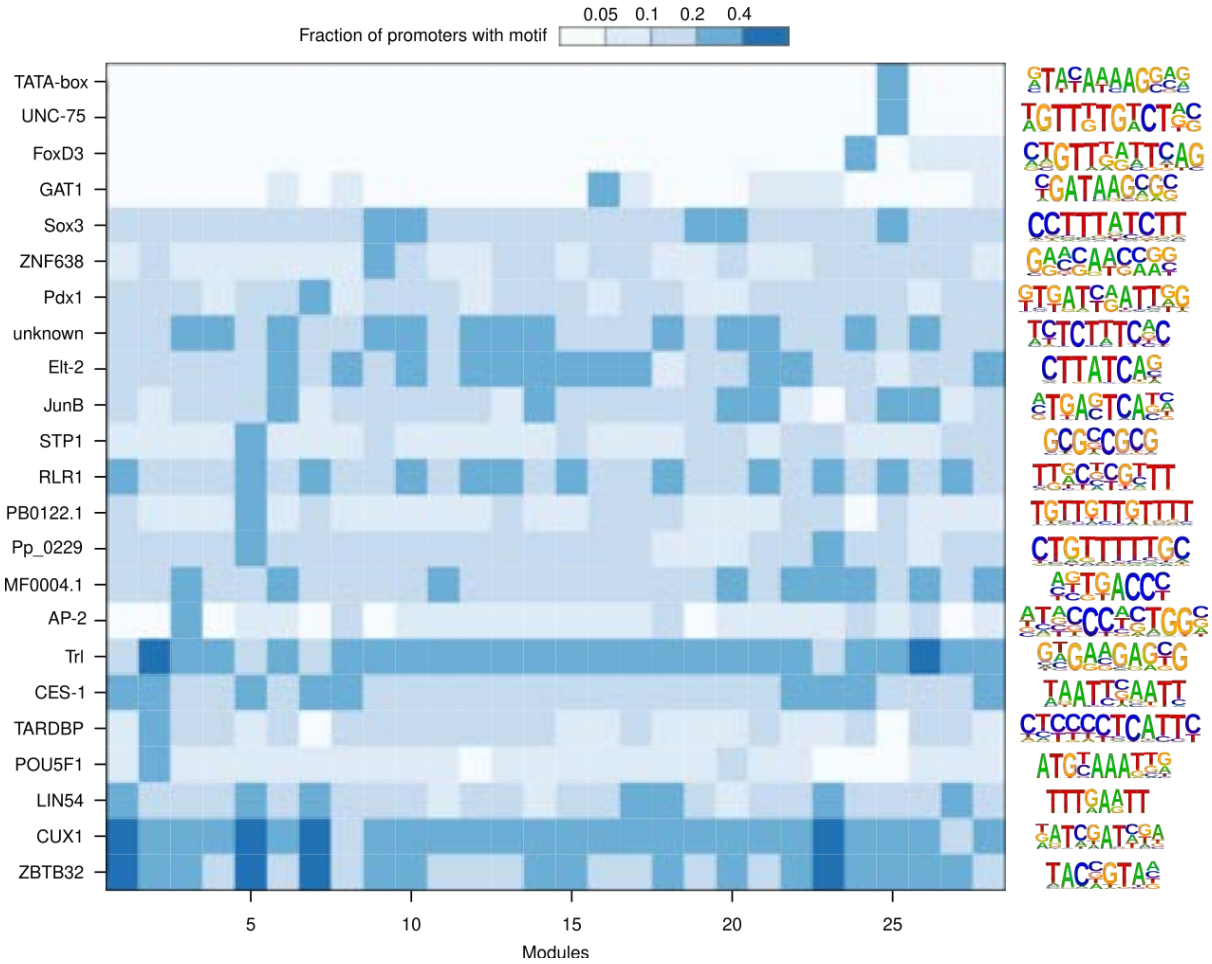
S3 Fig. Developmental timing on different bacteria. The distribution of developmental stages 56h after J2 synchronization is shown for *P. pacificus* worms on *E. coli* OP50 and two *Wautersiella* bacteria (5 biological replicates). Nematodes grow slower on both *Wautersiella* strains. The significance level was computed by a χ^2-test (mean *P*-value from all pairwise comparisons).



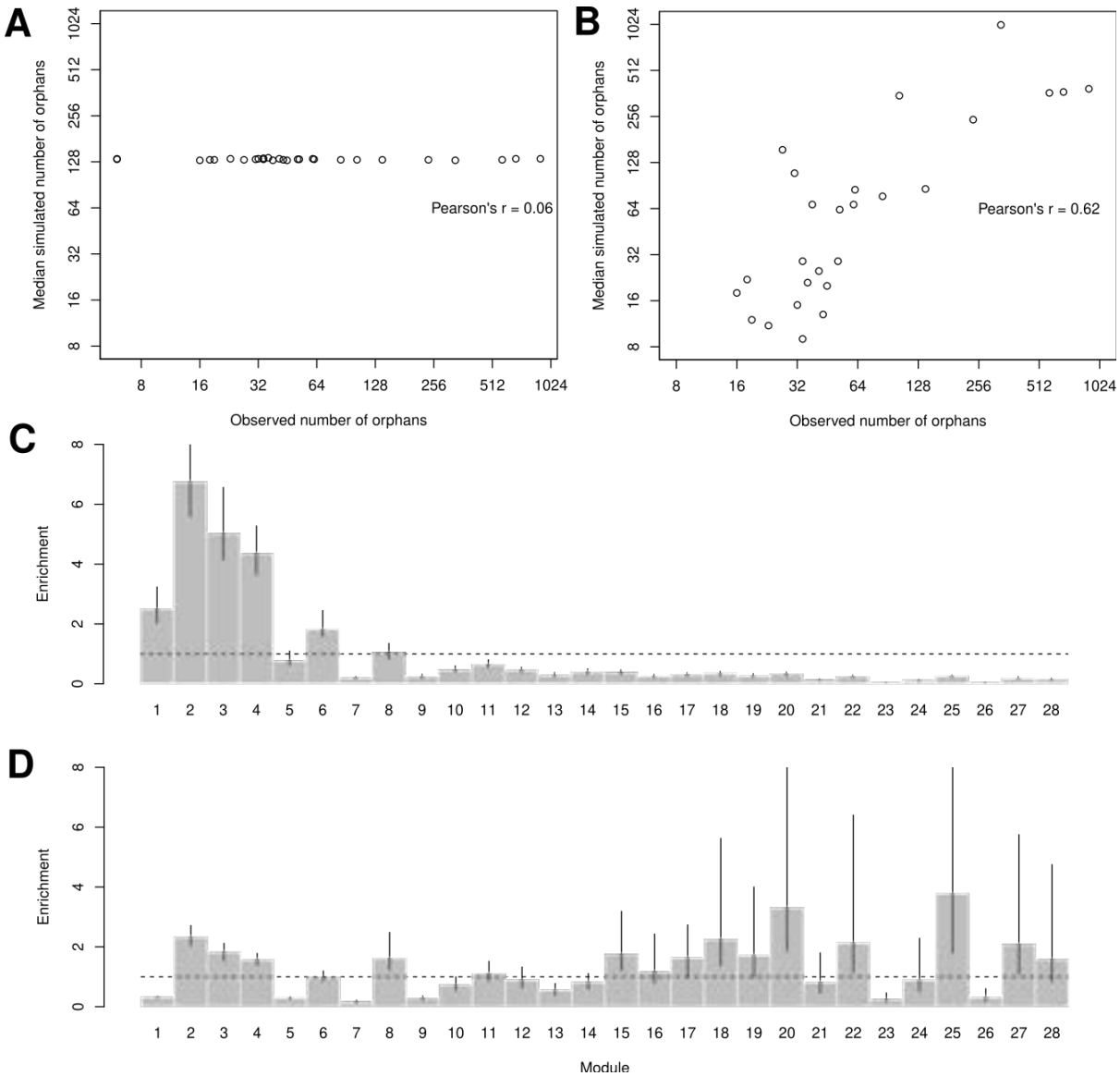
S4 Fig. Comparison of MCL networks constructed with or without LRB104. The heatmap shows fraction genes from modules of the complete coexpression network (including LRB104) that overlap within a given module from the coexpression network without LRB104. For most modules there exists a 1-1 correspondence between both networks, indicating that the network structure is robust with regard to the *Wautersiella* LRB104 data set.



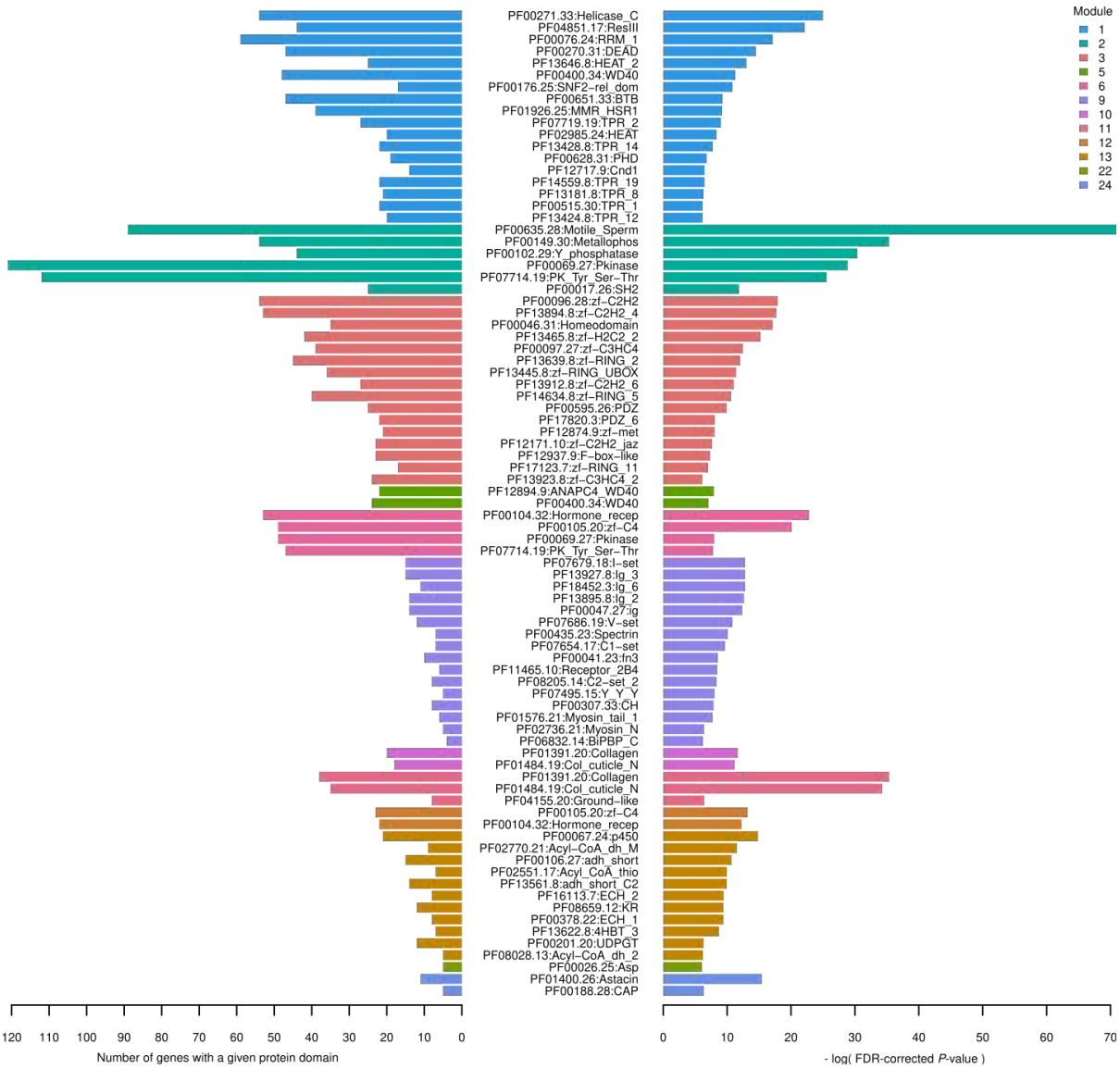
S5 Fig. Comparison of Coexpression networks on subsampled RNA-seq data. Taking the full coexpression network calculated from 24 RNA-seq samples as reference, we evaluated the classification of 10,000 randomly chosen gene pairs using subsampled RNA-seq data. For the full and the subsampled data set, a gene pair was either classified to be part of the same module or not. This allowed us to calculate the positive predictive value (PPV, panel A) and negative predictive value (NPV, panel B) for 10 randomly subsampled data sets of a fixed size. While the NPV is always close to 1, the PPV shows drastic differences between the full and subsampled data. This suggests that with fewer RNA-seq samples, additional gene pairs are assigned to the same coexpression module. However, with additional samples, such modules may be split into smaller components. Consistently, the number of genes in large modules (N>50) is much higher for smaller sample sizes (panel C). Panel D shows the number of diplogastrid-specific orphan genes in large modules.



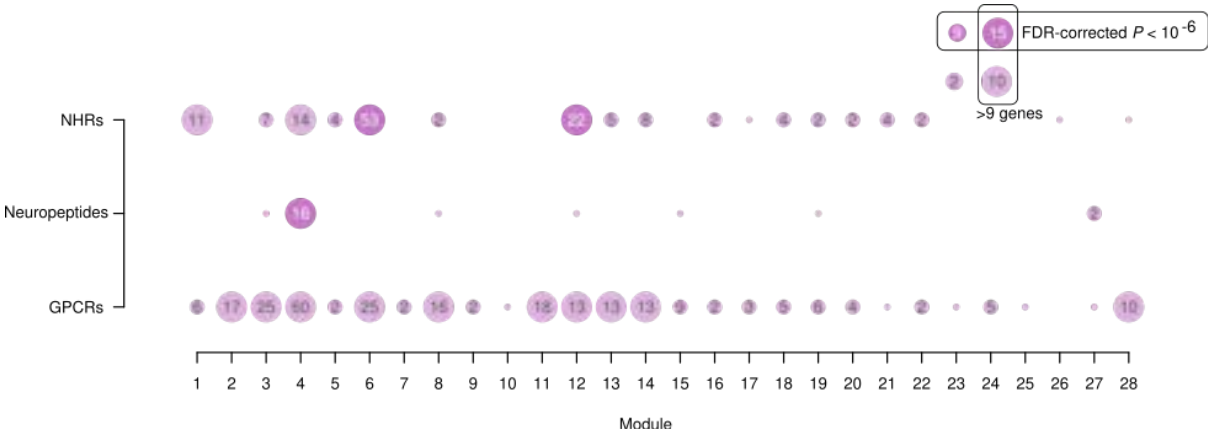
S6 Fig.. Regulatory architecture of coexpression modules. Significantly overrepresented motifs were identified for each module by the HOMER software. We arbitrarily selected motifs that occurred in at least 20% of promoters of a given module and visualized their distribution across all modules. Note that the complementary regulation by less frequent motifs and other regulatory mechanisms such as microRNAs are not considered here. Sequence logos for each motif are shown at the right and the labels to the left indicate the best motif match among known motifs.



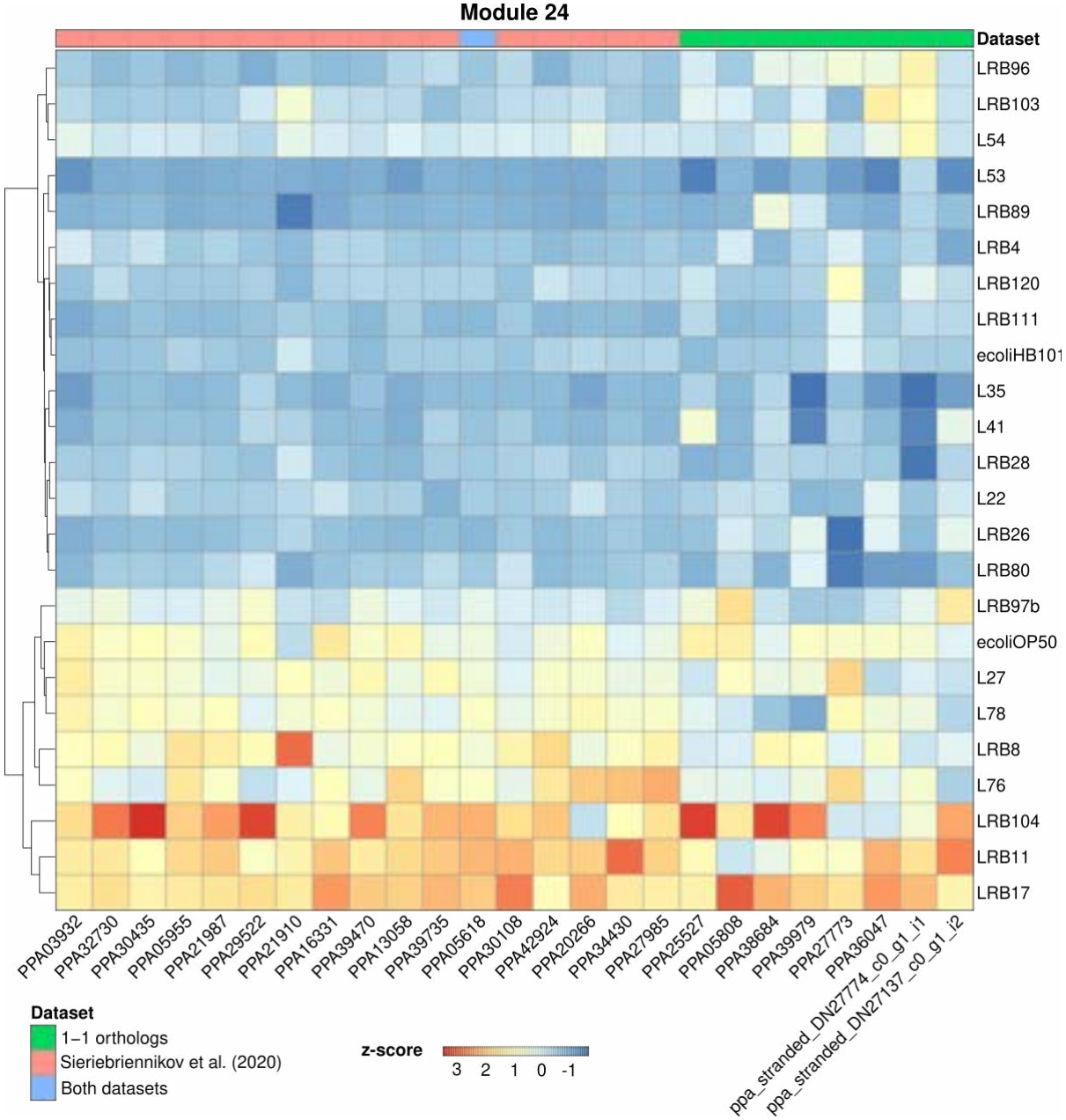
S7 Fig. Enrichment analysis for simulated network evolution. We simulated the integration of novel genes into existing networks by estimating ancestral module sizes based on the number of ancient genes (non-orphan genes) and then assigning an equivalent number of orphan genes to existing modules with equal probabilities (panel A and C) and with probabilities that were proportional to the module size (panel B and D). The scatterplots show the observed and simulated number of orphan genes per module. The barplots show the median enrichment of the observed relative to the simulated number of orphan genes (error bars indicate the minimal and maximal values from 100 simulations).



S8 Fig. Overrepresented protein domains in coexpression modules. Specific protein domains are strongly overrepresented in twelve coexpression modules. The left barplot shows the number of genes with a given protein domain for each module and the the right bar plot shows the negative logarithm of the FDR corrected P-value (Fisher's exact test). The most significant association is between Motile sperm proteins (PF00635) and coexpression module 2.



S9 Fig. Distribution of GPCRs, NHRs, and neuropeptides across the coexpression modules.



S10 Fig. Expression of selected module 24 genes. The heatmap shows the expression of genes that are shared between module 24 and the target genes of mouth form regulators and additional module 24 genes with 1-1 orthologs in *C. elegans*. The *C. elegans* ortholog of PPA25527 (D1044.3) is reported to be expressed in the gland cell and reporter lines of multiple candidate genes show expression in the gland of *P. pacificus* (Sieribriennikov et al. 2020) [1].

References

1. Sieriebriennikov B, Sun S, Lightfoot JW, Witte H, Moreno E, et al. Conserved nuclear hormone receptors controlling a novel plastic trait target fast-evolving genes expressed in a single cell. PLoS Genet. 2020;16: e1008687.

The genome assembly of *Rhabditoides inermis* from a complex microbial community reveals further evidence for parallel gene family expansions across multiple nematodes

Christian Rödelsperger[1,*], Waltraud Röseler[1], Marina Athanasouli[1], Sara Wighard[1,2], Matthias Herrmann[1], Ralf J. Sommer[1,*]

[1] Department for Integrative Evolutionary Biology, Max Planck Institute for Biology, Max-Planck-Ring 9, 72076 Tübingen, Germany

[2] Current address: Institute of Molecular Biotechnology of the Austrian Academy of Sciences (IMBA), Vienna BioCenter (VBC), 1030 Vienna, Austria.

[*] Corresponding authors:

christian.roedelsperger@tuebingen.mpg.de

ralf.sommer@tuebingen.mpg.de

Abstract

Background Free-living nematodes such as *Caenorhabditis elegans* and *Pristionchus pacificus* are powerful model systems for linking specific traits to their underlying genetic basis. To trace the evolutionary history of a candidate gene, a robust phylogenomic framework is indispensable.

Results In this work, we generated a near chromosome-scale genome assembly of the nematode *Rhabditoides inermis* which had previously been proposed as the sister group of the family Diplogastridae to which *P. pacificus* belongs. The genome was assembled from a complex microbial community that consists of multiple bacteria and a fungus of the genus *Vanrija*. The *R. inermis* genome has five chromosomes that likely arose from recent fusions of different Nigon elements. Phylogenomic analysis grouped *R. inermis* within a clade including *C. elegans*, *Mesorhabditis belari* and other rhabditids and thus, did not support a sister group relationship between *R. inermis* and the family Diplogastridae. Comparative genomic analyses identified abundant lineage-specific orthogroups which reveal evidence for parallel expansions of environmentally responsive gene families.

Conclusions. Our work demonstrates the value of the *R. inermis* genome as a resource for future phylogenomic analysis and for studying gene family evolution.

Keywords: Comparative genomics, evolution, metagenome, *Rhabditoides regina*, *Pristionchus pacificus*, *Caenorhabditis, elegans*, *Vanrija albida*

Background

Nematodes are the most abundant animals on Earth and have colonized virtually all ecosystems including multiple animal and plant hosts [1–3]. In addition, with *Caenorhabditis elegans* being a nematode, a vast portion of our biological knowledge originates from basic nematode research. In order to investigate the extent in which biological processes known from *C. elegans* are also conserved in other animals, the free-living nematode *Pristionchus pacificus* had been introduced as a satellite model organism for comparative biology [4]. Over the past years, *P. pacificus* became a more independent model system as new research directions emerged that are unrelated to previous *C. elegans* research. This research focused on an evolutionary novelty of the family Diplogastridae to which *P. pacificus* belongs. Specifically, unlike *C. elegans* and many other nematodes that live in soil habitats and feed on bacteria, *P. pacificus* can form teeth-like denticles in its mouth that allow killing and predation [4]. This predatory form is one of two alternative mouth forms, an example of developmental (phenotypic) plasticity without any intermediate morphologies [5]. The *stenostomatous* (St) morph forms a single tooth with a narrow stoma resulting in strict bacterial feeding. In contrast, the *eurystomatous* (Eu) form has two teeth with a wider stoma, which, in addition to feeding on bacteria, can also kill other nematodes. Research of

developmental plasticity and predatory feeding behaviors have established *P. pacificus* as a completely independent model system [6, 7]. The mouth-form dimorphism has become a major model to investigate the molecular mechanisms regulating developmental plasticity, resulting in the identification of a complex gene regulatory network [6, 8–11]. Comparative studies revealed extensive morphological diversification of these feeding structures, highlighting the evolutionary and ecological significance of developmental plasticity. Specifically, more than 50 species of 29 genera of the Diplogastridae family were investigated by morphometric analysis providing multiple examples of structural diversification and evidence for the loss of the dimorphism in individual lineages [12]. Together, these studies have provided a powerful model for investigating plasticity, its molecular causes and organismal significance [13, 14].

The family Diplogastridae to which *P. pacificus* belongs, is a monophyletic taxon that has been subject to intense investigations in the last decade (reviewed in [15–17]). This included the isolation and characterization of new material resulting in the description of six new diplogastrid genera since 2011 [18–23]. These findings suggest that the current phylogenetic and taxonomic knowledge of Diplogastridae is likely still incomplete. In addition, the sister group of the Diplogastridae is unknown and multiple different suggestions have been made over the last two decades. For example, the family Bunonematidae was suggested as a sister group based on morphological comparisons [24] and small subunit (SSU) ribosomal RNA sequencing [25]. In contrast, other rDNA analyses suggested that *Rhabditoides inermis* represents the sister taxon [26–28]. However, a complementary phylogenetic study grouped *R. inermis* and *C. elegans* in a clade of species that are all members of the family Rhabditidae [12]. Recent phylogenomic analysis of the phylum Nematoda did not include *R. inermis* because no genome draft is available [29]. To

fill the existing knowledge gap, we here performed genome sequencing of *R. inermis* by PacBio sequencing and Hi-C scaffolding. Using the newly available *R. inermis* genome we performed phylogenomic analysis, which did not provide supporting evidence for a sister group relationship between Diplogastridae and *R. inermis*. Instead, phylogenomic analysis supports a sister group relationship between the Diplogastridae family with a clade of Rhabditidae that includes the genera *Oscheius*, *Caenorhabditis*, *Mesorhabditis*, *Rhabditoides* and *Haemonchus*.

Results

R. inermis can be repeatedly found in association with the burying beetle *N. vespilloides* in southern Germany

Multiple studies found *R. inermis* to be associated with the burying beetle *Nicrophorus vespilloides* [30, 31]. Burying beetles have a complex biology, e.g. they conceal small vertebrate carcasses underground and prepare these carcasses for consumption by their own offspring [32]. In 2020, we obtained *Nicrophorus vespilloides* beetles from three localities in southern Germany (Bayreuth, Freiburg and Tübingen). We isolated nematodes using standard procedures [33] and obtained more than 12 independent lines that were morphologically similar to the available reference strain *R. inermis* RS5678. Sequencing of the 18S rRNA revealed high similarity and mating experiments confirmed that all strains belong to the same species. Thus, *R. inermis* is found in association with *N. vespilloides* in high frequency and all obtained strains belong to the same species. Also, when using related burying beetle species, which we found in lower frequency, we did not obtain *Rhabditoides* material other than *R. inermis*.

All new isolates of *R. inermis* show a number of features that are different from many other Rhabditidae (including *C. elegans*), but are consistent with original comparative analyses [30]. Specifically, *R. inermis* grows extremely rapidly under laboratory conditions with a generation time of 2-3 days (20° C) when grown on *E. coli* OP50. Distinctively, *R. inermis* is viviparous, resulting in cultures with few if any eggs on the agar plates. Together, we document the consistent association of *R. inermis* with the burying beetle *N. vespilloides* in southern Germany. We established the isolate RS5678 from Great Britain as a reference strain and performed standard inbreeding procedures for 10 generations resulting in a strain RS5678B that was used for genome sequencing. We obtained additional whole genome sequences from two of the german isolates RS6171 and RS6371.

Genome sequencing data of *R. inermis* represents a complex microbial community

In order to sequence the genome of the inbred *R. inermis* strain RS5678B, we generated genomic DNA libraries from mixed-stage cultures growing on agar plates and sequenced them on the Pacific Biosciences Sequel II platform. This resulted in 35 Gb of HiFi reads with a median read length of 12 kb. This was assembled into a raw assembly of 2228 contigs spanning 419 Mb (N50=0.4Mb). To reduce the degree of allelism resulting from remaining heterozygosity in the *R. inermis* genome, we ran the HaploMerger2 software on the repeat masked raw assembly [34]. This reduced the number of contigs to 782 spanning 253 Mb with an N50 of 0.9 Mb. We further generated Hi-C data and ran the YaHS software to scaffold the *R. inermis* genome [35]. Visualization of the average GC content and Hi-C coverage for the resulting scaffolds identified five large scaffolds (>10Mb) with a GC content of 47% that span 157Mb. However, a subset of scaffolds also showed an unusually high or

low GC content and were mostly sequenced less deeply (Fig. 1A). The Hi-C contact map of the 22 largest scaffolds suggested spatial proximity between the five largest scaffolds, but almost no Hi-C signal between the smaller scaffolds (Fig. 1B). Hypothesizing that the smaller scaffolds could possibly represent microbial contamination on the plates, we annotated protein coding genes and performed a homology search against the NCBI nr database summarizing the phylogenetic distribution of the best hits for all predicted proteins of a given scaffold. This confirmed the microbial origin of these scaffolds (Fig. 1C). Interestingly, none of the microbial contaminants represented *E. coli* OP50 suggesting that all *E. coli* bacteria had been consumed by the time DNA was extracted. Additional sequencing data from two independently isolated *R. inermis* strains showed some sequencing coverage for all identified microbes, suggesting a stable association between *R. inermis* and its microbial community (Additional File 1, Fig. S1).

A complete fungal genome was assembled from sequencing data of *R. inermis* cultures

While bacterial protein predictions showed a median of more than 90% protein sequence identity with their best hits in the NCBI nr database, multiple scaffolds showed lower identity (~80%) to fungal sequences of the genus *Vanrija* (Fig. 1C) [36, 37]. This could potentially indicate that our sequencing data represents a novel fungal species or at least that no whole genome data for this species is publicly available. To better characterize our fungal genome (labeled CR7), we first assessed its degree of completeness. The eight scaffolds comprise 22.5Mb of genomic sequence with a GC content of around 65%, which is very comparable to genomic data of other *Vanrija* genomes (Fig. 2). Furthermore, benchmarking of universal single copy orthologs (BUSCO) identified 94.8% of the fungal orthologs (odb9) in our genome assembly [38, 39]. To get more insights into the

phylogenetic relationships between the new assembly and publicly available *Vanrija* genomes, we first generated evidence-based gene annotations based on RNA-seq data from *R. inermis* nematode cultures and existing protein annotations of *Vanrija humicola* and *Apiotrichum porosum* [37, 40] and then reconstructed a genome-wide phylogeny based on concatenated alignments of orthologous proteins (Fig. 2, see *Methods*). This confirmed that our fungal genome is closely related but distinct from any other *Vanrija* species, for which whole genome data is available. We then extracted large subunit (LSU) rRNA and internal transcribed spacer (ITS) sequences from the new assembly and performed a combined phylogenetic analysis together with publicly available sequences from diverse *Vanrija* species on NCBI (Additional file 1, Fig. S2). This grouped our sequences together with sequences from *V. albida*. Finally, to further support that our genome most likely represents *V. albida,* we downloaded 56 *V. albida* sequences from NCBI and performed a BLASTN search against our assembly. This showed a median of 100% identity strongly supporting that our genome assembly most likely represents an isolate of *V. albida*. To make this genome publicly available, we deposited it at DDBJ/ENA/GenBank under the accession JAZAQS000000000.

R. inermis has five chromosomes with near-complete coverage by the five largest scaffolds

To compile a clean version of the *R. inermis* genome, we manually inspected the Hi-C data, GC content, and coverage data for the different scaffolds and performed additional BLASTN searches against the NCBI database. This resulted in a classification of the scaffolds into 213 candidates that are presumably of nematode origin and a subset of scaffolds that represents the microbial community on which the worms were grown. The completeness of the *R. inermis* genome was estimated by BUSCO to be 88.8% (single-copy and duplicates) using the

nematode ortholog data set (odb9) [38, 39]. This value is comparable to recently generated nematode genome assemblies from cultures without large-scale microbial contamination [41, 42]. Note that the expected BUSCO completeness value should be around 90%, but will also depend on evolutionary distance from the set of reference species in the ortholog set. The complete *R. inermis* assembly spans 173.4 Mb with an N50 value of 31.4 Mb. The five largest scaffolds comprise 156.7 Mb and thus represent more than 90% of the total assembly. This suggests that *R. inermis* has five chromosomes. To obtain further support for this, we performed a Hoechst staining of the gonads as described previously [42]. This indeed revealed additional evidence for the presence of five chromosomes in *R. inermis* oocytes (Fig. 3A). Thus, genomic data as well as karyotyping analysis support that *R. inermis* has five chromosomes.

Nigon elements are largely intact in the *R. inermis* genome

The structure of most nematode genomes with chromosome-scale assemblies reflects a specific configuration of seven ancestral linkage blocks that had been called Nigon elements A,B,C,D,E,N and X [43, 44]. These Nigon elements constitute large blocks of macrosynteny between nematode genomes. This means that genes within a Nigon element are frequently reshuffled, whereas rearrangements of genes from different elements are rare. Most nematode chromosomes, specifically in the suborder Rhabditina, correspond to exactly one Nigon element, whereas some other chromosomes are fusions between different Nigon elements. Examples of fused chromosomes are the *Caenorhabditis* chromosome X, *P. pacificus* chromosome I and many others [44–47]. The observation of five chromosomes in *R. inermis* suggests that two fusions of Nigon elements occurred. We generated evidence-based gene annotations based on protein homology and transcriptome data with a

BUSCO completeness of 86%. We used these protein annotations to visualize Nigon elements across the *R. inermis* genome, which revealed one fusion of Nigon element A and C resulting in *R. inermis* chromosome I and a second fusion between elements E and X resulting in *R. inermis* chromosome X (Fig. 3B). Note that we labeled chromosome X based on the observation that Nigon element X tends to form the X chromosome [44]. However, all chromosomes have very similar coverage (Fig. 1A), which makes the identification of a good candidate for a sex chromosome difficult. Based on these observations, we would speculate that the sex chromosome of *R. inermis* was formed recently and there is very little differentiation between the X and Y chromosome.

Analysis of coding density, repeat content and genetic diversity reveals evidence for the presence of chromosome centers

Many nematode chromosomes also display an internal structure with chromosomal centers and arms [45, 48]. As the chromosomes of *C. elegans* and *P. pacificus* are holocentric [49, 50], *i.e.* centromeric function is distributed along the whole length of the chromosomes, chromosome centers are not centromeres. The chromosome centers are rather distinguished by higher gene density and a bias towards older gene classes, lower recombination, less genetic diversity, and lower repeat content as opposed to the chromosome arms [45]. These patterns are thought to be generated by higher levels of background selection in the centers [51]. It has to be noted, that the distinction between arms and centers is not universal across nematode chromosomes and it has been shown that chromosome fusions with associated shifts in the recombinational landscapes can drastically alter intra-chromosomal structures [47]. To test for the presence of center-like regions in the *R. inermis* genome, we calculated the coding density (percentage of protein

coding sequence), repeat content and genetic diversity in 100-kb windows (Fig. 3C). This revealed some central peaks of high coding density that coincided with low repeat content on chromosomes I,II,III, and X. With regard to genetic diversity, the pattern becomes less clear. While the more closely related strain RS6171 also shows lower SNP density in central regions of chromosomes I,II, and X, these patterns appear to be reversed in the more divergent strain RS6371 (Fig. 3C). This could potentially be explained by different recombinational landscapes between the strains and could be tested in future by experimental crosses. Nevertheless, the genetic diversity in the more closely related RS6171 together with the analysis of coding density and repeat content point towards the presence of center-like regions on *R. inermis* chromosomes I, II, and X.

Phylogenomic analysis does not support *R. inermis* as a sister taxon to Diplogastridae

Previous phylogenetic studies suggested a potential sister taxon relationship between *R. inermis* and the family Diplogastridae [26, 27]. To resolve the phylogenetic relationship between *R. inermis* and diplogastrid nematodes, we compiled a phylogenomic data set of annotated proteins and transcriptomes from 40 species [41, 42, 51–69]. We employed a previously established phylogenomic pipeline for reconstructing a species tree from this type of data [70]. This resulted in a concatenated alignment of 1402 single-copy orthologous genes that was taken to compute a maximum-likelihood tree (Fig. 4, see *Methods*). This phylogeny groups *R. inermis* within a clade of rhabditids including *C. elegans* and *M. belari*. Thus, this grouping did not support the position of *R. inermis* as sister group to the diplogastrids. Importantly, the genera *Poikilolaimus* and *Bunonema* are excluded from this clade and appear as next outgroups. In order to explore potential reasons for the misplacement of *R. inermis* in previous phylogenies, we reanalyzed available nematode

sequences for the RNApolII and the 18S and 28S ribosomal RNA markers. Of all three markers, only the 18S sequences supported a sister group relationship (Additional file 1, Fig. S3). This suggests that previous phylogenetic analysis is reproducible and that the 18S sequence largely contributed to the previous misplacement of *R. inermis*.

Comparative genomic analysis reveals parallel gene family expansions across multiple nematode lineages

To further characterize the gene set of *R. inermis*, we performed an orthology analysis with *C. elegans*, *P. pacificus*, *M. belari*, *H. contortus*, and *O. tipulae*. The program Orthofinder identified 17,464 orthogroups with a minimum size of two genes. 6135 orthogroups have at least one ortholog in all six species (Fig. 5A). The presence of orthologs in all six species is the most abundant pattern among all orthogroups. Orthogroups reflecting species-specific absence or presence represent the next most frequent patterns (Fig. 5A). This could be expected since none of the species are closely related and the terminal branches in the underlying species tree are relatively long. As the minimum size of an orthogroup is two, species-specific orthogroups can be interpreted as lineage specific duplication events. Thus, the number of 1752 orthogroups that are specific to the *R. inermis* branch translates to 5985 *R. inermis* genes. We tried to further characterize this gene set by screening for overrepresented protein domains. This revealed a drastic expansion of genes with a BTB domain and an associated Kelch domain (Fig. 5B). These domains are typically found in adapter proteins regulating protein degradation. In *P. pacificus*, BTB domain containing proteins were found to be enriched in the germline [71] and multiple studies reported a loss of BTB genes following the transition to hermaphrodites [42, 62, 70, 72]. The expansion of this gene family in the *R. inermis* lineage apparently occurred in parallel to the expansions in

the lineages leading to *C. elegans* and *Panagrellus redivivus* [73, 74]. This could imply that the expansion of the BTB gene family may be a general pattern in nematodes. To test if even more lineages share the gene expansions of BTBs and other gene families, we repeated this analysis with all other sets of species-specific orthogroups. This revealed 22 protein domains that show significant enrichments in at least two lineages (Bonferroni corrected P-value < 0.05, Fig. 5C). While no gene family is significantly enriched in all six species, G-protein coupled receptors (GPCRS) show up in five lineages. Furthermore BTB, F-box, and C-type lectins show significant enrichments in four lineages. This suggests that multiple gene families underwent parallel expansions in different nematode lineages.

Discussion

Free-living nematodes like *C. elegans* and *P. pacificus* are not only powerful model systems for genetic studies, they are also an excellent subject for evolutionary genomic studies as their species richness allows us to test various hypotheses across different clades and multiple time-scales [51, 72]. Two major factors are essential for such studies. First, deep taxon sampling is needed to facilitate comparisons at high phylogenetic resolution. Over the last two decades, continued sampling efforts have increased the number of known species in the genera *Caenorhabditis* and *Pristionchus* from 10 and 18, respectively to over 50 for both genera [59, 75–77]. Second, we need a robust phylogenetic framework to correctly infer the evolutionary history of genomic features. Abundant genomic and transcriptomic resources have helped to resolve various ambiguities about the phylogeny of the nematode phylum [29, 78, 79]. In this context, the sequencing of the *R. inermis* genome was needed to clarify the question about the sister group to the family Diplogastridae. Our phylogenomic

analysis does not support *R. inermis* as sister to Diplogastridae and re-analysis of previous data suggests that the misplacement of *R. inermis* is largely due to phylogenetic signals coming from the 18S rRNA sequence. Although the phylogenetic position of *R. inermis* makes its genome not particularly interesting for studies focusing on *P. pacificus*, we demonstrate its value for detecting parallel gene family expansions across multiple lineages. It should be noted that additional analysis using more deeply sampled taxa is needed to pinpoint the exact timing of the expansions. Thus, a more highly resolved study in diplogastrid nematodes dated the origin of many BTB orthologs to the ancestor of the *Pristionchus* genus [80]. Given that GPCRs are expressed in chemosensory neurons [81], C-type lectins were shown to carry out important functions in the immune response [82], and also F-box and BTB proteins are also thought be involved in the response to pathogens [73], all identified gene families seem to be involved in the response to environmental stimuli. Thus, it is tempting to speculate that the dynamic evolution of these gene families reflects adaptation to new environments. *R. inermis* has a tight association with the burying beetle *N. vespilloides.* Note that *Rhabditoides* is the only nematode genus known to be reliably associated with burying beetles. Our genome data suggests that this ecological interaction also involves a complex microbial community containing bacteria and fungi. Currently, we do not know which of the microbes forms a component of the nematode's gut microbiome and whether there are any synergistic effects between different species [83]. Thus, the *R. inermis* associated microbial community might be an interesting system to investigate complex interplay between nematodes, bacteria, and fungi.

Conclusions

The main motivation of this study was the identification of the potential sister group of the nematode family Diplogastridae that has long been debated. Previous genomic studies remained inconclusive because not all known taxa had available genome drafts. With this study on *Rhabditoides* one knowledge gap is filled. This work, and the new phylogeny provided in here, supports a sister group relationship between the monophyletic Diplogastridae and a group of genera of Rhabditidae including *C. elegans* and *R. inermis*. In contrast to previous phylogenetic analysis, no single genus or species shows a sister group relationship with Diplogastridae and instead, a sub-clade of rhabditids with substantial morphological and ecological diversification appears as a sister taxon. While the future identification and characterization of additional, currently unknown taxa, might change this phylogeny, the new pattern is consistent with the massive radiation and ecological diversification of two major nematode lineages in soil and associated habitats.

Methods

Nematode culturing and preparation of sequencing libraries

Nematodes of *R. inermis* were grown on standard nematode growth medium as for *C. elegans* and *P. pacificus* [84]. Worms were washed off of 50 fully grown plates using M9 buffer, filtered by using a 5µm filter in order to remove remaining bacteria and then gently

pelleted by centrifugation (1,300 g, 1min). Pelleted worms were frozen in liquid nitrogen and ground to a fine powder using a mortar and pestle. The powder was directly transferred into the lysis buffer used in combination with Qiagen genomic tip columns (500/G) (Qiagen, Hamburg, Germany). The protocol was performed following the manufacturer's instructions. All steps involving sample vortexing were replaced by sample inverting to limit unwanted DNA shearing. DNA quality and quantity were determined with a NanoDrop ND 1000 spectrometer (PeqLab, Erlangen, Germany), a Qubit 2.0 Fluorometer (ThermoFisher Scientific, Waltham, USA), and by a Femto pulse system (Agilent, CA, USA). A total of 10 µg genomic DNA was sheared to a target fragment size of 19 kb using a Megaruptor 2 device (Diagenode, Denville, USA). In order to remove small fragments (<8kb), the BluePippin size-selection system was used according to the manufacturer's protocol (P/N 100-286-000-07, Pacific Biosciences, USA). The final library was sequenced on a SMRT cell of a Pacific Biosciences Sequel II instrument following the Magbead loading protocol. RNA-seq data of mixed-stage cultures was generated as described previously. For Hi-C analysis, *R. inermis* nematodes were grown on standard nematode growth medium and washed off with M9 Buffer. 50µl frozen worm pellet was ground with a motor in an Eppendorf tube and used for Hi-C Library preparation following the manufacturer's instructions for small animals (Arima Genomics, California, USA). RNA-seq libraries were prepared from mixed-stage cultures as described previously [70].

Genome assembly

Raw PacBio reads were converted into Hifi reads by the ccs program (version 6.4.0) with –min-rq 0.99 option and then demultiplexed by the lima program (version 2.6.0). The Hifi reads were then assembled by the Canu program (version v1.4) with options:

-pacbio-corrected -genomeSize=400m [85]. Haplotypes were merged by the program Haplomerger2 (version 20180603) after softmasking repeats in the draft assembly with the help of the Red software (version 05/22/2015) [86, 87]. Hi-C data was aligned against the haplomerged assembly by BWA mem (version 0.7.17-r1188) and scaffolded using the yahs tool (version 1.2a.2) [35, 88]. For metagenomic classification, we annotated scaffolds with Prokka (version 1.14.6) and performed DIAMOND (version 2.1.4.158) searches against the NCBI NR database (downloaded August 2023) [89, 90]. The taxonomic distributions of the best hits were summarized using the taxonomizer library in R. Based on manual inspection of Hi-C data, GC content, coverage data, and additional BLASTN searches against the NCBI database, we classified the scaffolds into 213 candidates that presumably are of nematode origin and a subset of scaffolds that represents the microbial community on which the worms were grown. At this step, we relabeled the five largest scaffolds as chromosomes (ChrI-ChrX) and manually extracted some regions from scaffolds 1 and 4 that showed some Hi-C signal and merged them as a separate scaffold (ChrUn). Karyotyping analysis was performed as described previously [42]. In order to visualize the repeat content across the *R. inermis* genome, the software Red was run to detect repetitive sequences in the final assembly.

Gene annotation

A *de novo* transcriptome assembly was assembled from mixed-stage RNA-seq data using trinity (version 2.2.0) with the normalize_reads option [91]. This transcriptome assembly and the current *P. pacificus* gene annotations (version El Paco gene annotations 3 [92]) were used to generate a set of evidence-based gene annotations using the PPCAC pipeline [93]. In short, transcribed open reading frames and proteins were mapped against the *R. inermis*

genome by the exonerate protein2genome tool (version 2.2.0) with options: --bestn 2 --dnawordlen 20 --maxintron 20000 [94]. Subsequently, per locus (100-bp window) and strand, the gene model with the longest ORF was chosen as representative gene model. We then applied a filtering step to remove spurious annotations that likely result from falsely called ORFs in our transcriptome data. Such spurious annotations will result in species-specific orphan genes (SSOGs) without protein homology in other genomes that are located in the antisense direction of conserved genes, i.e. genes with homologs in other species [80, 92]. In short, *R. inermis* SSOGs were defined by blastp searches against the protein sets of *C. elegans*, *O. tipulae*, *P. pacificus*, *H. contortus*, and *M. belari* (e-value < 0.001) and SSOGs overlapping conserved genes on the antisense strand were removed from the annotation. This resulted in 23,442 gene models with a BUSCO (version 3) completeness score of 85.9% (complete + duplicated, against the odb9 nematode data set) [38, 39]. Coding densities, i.e. the percentage of protein-coding sequences within 100-kb windows were extracted from the resulting annotation files. For coloring the Nigon elements, best reciprocal BLAST hits were computed by the get_BRH.pl of the PPCAC pipeline using the pairwise BLAST output files between *R. inermis* and *P. pacificus* (e-value <0.001).

Population genomic analysis

Raw sequencing data for the *R. inermis* strains RS6171 and RS6371 were aligned against the *R. inermis* genome by the BWA mem alignment program (version 0.7.17-r1188). Single nucleotide variants (SNVs) calls were generated by mpileup and vcfutils.pl varFilter (options: -w0 -D1000) commands of the samtools software suite (version 0.1.8). Only variants with a quality score of at least 20 were taken for visualization of genetic diversity across the *R. inermis* genome.

Phylogenomic species tree

For most species, we obtained protein-coding gene predictions from WormBase ParaSite (Table S1) [95]. For genes, with multiple annotated isoforms, we used the longest protein as a representative sequence. For other species, we generate protein predictions from available transcriptome data. In such cases, either transcriptome assemblies were downloaded from the European Nucleotide Archive or were generated by the Trinity assembler (version 2.2.0 with –normalize_reads option). Subsequently, the longest partial or complete open reading frame was called per sequence and potential redundant sequences resulting from alternative splicing were removed by running the cd-hit clustering program (version 4.3 with default 90% sequence identity threshold). For the reconstruction of a species tree, we followed a similar procedure as described previously [70]. In short, orthologous groups were calculated by the orthagogue [96]. Orthogroups with at most one sequence per species were taken to generate multiple sequence alignments using the MUSCLE program (version 3.8.1551). The concatenated alignments were taken as input to the RAxML software (version 8.2.12) which reconstructed a maximum-likelihood tree (options: -m PROTGAMMAILG -f a) [97].

Orthology and protein domain enrichment

We ran the OrthoFinder pipeline (version 2.5.2) [98] on the *R. inermis*, *M. belari*, *P. pacificus*, *C. elegans*, *O. tipulae*, and *H. contortus* data sets (Additional File 1, Table S1) and extracted sets of orthogroups with specific absence and presence patterns from the Orthogroups output folder. To test for overrepresented protein domains among species-specific orthogroups, we scanned for protein domains with the hmmsearch program (Version 3.3

with -R 0.001 option) using the Pfam-A database (version 3.3) [99, 100]. Tests for overrepresentation were performed by Fisher's exact-test with Bonferroni correction for multiple testing.

Gene tree analysis

For each tested marker (ITS and LSU from fungi and RNApolII, 18S, 28S from nematodes), we downloaded corresponding sequences for representative species from NCBI and generated a multiple sequence alignment with MUSCLE (version 3.8.1551) [101]. Phylogenetic trees were calculated using the R phangorn package [102]. Specifically, the modelTest function was run on each individual data set and a final maximum likelihood tree was calculated using bootstrap.pml function selecting the best model based on the Bayes information criterion.

Abbreviations

ITS: internal transcribed spacer; **LSU:** large subunit ribosomal RNA**; SNV:** Single nucleotide variant; **SSOGs:** Species-specific orphan gene; **SSU:** small subunit ribosomal RNA

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The genome assembly and sequencing reads have been uploaded to the European Nucleotide archive under the study accession PRJEB71643. The fungal genome of the genus *Vanrija* has been deposited at DDBJ/ENA/GenBank under the accession JAZAQS000000000.

Competing interests

The author declares that he has no competing interests.

Funding

This work was funded by the Max Planck Society. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors' contributions

Conceptualization, R.J.S.; Investigation, C.R., W.R., M.A., S.W., M.H. and R.J.S.; Data curation, C. R.; Visualization, M.A., S.W. and C. R.; Writing original draft, C.R.; Writing – review & editing, C.R. and R.J.S; Project administration, C.R. and R.J.S.; Supervision, C.R and R.J.S.

Acknowledgements

We would like to thank the whole Sommer lab for helpful discussions.

Supplementary Information

Additional file 1

Fig. S1: Coverage of microbial scaffolds in different *R. inermis* isolates.

Fig. S2. Phylogenetic relationships between different *Vanrija* species.

Fig. S3. Phylogenetic relationships between different nematodes.

Table S1 Nematode genomic and transcriptomic data sets.

References

1. van den Hoogen J, Geisen S, Routh D, Ferris H, Traunspurger W, Wardle DA, et al. Soil nematode abundance and functional group composition at a global scale. Nature. 2019;572:194–8.

2. Shih P-Y, Lee JS, Shinya R, Kanzaki N, Pires-daSilva A, Badroos JM, et al. Newly Identified Nematodes from Mono Lake Exhibit Extreme Arsenic Resistance. Curr Biol. 2019;29:3339–44.e4.

3. Blaxter ML, De Ley P, Garey JR, Liu LX, Scheldeman P, Vierstraete A, et al. A molecular evolutionary framework for the phylum Nematoda. Nature. 1998;392:71–5.

4. Sommer RJ, Carta LK, Kim S-Y, Sternberg PW. Morphological, genetic and molecular description of Pristionchus pacificus sp. n.(Nematoda: Neodiplogasteridae). Fundam Appl Nematol. 1996;19:511–22.

5. Bento G, Ogawa A, Sommer RJ. Co-option of the hormone-signalling module dafachronic acid-DAF-12 in nematode evolution. Nature. 2010;466:494–7.

6. Ragsdale EJ, Müller MR, Rödelsperger C, Sommer RJ. A developmental switch coupled to the evolution of plasticity acts through a sulfatase. Cell. 2013;155:922–33.

7. Lightfoot JW, Wilecki M, Rödelsperger C, Moreno E, Susoy V, Witte H, et al. Small peptide-mediated self-recognition prevents cannibalism in predatory nematodes. Science. 2019;364:86–9.

8. Kieninger MR, Ivers NA, Rödelsperger C, Markov GV, Sommer RJ, Ragsdale EJ. The Nuclear Hormone Receptor NHR-40 Acts Downstream of the Sulfatase EUD-1 as Part of a Developmental Plasticity Switch in Pristionchus. Curr Biol. 2016;26:2174–9.

9. Bui LT, Ivers NA, Ragsdale EJ. A sulfotransferase dosage-dependently regulates mouthpart polyphenism in the nematode Pristionchus pacificus. Nat Commun. 2018;9:4119.

10. Sieriebriennikov B, Sun S, Lightfoot JW, Witte H, Moreno E, Rödelsperger C, et al. Conserved nuclear hormone receptors controlling a novel plastic trait target fast-evolving genes expressed in a single cell. PLoS Genet. 2020;16:e1008687.

11. Ishita Y, Onodera A, Ekino T, Chihara T, Okumura M. Co-option of an Astacin Metalloprotease Is Associated with an Evolutionarily Novel Feeding Morphology in a Predatory Nematode. Mol Biol Evol. 2023;40.

12. Susoy V, Ragsdale EJ, Kanzaki N, Sommer RJ. Rapid diversification associated with a macroevolutionary pulse of developmental plasticity. Elife. 2015;4.

13. Sommer RJ. Phenotypic Plasticity: From Theory and Genetics to Current and Future Challenges. Genetics. 2020;215:1–13.

14. Levis NA, Ragsdale EJ. Linking Molecular Mechanisms and Evolutionary Consequences of Resource Polyphenism. Front Integr Neurosci. 2022;16:805061.

15. Kanzaki N, Ragsdale EJ, Giblin-Davis RM. Revision of the paraphyletic genus Koerneria Meyl, 1960 and resurrection of two other genera of Diplogastridae (Nematoda). Zookeys. 2014;:17–30.

16. Mahboob M, Tahseen Q. Diversity, prevalence and microhabitat specificity of nematodes (Rhabditidae Örley, 1880 and Diplogastridae Micoletzky, 1922) associated with insects: an overview. Int J Pest Manage. https://doi.org/10.1080/09670874.2021.1969470

17. Hodda M. Phylum Nematoda: a classification, catalogue and index of valid genera, with a census of valid species. Zootaxa. 2022;5114:1–289.

18. von Lieven AF, Uni S, Ueda K, Barbuto M, Bain O. Cutidiplogaster manati n. gen., n. sp. (Nematoda: Diplogastridae) from skin lesions of a West Indian manatee (Sirenia) from the Okinawa Churaumi Aquarium. Nematology. 2011;13:51–9.

19. Kanzaki N, Ragsdale EJ, Herrmann M, Mayer WE, Tanaka R, Sommer RJ. Parapristionchus giblindavisi n. gen., n. sp. (Rhabditida: Diplogastridae) isolated from stag beetles (Coleoptera: Lucanidae) in Japan. Nematology. 2012;14:933–47.

20. Herrmann M, Ragsdale EJ, Kanzaki N, Sommer RJ. Sudhausia aristotokia n. gen., n. sp. and S. crassa n. gen., n. sp. (Nematoda: Diplogastridae): viviparous new species with precocious gonad development. Nematology. 2013;15:1001–20.

21. Kanzaki N, Ragsdale EJ, Susoy V, Sommer RJ. Leptojacobus dorci n. gen., n. sp. (Nematoda:

Diplogastridae), an Associate of Dorcus Stag Beetles (Coleoptera: Lucanidae). J Nematol. 2014;46:50–9.

22. Ragsdale EJ, Kanzaki N, Sommer RJ. Levipalatum texanum n. gen., n. sp. (Nematoda: Diplogastridae), an androdioecious species from the south-eastern USA. Nematology. 2014;16:695–709.

23. Kanzaki N, Ikeda Y, Shinya R. Onthodiplogaster japonica n. gen., n. sp. (Rhabditida: Diplogastridae) isolated from Onthophagus sp. (Coleoptera: Scarabaeidae) from Japan. Sci Rep. 2023;13:6470.

24. von Lieven AF. The sister group of the Diplogastrina (Nematoda). Russ J Nematol. 2002;10:127–38.

25. Meldal BHM, Debenham NJ, De Ley P, De Ley IT, Vanfleteren JR, Vierstraete AR, et al. An improved molecular phylogeny of the Nematoda with special emphasis on marine taxa. Mol Phylogenet Evol. 2007;42:622–36.

26. Kiontke K, Barrière A, Kolotuev I, Podbilewicz B, Sommer R, Fitch DHA, et al. Trends, stasis, and drift in the evolution of nematode vulva development. Curr Biol. 2007;17:1925–37.

27. van Megen H, van den Elsen S, Holterman M, Karssen G, Mooyman P, Bongers T, Holovachov O, Bakker J, Helder J. A phylogenetic tree of nematodes based on about 1200 full-length small subunit ribosomal DNA sequences. Nematology. 2009;11**:** 927-950.

28. Sudhaus W. An update of the catalogue of paraphyletic “Rhabditidae” (Nematoda) after eleven years. Soil Org. 2023;95:95–116.

29. Ahmed M, Roberts NG, Adediran F, Smythe AB, Kocot KM, Holovachov O. Phylogenomic Analysis of the Phylum Nematoda: Conflicts and Congruences With Morphology, 18S rRNA, and Mitogenomes. Frontiers in Ecology and Evolution. 2022;9:769565.

30. Sudhaus W. Vergleichende Untersuchungen zur Phylogenie, Systematik, Ökologie, Biologie und Ethologie der Rhabditidae (Nematoda). E. Schweizerbart; 1976;125:1-229.

31. Wang Y, Rozen DE. Gut microbiota in the burying beetle, Nicrophorus vespilloides, provide colonization resistance against larval bacterial pathogens. Ecol Evol. 2018;8:1646–54.

32. Scott MP. The ecology and behavior of burying beetles. Annu Rev Entomol. 1998;43:595–618.

33. Herrmann M, Mayer WE, Sommer RJ. Nematodes of the genus Pristionchus are closely associated with scarab beetles and the Colorado potato beetle in Western Europe. Zoology . 2006;109:96–108.

34. Huang S, Kang M, Xu A. HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. Bioinformatics. 2017;33:2577–9.

35. Zhou C, McCarthy SA, Durbin R. YaHS: yet another Hi-C scaffolding tool. Bioinformatics. 2023;39:btac808.

36. Meng L, Liu C, Zhu L, Wang X, Li C. Draft genome sequence of yeast Vanrija sp. strain TS01, isolated from leukemia patient’s urine. Microbiol Resour Announc. 2023;12:e0015223.

37. Imanishi D, Abe K, Kera Y, Takahashi S. Draft Genome Sequence of the Yeast Vanrija humicola (Formerly Cryptococcus humicola) Strain UJ1, a Producer of d-Aspartate Oxidase. Genome Announc. 2018;6:e00068-18.

38. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.

39. Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simão FA, Ioannidis P, et al. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. Nucleic Acids Res. 2017;45:D744–9.

40. Gorte O, Aliyu H, Neumann A, Ochsenreither K. Draft Genome Sequence of the Oleaginous Yeast Apiotrichum porosum (syn. Trichosporon porosum) DSM 27194. J Genomics. 2019;7:11–3.

41. Röseler W, Collenberg M, Yoshida K, Lanz C, Sommer RJ, Rödelsperger C. The improved genome of the nematode Parapristionchus giblindavisi provides insights into lineage-specific gene family evolution. G3 . 2022;12:jkac215.

42. Wighard SS, Athanasouli M, Witte H, Rödelsperger C, Sommer RJ. A New Hope: A Hermaphroditic Nematode Enables Analysis of a Recent Whole Genome Duplication Event. Genome Biol Evol. 2022;14:evac169.

43. Tandonnet S, Koutsovoulos GD, Adams S, Cloarec D, Parihar M, Blaxter ML, et al. Chromosome-Wide Evolution and Sex Determination in the Three-Sexed Nematode Auanema rhodensis. G3 . 2019;9:1211–30.

44. Gonzalez de la Rosa PM, Thomson M, Trivedi U, Tracey A, Tandonnet S, Blaxter M. A telomere-to-telomere assembly of Oscheius tipulae and the evolution of rhabditid nematode chromosomes. G3 . 2021;11:jkaa020.

45. Rödelsperger C, Meyer JM, Prabh N, Lanz C, Bemm F, Sommer RJ. Single-Molecule Sequencing Reveals the Chromosome-Scale Genomic Architecture of the Nematode Model Organism Pristionchus pacificus. Cell Rep. 2017;21:834–44.

46. Cotton JA, Bennuru S, Grote A, Harsha B, Tracey A, Beech R, et al. The genome of Onchocerca volvulus, agent of river blindness. Nat Microbiol. 2016;2:16216.

47. Yoshida K, Rödelsperger C, Röseler W, Riebesell M, Sun S, Kikuchi T, et al. Chromosome fusions repatterned recombination rate and facilitated reproductive isolation during Pristionchus nematode speciation. Nat Ecol Evol. 2023;7:424–39.

48. The C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science. 1998;282:2012–8.

49. Albertson DG, Thomson JN. Segregation of holocentric chromosomes at meiosis in the nematode, Caenorhabditis elegans. Chromosome Res. 1993;1:15–26.

50. Rillo-Bohn R, Adilardi R, Mitros T, Avşaroğlu B, Stevens L, Köhler S, et al. Analysis of meiosis in Pristionchus pacificus reveals plasticity in homolog pairing and synapsis in the nematode lineage. Elife. 2021;10.

51. Prabh N, Roeseler W, Witte H, Eberhardt G, Sommer RJ, Rödelsperger C. Deep taxon sampling reveals the evolutionary dynamics of novel gene families in Pristionchus nematodes. Genome Res. 2018;28:1664–74.

52. Stevens L, Félix M-A, Beltran T, Braendle C, Caurcel C, Fausett S, et al. Comparative genomics of 10 new Caenorhabditis species. Evol Lett. 2019;3:217–36.

53. Casasa S, Biddle JF, Koutsovoulos GD, Ragsdale EJ. Polyphenism of a Novel Trait Integrated Rapidly Evolving Genes into Ancestrally Plastic Networks. Mol Biol Evol. 2021;38:331–43.

54. Schwarz EM, Hu Y, Antoshechkin I, Miller MM, Sternberg PW, Aroian RV. The genome and transcriptome of the zoonotic hookworm Ancylostoma ceylanicum identify infection-specific gene families. Nat Genet. 2015;47:416–22.

55. Kikuchi T, Cotton JA, Dalzell JJ, Hasegawa K, Kanzaki N, McVeigh P, et al. Genomic insights into the origin of parasitism in the emerging plant pathogen Bursaphelenchus xylophilus. PLoS Pathog. 2011;7:e1002219.

56. Mitreva M, Jasmer DP, Zarlenga DS, Wang Z, Abubucker S, Martin J, et al. The draft genome of the parasitic nematode Trichinella spiralis. Nat Genet. 2011;43:228–35.

57. Fierst JL, Willis JH, Thomas CG, Wang W, Reynolds RM, Ahearne TE, et al. Reproductive Mode and the Evolution of Genome Size and Structure in Caenorhabditis Nematodes. PLoS Genet. 2015;11:e1005323.

58. Ghedin E, Wang S, Spiro D, Caler E, Zhao Q, Crabtree J, et al. Draft genome of the filarial nematode parasite Brugia malayi. Science. 2007;317:1756–60.

59. Stevens L, Rooke S, Falzon LC, Machuka EM, Momanyi K, Murungi MK, et al. The Genome of Caenorhabditis bovis. Curr Biol. 2020;30:1023–31.e4.

60. Doyle SR, Tracey A, Laing R, Holroyd N, Bartley D, Bazant W, et al. Genomic and transcriptomic variation defines the chromosome-scale assembly of Haemonchus contortus, a model gastrointestinal worm. Commun Biol. 2020;3:656.

61. Kanzaki N, Tsai IJ, Tanaka R, Hunt VL, Liu D, Tsuyama K, et al. Biology and genome of a newly discovered sibling species of Caenorhabditis elegans. Nat Commun. 2018;9:3216.

62. Yin D, Schwarz EM, Thomas CG, Felde RL, Korf IF, Cutter AD, et al. Rapid genome shrinkage in a self-fertile nematode reveals sperm competition proteins. Science. 2018;359:55–61.

63. Besnard F, Koutsovoulos G, Dieudonné S, Blaxter M, Félix M-A. Toward Universal Forward Genetics: Using a Draft Genome Sequence of the Nematode Oscheius tipulae To Identify Mutations Affecting Vulva Development. Genetics. 2017;206:1747–61.

64. Guo F, Slos D, Du H, Li K, Li H, Qing X. Transcriptomics of Cruznema velatum (Nematoda: Rhabditidae) with a redescription of the species. J Helminthol. 2023;97:e57.

65. Vadnal J, Granger OG, Ratnappan R, Eleftherianos I, O'Halloran DM, Hawdon JM. Refined ab initio gene predictions of Heterorhabditis bacteriophora using RNA-seq. Int J Parasitol. 2018;48:585–90.

66. Beltran T, Barroso C, Birkle TY, Stevens L, Schwartz HT, Sternberg PW, et al. Comparative Epigenomics Reveals that RNA Polymerase II Pausing and Chromatin Domain Organization Control Nematode piRNA Biogenesis. Dev Cell. 2019;48:793–810.e6.

67. Lee RYN, Howe KL, Harris TW, Arnaboldi V, Cain S, Chan J, et al. WormBase 2017: molting into a new stage. Nucleic Acids Res. 2018;46:D869–74.

68. Slos D, Sudhaus W, Stevens L, Bert W, Blaxter M. Caenorhabditis monodelphis sp. n.: defining the stem morphology and genomics of the genus Caenorhabditis. BMC Zoology. 2017;2:1–15.

69. Grosmaire M, Launay C, Siegwald M, Brugière T, Estrada-Virrueta L, Berger D, et al. Males as somatic investment in a parthenogenetic nematode. Science. 2019;363:1210–3.

70. Rödelsperger C, Röseler W, Prabh N, Yoshida K, Weiler C, Herrmann M, et al. Phylotranscriptomics of Pristionchus Nematodes Reveals Parallel Gene Loss in Six Hermaphroditic Lineages. Curr Biol. 2018;28:3123–7.e5.

71. Athanasouli M, Akduman N, Röseler W, Theam P, Rödelsperger C. Thousands of Pristionchus pacificus orphan genes were integrated into developmental networks that respond to diverse environmental microbiota. PLoS Genet. 2023;19:e1010832.

72. Adams PE, Eggers VK, Millwood JD, Sutton JM, Pienaar J, Fierst JL. Genome Size Changes by Duplication, Divergence, and Insertion in Caenorhabditis Worms. Mol Biol Evol. 2023;40:msad039.

73. Thomas JH. Adaptive evolution in two large families of ubiquitin-ligase adapters in nematodes and plants. Genome Res. 2006;16:1017–30.

74. Srinivasan J, Dillman AR, Macchietto MG, Heikkinen L, Lakso M, Fracchia KM, et al. The draft genome and transcriptome of Panagrellus redivivus are shaped by the harsh demands of a free-living lifestyle. Genetics. 2013;193:1279–95.

75. Kiontke K, Fitch DHA. The phylogenetic relationships of Caenorhabditis and other rhabditids. WormBook. 2005; doi: 10.1895/wormbook.1.11.1.

76. Mayer WE, Herrmann M, Sommer RJ. Phylogeny of the nematode genus Pristionchus and implications for biodiversity, biogeography and the evolution of hermaphroditism. BMC Evol Biol. 2007;7:104.

77. Kanzaki N, Herrmann M, Weiler C, Röseler W, Theska T, Berger J, et al. Nine new Pristionchus (Nematoda: Diplogastridae) species from China. Zootaxa. 2021;4943:zootaxa.4943.1.1.

78. Smythe AB, Holovachov O, Kocot KM. Improved phylogenomic sampling of free-living nematodes enhances resolution of higher-level nematode phylogeny. BMC Evol Biol. 2019;19:121.

79. Lee Y-C, Ke H-M, Liu Y-C, Lee H-H, Wang M-C, Tseng Y-C, et al. Single-worm long-read sequencing reveals genome diversity in free-living nematodes. Nucleic Acids Res. 2023;51:8035–47.

80. Prabh N, Rödelsperger C. Multiple Pristionchus pacificus genomes reveal distinct evolutionary dynamics between de novo candidates and duplicated genes. Genome Res. 2022;32:1315–27.

81. Bargmann CI. Chemosensation in C. elegans. WormBook. 2006; doi: 10.1895/wormbook.1.123.1.

82. Pees B, Yang W, Kloock A, Petersen C, Peters L, Fan L, et al. Effector and regulator: Diverse functions of C. elegans C-type lectin-like domain proteins. PLoS Pathog. 2021;17:e1009454.

83. Lo W-S, Han Z, Witte H, Röseler W, Sommer RJ. Synergistic interaction of gut microbiota enhances the growth of nematode through neuroendocrine signaling. Curr Biol. 2022;32:2037–50.e4.

84. Pires-daSilva A. Pristionchus pacificus protocols. WormBook. 2013; doi: 10.1895/wormbook.1.114.2.

85. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27:722–36.

86. Girgis HZ. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. BMC Bioinformatics. 2015;16:227.

87. Huang S, Chen Z, Huang G, Yu T, Yang P, Li J, et al. HaploMerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. Genome Res. 2012;22:1581–8.

88. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics. 2010;26:589–95.

89. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9.

90. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021;18:366–8.

91. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52.

92. Athanasouli M, Witte H, Weiler C, Loschko T, Eberhardt G, Sommer RJ, et al. Comparative genomics and community curation further improve gene annotations in the nematode Pristionchus pacificus. BMC Genomics. 2020;21:708.

93. Rödelsperger C. The community-curated Pristionchus pacificus genome facilitates automated gene annotation improvement in related nematodes. BMC Genomics. 2021;22:216.

94. Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005;6:31.

95. Howe KL, Bolt BJ, Shafie M, Kersey P, Berriman M. WormBase ParaSite - a comprehensive resource for helminth genomics. Mol Biochem Parasitol. 2017;215:2–10.

96. Ekseth OK, Kuiper M, Mironov V. orthAgogue: an agile tool for the rapid prediction of orthology relations. Bioinformatics. 2014;30:734–6.

97. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006;22:2688–90.

98. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16:157.

99. Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer EL. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. Nucleic Acids Res. 1999;27:260–2.

100. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011;7:e1002195.

101. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004;5:113.

102. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27:592–3.

Figure legends

Fig. 1. Genome assembly of a complex microbial community. A Visualization of GC content and Hi-C coverage indicates that the five largest scaffolds appear to be different from the smaller ones. **B** The Hi-C contact map of the 22 largest scaffolds shows a background level of Hi-C signal across the five largest scaffolds but not between the smaller scaffolds. **C** For scaffolds 6-22 we performed homology searches against the NCBI nr database. The pie charts summarize the taxonomic distribution for the best hits at the genus level.

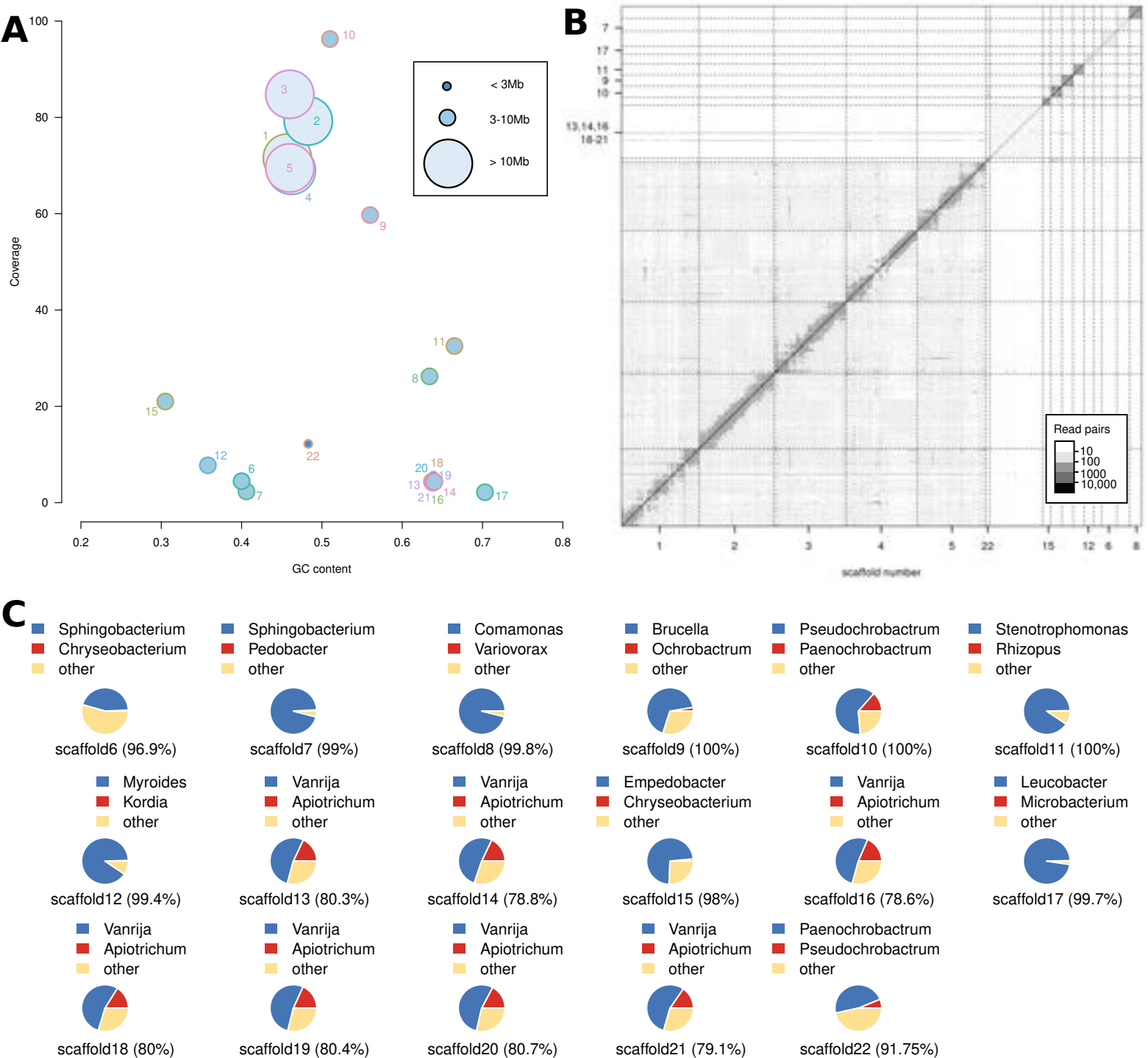
Fig. 2. Phylogenomics of fungal genomes from the order Trichosporonales. The phylogeny represents a Maximum-likelihood tree calculated from a concatenated alignment of orthologous proteins. The barplots show the values of various features across the phylogenomic data set.

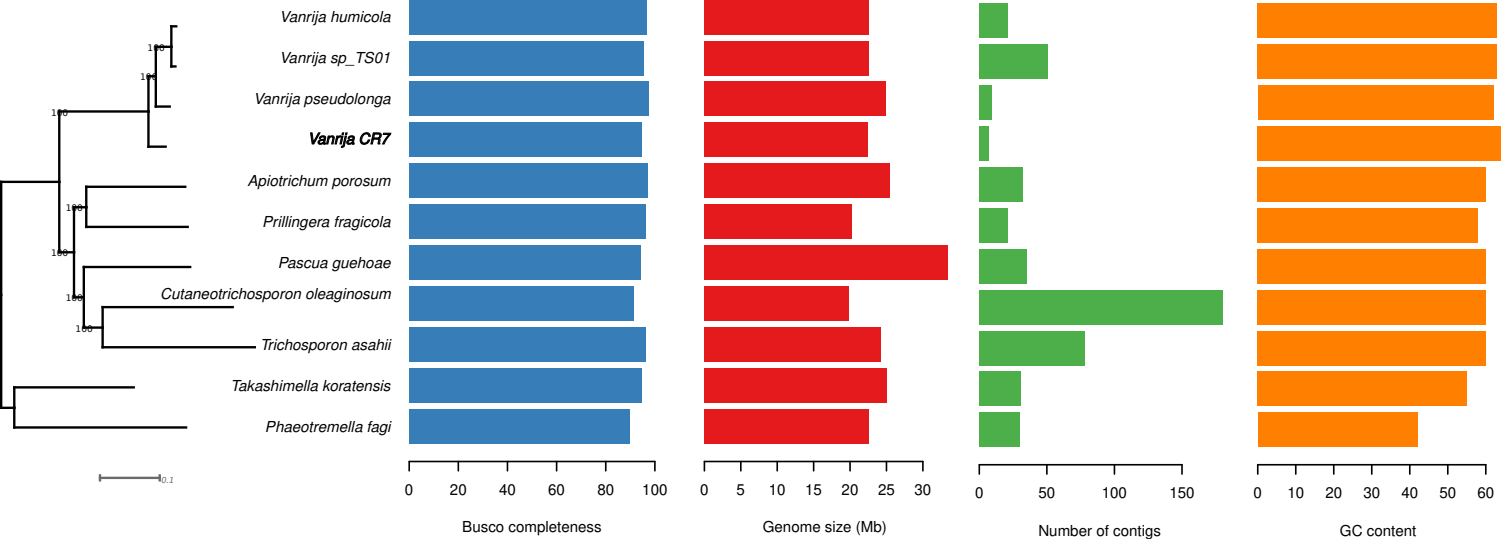
Fig. 3. Chromosomal configuration of *R. inermis*. A Karyotyping of *R. inermis* worms identified oocytes with five chromosomes. **B** The bars represent 500-kb windows along the five largest scaffolds/chromosomes that show the distribution of 1-1 orthologs color-coded by their chromosomal location in the *P. pacificus* genome which can be used as a proxy to indicate the Nigon elements. **C** The plots show the distribution of coding density, repeat content, and genetic diversity across the *R. inermis* genome. Features were calculated in non-overlapping 100-kb windows and averaged over 20 windows.

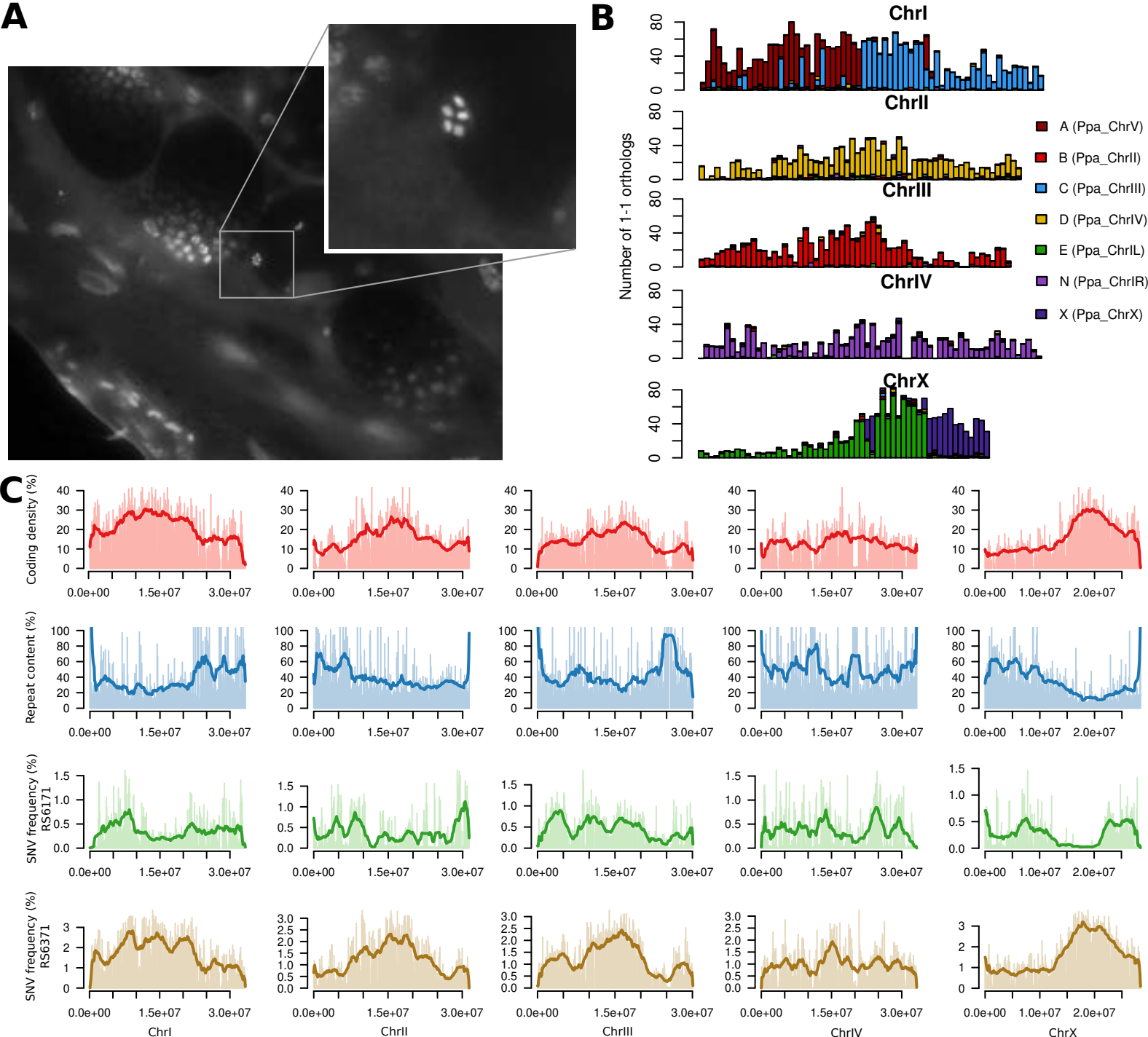
Fig. 4. Phylogenomics of nematodes of the suborder Rhabditina. Phylogenomic and transcriptomic data of diplogastrid and rhabditid nematodes were used to reconstruct a maximum-likelihood tree for the suborder Rhabditina. Note that the family Rhabditidae is

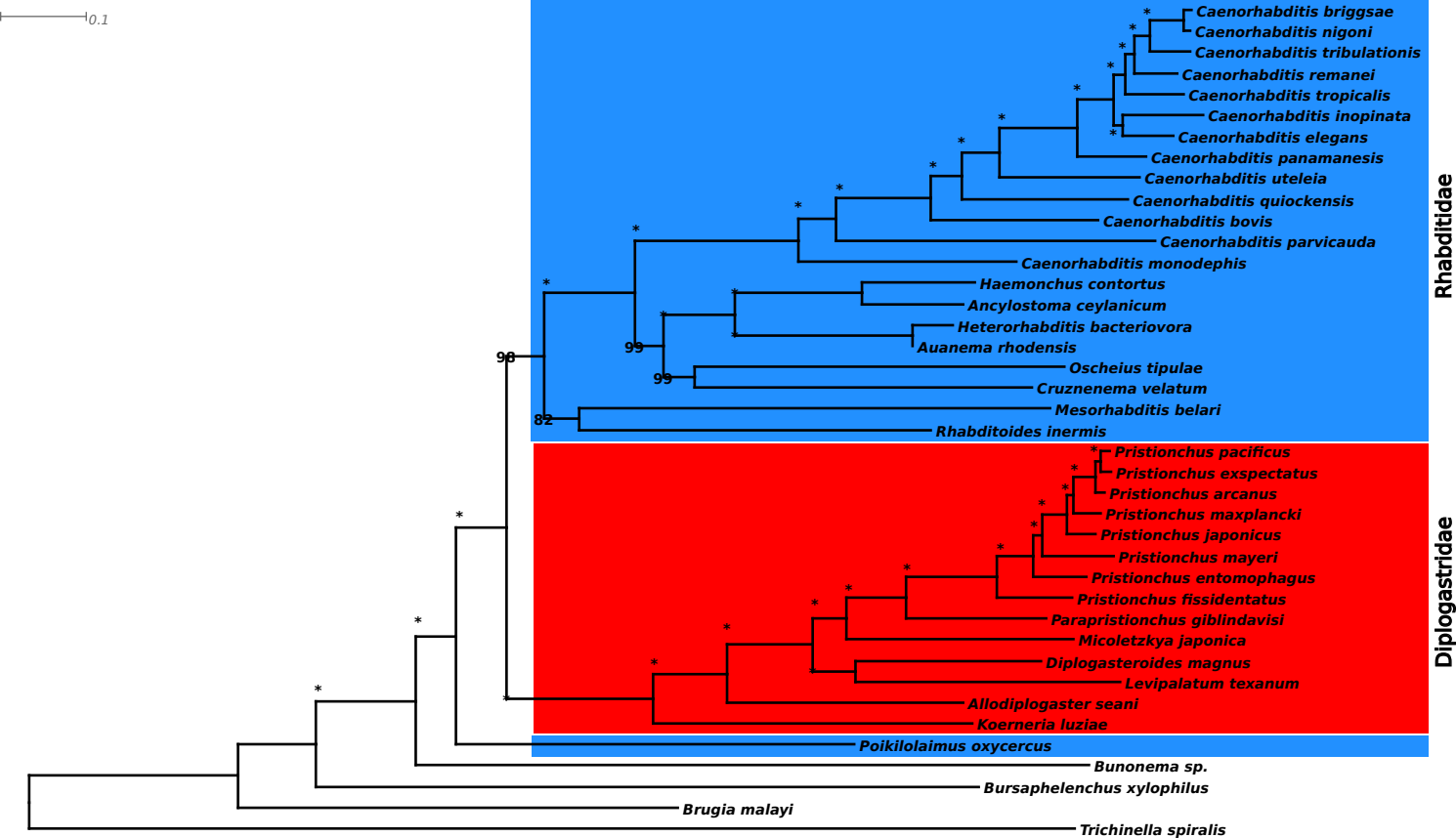
known to be paraphyletic [25]. The species *B. xylophilus*, *B. malayi,* and *T. spiralis* were included as outgroups. Stars indicate a bootstrap support of 100/100.

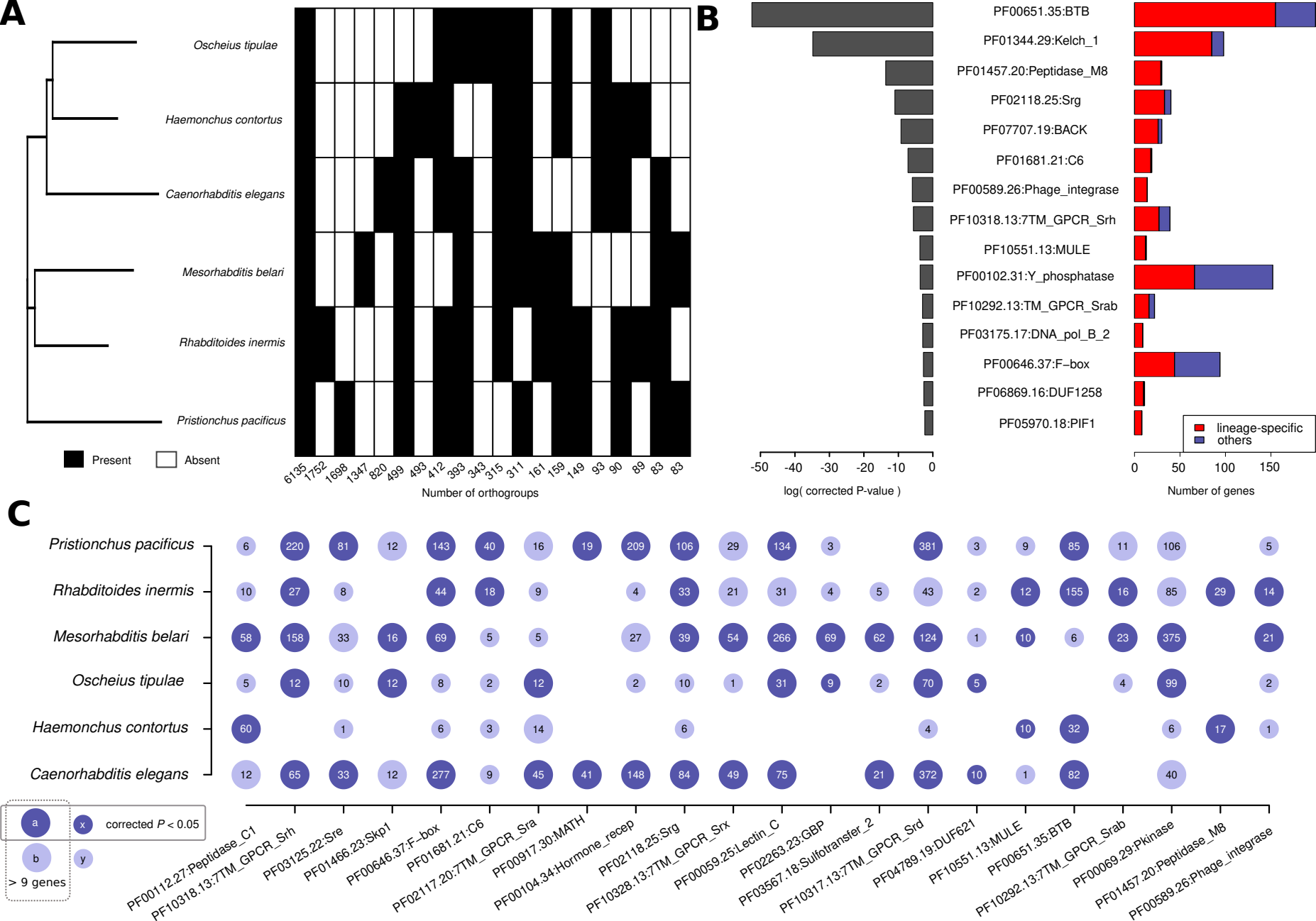
Fig. 5 Parallel gene family expansions across multiple nematode lineages. A Orthogroups were computed by the OrthoFinder pipeline and classes of orthogroups were defined based on the absence/presence patterns and were sorted by abundance. Orthology classes are visualized as a heatmap next to the phylogeny. With exception of the first class of 6135 orthogroups that are shared between all species, the most abundant patterns mostly reflect lineage-specific patterns such as species-specific orthogroups or orthogroup that are only missing in a single data set. **B** The barplots indicate the significance of enrichment and the total number of *R. inermis* genes with protein domains. **C** We repeated the protein domain enrichment analysis for all species-specific orthogroups. The bubble plot visualizes the significance and the number of genes in species-specific orthogroups per domain and species.


A
Coverage
GC content
< 3Mb
3-10Mb
> 10Mb
B
Read pairs
10
100
1000
10,000
scaffold number
C
scaffold6 (96.9%)
Sphingobacterium
Chryseobacterium
other
scaffold7 (99%)
Sphingobacterium
Pedobacter
other
scaffold8 (99.8%)
Comamonas
Variovorax
other
scaffold9 (100%)
Brucella
Ochrobactrum
other
scaffold10 (100%)
Pseudochrobactrum
Paenochrobactrum
other
scaffold11 (100%)
Stenotrophomonas
Rhizopus
other
scaffold12 (99.4%)
Myroides
Kordia
other
scaffold13 (80.3%)
Vanrija
Apiotrichum
other
scaffold14 (78.8%)
Vanrija
Apiotrichum
other
scaffold15 (98%)
Empedobacter
Chryseobacterium
other
scaffold16 (78.6%)
Vanrija
Apiotrichum
other
scaffold17 (99.7%)
Leucobacter
Microbacterium
other
scaffold18 (80%)
Vanrija
Apiotrichum
other
scaffold19 (80.4%)
Vanrija
Apiotrichum
other
scaffold20 (80.7%)
Vanrija
Apiotrichum
other
scaffold21 (79.1%)
Vanrija
Apiotrichum
other
scaffold22 (91.75%)
Paenochrobactrum
Pseudochrobactrum
other


Vanrija humicola
Vanrija sp_TS01
Vanrija pseudolonga
Vanrija CR7
Apiotrichum porosum
Prillingera fragicola
Pascua guehoae
Cutaneotrichosporon oleaginosum
Trichosporon asahii
Takashimella koratensis
Phaeotremella fagi
100
100
100
100
100
100
0.1
0 20 40 60 80 100
Busco completeness
0 5 10 15 20 25 30
Genome size (Mb)
0 50 100 150
Number of contigs
0 10 20 30 40 50 60
GC content


A
B
ChrI
ChrII
ChrIII
ChrIV
ChrX
Number of 1-1 orthologs
A (Ppa_ChrV)
B (Ppa_ChrII)
C (Ppa_ChrIII)
D (Ppa_ChrIV)
E (Ppa_ChrIL)
N (Ppa_ChrIR)
X (Ppa_ChrX)
C
Coding density (%)
Repeat content (%)
SNV frequency (%) RS6171
SNV frequency (%) RS6371
ChrI
ChrII
ChrIII
ChrIV
ChrX


0.1
Caenorhabditis briggsae
Caenorhabditis nigoni
Caenorhabditis tribulationis
Caenorhabditis remanei
Caenorhabditis tropicalis
Caenorhabditis inopinata
Caenorhabditis elegans
Caenorhabditis panamanesis
Caenorhabditis uteleia
Caenorhabditis quiockensis
Caenorhabditis bovis
Caenorhabditis parvicauda
Caenorhabditis monodephis
Haemonchus contortus
Ancylostoma ceylanicum
Heterorhabditis bacteriovora
Auanema rhodensis
Oscheius tipulae
Cruznenema velatum
Mesorhabditis belari
Rhabditoides inermis
Rhabditidae
Pristionchus pacificus
Pristionchus exspectatus
Pristionchus arcanus
Pristionchus maxplancki
Pristionchus japonicus
Pristionchus mayeri
Pristionchus entomophagus
Pristionchus fissidentatus
Parapristionchus giblindavisi
Micoletzkya japonica
Diplogasteroides magnus
Levipalatum texanum
Allodiplogaster seani
Koerneria luziae
Diplogastridae
Poikilolaimus oxycercus
Bunonema sp.
Bursaphelenchus xylophilus
Brugia malayi
Trichinella spiralis
98
99
99
82


A
Oscheius tipulae
Haemonchus contortus
Caenorhabditis elegans
Mesorhabditis belari
Rhabditoides inermis
Pristionchus pacificus
Present
Absent
6135 1752 1698 1347 820 499 493 412 393 343 315 311 161 159 149 93 90 89 83 83
Number of orthogroups
B
PF00651.35:BTB
PF01344.29:Kelch_1
PF01457.20:Peptidase_M8
PF02118.25:Srg
PF07707.19:BACK
PF01681.21:C6
PF00589.26:Phage_integrase
PF10318.13:7TM_GPCR_Srh
PF10551.13:MULE
PF00102.31:Y_phosphatase
PF10292.13:TM_GPCR_Srab
PF03175.17:DNA_pol_B_2
PF00646.37:F-box
PF06869.16:DUF1258
PF05970.18:PIF1
-50 -40 -30 -20 -10 0
log(corrected P-value)
lineage-specific
others
0 50 100 150
Number of genes
C
Pristionchus pacificus 6 220 81 12 143 40 16 19 209 106 29 134 3 381 3 9 85 11 106 5
Rhabditoides inermis 10 27 8 44 18 9 4 33 21 31 4 5 43 2 12 155 16 85 29 14
Mesorhabditis belari 58 158 33 16 69 5 5 27 39 54 266 69 62 124 1 10 6 23 375 21
Oscheius tipulae 5 12 10 12 8 2 12 2 10 1 31 9 2 70 5 4 99 2
Haemonchus contortus 60 1 6 3 14 6 4 10 32 6 17 1
Caenorhabditis elegans 12 65 33 12 277 9 45 41 148 84 49 75 21 372 10 1 82 40
a
x
corrected P < 0.05
b
y
> 9 genes
PF00112.27:Peptidase_C1
PF10318.13:7TM_GPCR_Srh
PF03125.22:Sre
PF01466.23:Skp1
PF00646.37:F-box
PF01681.21:C6
PF02117.20:7TM_GPCR_Sra
PF00917.30:MATH
PF00104.34:Hormone_recep
PF02118.25:Srg
PF10328.13:7TM_GPCR_Srx
PF00059.25:Lectin_C
PF02263.23:GBP
PF03567.18:Sulfotransfer_2
PF10317.13:7TM_GPCR_Srd
PF04789.19:DUF621
PF10551.13:MULE
PF00651.35:BTB
PF10292.13:7TM_GPCR_Srab
PF00069.29:Pkinase
PF01457.20:Peptidase_M8
PF00589.26:Phage_integrase

Supplementary Material to

The genome assembly of *Rhabditoides inermis* from a complex microbial community reveals further evidence for parallel gene family expansions across multiple nematodes

Christian Rödelsperger[1,*], Waltraud Röseler[1], Marina Athanasouli[1], Sara Wighard[1,2], Matthias Herrmann[1], Ralf J. Sommer[1,*]

[1] Department for Integrative Evolutionary Biology, Max Planck Institute for Biology, Max-Planck-Ring 9, 72076 Tübingen, Germany

[2] Current address: Institute of Molecular Biotechnology of the Austrian Academy of Sciences (IMBA), Vienna BioCenter (VBC), 1030 Vienna, Austria.

[*] Corresponding authors:

christian.roedelsperger@tuebingen.mpg.de

ralf.sommer@tuebingen.mpg.de

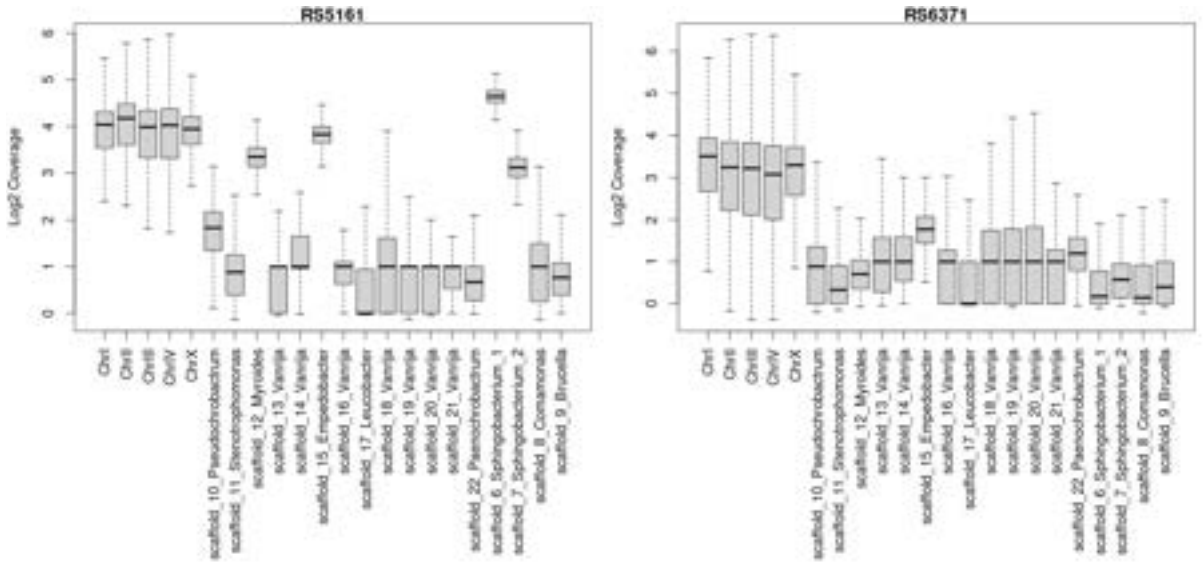


Fig. S1. Coverage of microbial scaffolds in different *R. inermis* isolates. Coverage analysis reveals evidence for the presence of all microbes in sequencing data of two independent isolates of *R. inermis*. The y-axis indicates the average Log2 coverage in 1-kb windows for the largest scaffolds/chromosomes.

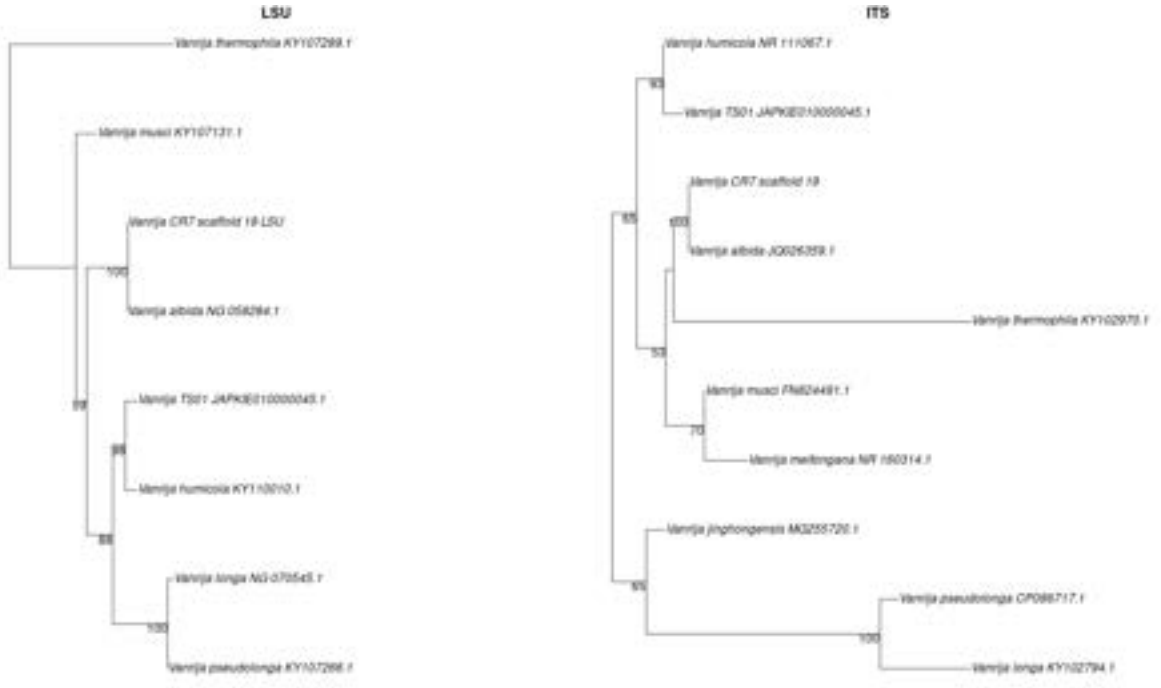


Fig. S2. Phylogenetic relationships between different *Vanrija* species. LSU and ITS sequences from different *Vanrija* species were downloaded from NCBI Genbank, aligned with MUSCLE, and maximum likelihood trees were calculated by the phangorn R package. For both markers, sequences from our genome assembly (CR7) group together with sequences from *V. albida*.

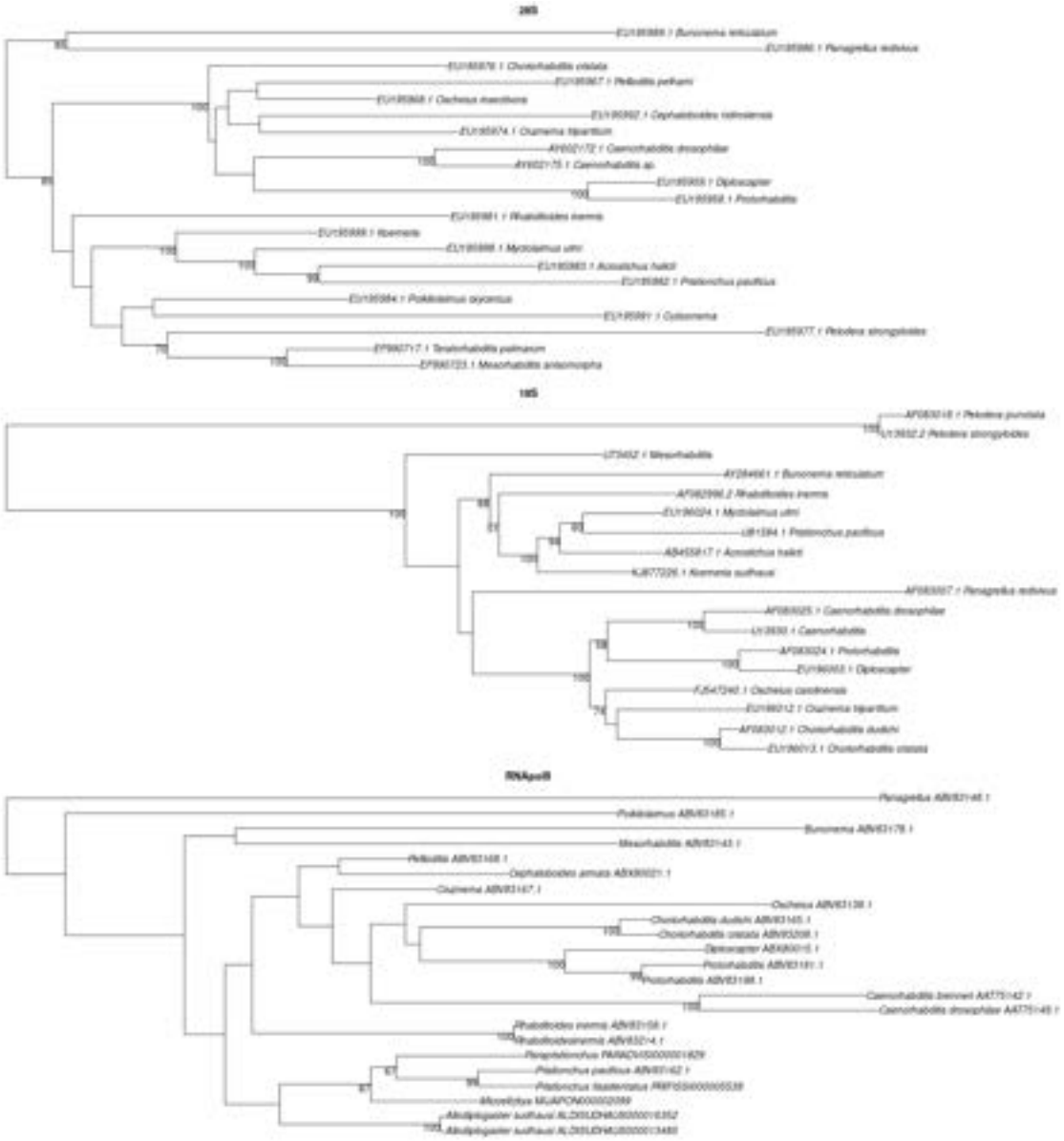


Fig. S3. Phylogenetic relationships between different nematodes. Phylogenetic trees were reconstructed for RNApolII, 18S and 28S ribosomal RNAs from different nematodes. Only the 18S data suggests a sister group relationship between R. inermis and the family Diplogastridae.

Table S1 - Nematode genomic and transcriptomic data sets

Species	**Source accession / version**	**Type**	**Reference**
Allodiplogaster seani	European Nucleotide Archive: HCAZ01000000	RNA	Wighard et al. 2022

Ancylostoma ceylanicum	WormBase ParaSite WBPS18 PRJNA231479	Protein	Schwarz et al. 2015
Auanema rhodensis	European Nucleotide Archive: ERR3150287	RNA	Tandonnet et al. 2019
Bunoema sp.	European Nucleotide Archive: GITZ01000000	RNA	Casasa et al. 2021
Brugia malayi	WormBase ParaSite WBPS14 PRJNA10729	Protein	Ghedin et al. 2007
Bursaphelenchelus xlyophilus	WormBase ParaSite WBPS14 PRJEA64437	Protein	Kikuchi et al. 2011
Caenorhabditis bovis	WormBase ParaSite WBPS18 PRJEB34497	Protein	Stevens et al. 2019
Caenorhabditis briggsae	WormBase ParaSite WBPS14 PRJNA10731	Protein	-
Caenorhabditis elegans	WormBase ParaSite WBPS14 PRJNA13758	Protein	-
Caenorhabditis inopinata	WormBase ParaSite WBPS18 PRJDB5687	Protein	Kanzaki et al. 2018
Caenorhabditis monodelphis	caenorhabditis.org	Protein	Slos et al. 2017
Caenorhabditis nigoni	WormBase ParaSite WBPS18 PRJNA384657	Protein	Yin et al. 2018
Caenorhabditis panamensis	WormBase ParaSite WBPS18 PRJEB28259	Protein	Stevens et al. 2019
Caenorhabditis parvicauda	WormBase ParaSite WBPS18 PRJEB12595	Protein	Stevens et al. 2019
Caenorhabditis quiockensis	WormBase ParaSite WBPS18 PRJEB11354	Protein	Stevens et al. 2019
Caenorhabditis remanei	WormBase ParaSite WBPS18 PRJNA248911	Protein	Fierst et al. 2015
Caenorhabditis tribulationis	WormBase ParaSite WBPS18 PRJEB12608	Protein	Stevens et al. 2019
Caenorhabditis tropicalis	WormBase ParaSite WBPS18 PRJNA53597	Protein	-
Caenorhabditis uteleia	WormBase ParaSite WBPS18 PRJEB12600	Protein	Stevens et al. 2019
Cruznema velatum	European Nucleotide Archive: SRR23934717	RNA	Guo et al. 2023
Diplogasteroides magnus	European Nucleotide Archive: GITX01000000	RNA	Casasa et al. 2021

Haemonchus contortus	WormBase ParaSite WBPS14 PRJEB506	Protein	Doyle et al. 2020
Heterorhabditis bacteriovora	European Nucleotide Archive: SRR6294669	RNA	Vadnal et al. 2018
Koerneria luziae	European Nucleotide Archive: GIUA01000000	RNA	Casasa et al. 2021
Levipalatum texanum	European Nucleotide Archive: GITY01000000	RNA	Casasa et al. 2021
Micoletzkya japonica	pristionchus.org	Protein	Prabh et al. 2018
Oscheius tipulae	WormBase ParaSite WBPS14 PRJEB15512	Protein	Besnard et al. 2017
Poikilolaimus oxycercus	European Nucleotide Archive: SRR6049087	Protein	Beltran et al. 2018
Pristionchus arcanus	pristionchus.org PPCAC (version 1)	Protein	Prabh et al. 2018
Pristionchus entomophagus	pristionchus.org PPCAC (version 1)	Protein	Prabh et al. 2018
Pristionchus exspectatus	pristionchus.org Yoshida et al.	Protein	Yoshida et al. 2023
Pristionchus fissidentatus	pristionchus.org PPCAC (version 1)	Protein	Prabh et al. 2018
Pristionchus japonicus	pristionchus.org PPCAC (version 1)	Protein	Prabh et al. 2018
Pristionchus mayeri	pristionchus.org PPCAC (version 1)	Protein	Prabh et al. 2018
Pristionchus maxplancki	pristionchus.org PPCAC (version 1)	Protein	Prabh et al. 2018
Pristionchus pacificus	pristionchus.org, El Paco gene annotation 3	Protein	Athanasouli et al. 2020
Parapristionchus giblindavisi	pristionchus.org (version 2, 2022)	Protein	Röseler et al. 2022
Rhabditoides inermis	this study	Protein	this study
Trichinella spiralis	WormBase ParaSite WBPS14 PRJNA12603	Protein	Mitreva et al. 2011

High nutritional conditions influence feeding plasticity in *Pristionchus pacificus* and render worms non-predatory

Veysi Piskobulu[1], Marina Athanasouli[1], Hanh Witte[1], Christian Feldhaus[2], Adrian Streit[1] & Ralf J. Sommer[1,*]

[1] Max-Planck Institute for Biology Tübingen, Department for Integrative Evolutionary Biology, Max-Planck Ring 9, D-72076 Tübingen, Germany

[2] Max-Planck Institute for Biology Tübingen, BioOptics Facility, Max-Planck Ring 5, D-72076 Tübingen, Germany

*author for correspondence at: ralf.sommer@tuebingen.mpg.de

Abstract

Developmental plasticity, the ability of a genotype to produce different phenotypes in response to environmental conditions, has been subject to intense studies in the last four decades. The self-fertilizing nematode *Pristionchus pacificus* has been developed as a genetic model system for studying developmental plasticity due to its mouth-form polyphenism that results in alternative feeding strategies with a facultative predatory and a non-predatory mouth form. Many studies linked molecular aspects of the regulation of mouth-form polyphenism with investigations of its evolutionary and ecological significance. Also, several environmental factors influencing *P. pacificus* feeding structure expression were identified including temperature, culture condition and population density. However, the nutritional plasticity of the mouth form has never been properly investigated although polyphenisms are known to be influenced by changes in nutritional conditions. For instance, studies in eusocial insects and scarab beetles have provided significant mechanistic insights into the nutritional regulation of polyphenisms but also other forms of plasticity. Here, we study the influence of nutrition on mouth-form polyphenism in *P. pacificus* through experiments with monosaccharide and fatty acid supplementation. We show that in particular glucose supplementation renders worms non-predatory. Subsequent transcriptomic and mutant analyses indicate that *de novo* fatty acid synthesis and peroxisomal beta-oxidation pathways play an important role in the mediation of this plastic response. Finally, the analysis of fitness consequences through fecundity counts suggests that non-predatory animals have an advantage over predatory animals grown in the glucose-supplemented condition.

Keywords: *Pristionchus pacificus*, developmental plasticity, polyphenism, nutrition, glucose, fatty acids

Research highlights

This study represents the first systematic attempt to investigate the influence of nutrition on mouth-form polyphenism in the genetic model organism *Pristionchus pacificus*. Through glucose and oleic acid supplementation we show that high nutritional conditions influence feeding plasticity and render worms non-predatory. Mutant analysis indicates a role of *de novo* fatty acid synthesis and peroxisomal beta-oxidation pathways for these responses.

Introduction

Developmental plasticity allows organisms to alter their developmental trajectory to execute distinct phenotypes in response to environmental cues with polyphenisms representing the most remarkable form of plasticity by exhibiting environment-sensitive alternative phenotypes without intermediate forms (West-Eberhard, 2003). In general, developmentally plastic responses can be observed at physiological, morphological, and behavioural levels, including life history traits of organisms. Many case studies have shown developmental plasticity to be ubiquitous in nature. It has been suggested to facilitate adaptation to novel or heterogenous environments and to play an important role for evolutionary diversification and the creation of novelty (Ghalambor et al., 2007; Pigliucci, 2001; West-Eberhard, 2003). Multiple environmental factors have been shown to influence polyphenisms including temperature, seasonal changes, and population density (see reviews; Simpson et al., 2011; Yang & Pospisilik, 2019). Specifically, nutrition is an important regulator of polyphenisms. Several studies in insects have highlighted the importance of nutritional conditions during development for the generation of discrete phenotypes, revealing genetic and epigenetic mechanisms involved (Casasa et al., 2020; Chandra et al., 2018; Gotoh et al., 2014; Kamakura, 2011; Kucharski et al., 2008). For instance, caste polyphenism in honeybee and horn polyphenism in dung beetles represent some of the extreme cases of nutrition-induced developmental plasticity (Emlen, 1997; Haydak, 1970; Moczek, 1998). Nonetheless, our understanding of how nutrition affects the expression of polyphenic traits in other taxa remains limited.

Besides insects, nematodes provided substantial mechanistic insights into the regulation of polyphenisms in the last decade. In particular, the free-living hermaphrodite *Pristionchus pacificus* has been a model system for the studies of developmental plasticity. Research in *P. pacificus* has provided insights into the genetic and epigenetic regulation as well as the ecological and evolutionary significance of plasticity (Schroeder, 2021; Sommer, 2020). In nature, *P. pacificus* can be found in soil and on scarab beetles (chafers, stag beetles, dung beetles), which, after their death, provide a nutritious environment for the nematode through proliferation of bacteria and fungi on the carcass (Herrmann et al., 2006; Meyer et al., 2017; Ragsdale et al., 2015; Renahan et al., 2021; Renahan & Sommer, 2022). In contrast to *Caenorhabditis elegans* and most other free-living nematodes, *P. pacificus* shows a remarkable mouth-form dimorphism, which influences its dietary niche and behaviour. During postembryonic development, genetically identical *P. pacificus* worms form one of two, alternative morphs, the so-called eurystomatous (Eu) and stenostomatous (St) mouth forms (Figure 1a) (Bento et al., 2010). While the Eu morph enables facultative predatory feeding on other nematodes by exhibiting two “teeth” and a wide mouth, the St morph leads to strict microbial feeding and forms only one tooth and a narrow mouth (Theska et al., 2020). Major

components of the gene regulatory network that govern mouth-form plasticity have been identified in this model (Bui et al., 2018; Casasa et al., 2023; Levis & Ragsdale, 2023; Namdeo et al., 2018; Ragsdale et al., 2013; Serobyan et al., 2016; Sieriebriennikov et al., 2018, 2020; Werner et al., 2023). For instance, a supergene locus with the sulfatase *eud-1* and two alpha-N-acetylglucosaminidase genes (*nag-1* and *nag-2*), as well as the sulfotransferase, *seud-1*/*sult-1*, act as part of a main switch in mouth-form determination (Bui et al., 2018; Namdeo et al., 2018; Ragsdale et al., 2013; Sieriebriennikov et al., 2018). They are involved in the regulation of the evolutionarily conserved nuclear hormone receptors *nhr-40* and *nhr-1* that have been co-opted for the regulation of mouth-form plasticity and act downstream of *eud-1* and *seud-1/sult-1* (Sieriebriennikov et al., 2020; Theska & Sommer, 2024). It is important to note that the understanding of the molecular regulation of feeding structure polyphenism is far from completion with many epigenetic aspects still to be identified (Brown et al., 2024; Werner et al., 2023).

Mouth-form plasticity in *P. pacificus* is influenced by several environmental factors such as culture condition, population density, and temperature (Lenuzzi et al., 2023; Werner et al., 2017, 2018). Previous studies have also revealed prominent roles for nematode metabolites (e.g. ascarosides) and dietary bacteria in the mouth-form decision and predatory feeding (Akduman et al., 2020; Bose et al., 2012; Dardiry et al., 2023). In contrast, our knowledge of how changes to the nutritional status of worms affect mouth-form plasticity is still scarce. Bento et al. (2010) have shown that starvation promotes the development of the Eu morph. More recent genetic work by Casasa and co-workers identified the Mediator subunit MDT15/MED15 to regulate mouth-form polyphenism, which is known to have a conserved role in metabolic responses to nutritional stress (Casasa et al., 2023). These observations suggests that mouth-form plasticity is sensitive to nutritional conditions in particular to low nutrition. However, mouth-form and fitness consequences of a high nutritional condition have never been investigated in *P. pacificus*. One approach of studying the effect of nutrition on nematode plasticity is to introduce supplements, such as monosaccharides and fatty acids, into the standard diet to increase fat storage in nematodes, which is a strong alteration to the nutritional status. This aided research in *C. elegans* to discover novel functions of molecular factors involved in lipid metabolism and associated biological processes (Alcántar-Fernández et al., 2018; Han et al., 2017; Lee et al., 2015; Nomura et al., 2010; Wan et al., 2022). Note that while nutrient supplementation is well established in *C. elegans*, the role of direct and indirect effects of supplementation through the bacterial diet cannot be disentangled. Very likely, the effects observed in such experiments are due to both direct and indirect uptake of nutrients. Moreover, nutrient supplementation studies have prioritised investigating lipid storage as an indicator of nutritional status. Nematodes lack an adipose tissue for lipid storage; however, they generate lipid droplets in many cells including the intestine to serve this function. In

addition, storage lipids can be visualised through staining methods such as Oil Red O (ORO), which is a reliable post-fix technique, providing a proxy for lipid storage when it is quantitatively measured (O'Rourke et al., 2009).

Here, we study the influence of nutrition on mouth-form plasticity in *P. pacificus* by establishing experimental setups in which worms are grown in "high nutrition" dietary conditions that promote fat storage. These conditions are obtained through supplementation of monosaccharides and fatty acids (mainly glucose and oleic acid) into the nematode growth medium (NGM) (Figure 1b). We also examine the lipid storage in worms via ORO staining and imaging analysis. We show that fat storage-promoting conditions, in particular glucose supplementation, render worms non-predatory. Similarly, we test whether these dietary effects are mediated transgenerationally but find no such influence. In addition, we carry out transcriptomic and subsequent mutant analyses to explore associated molecular mechanisms. Results indicate that *de novo* fatty acid synthesis and peroxisomal beta-oxidation pathways, which are involved in storage and utilisation of lipids, are essential for nutrition-induced mouth-form plasticity. Finally, we examine worm fitness by using fecundity as proxy to explore whether the mouth-form response in the glucose-supplemented dietary condition benefit *P. pacificus*. Our findings suggest that non-predatory animals have a fitness advantage over predatory animals grown under the same nutritional condition. Overall, this work establishes nutritional status as an additional factor influencing mouth-form polyphenism and highlights the beneficial advantage of this plastic ability in *P. pacificus*.

Materials and methods

Nematode stock maintenance

All *P. pacificus* strains were cultured on 6cm NGM plates with 300µl *E. coli* OP50 provided as food at 20°C as previously described (Sommer et al., 1996). The wild type strain PS312 was used for experimentation unless otherwise noted. Stocks were maintained by passing three young adult hermaphrodites with eggs every five days. Mutant *P. pacificus* strains were similarly cultured and are available from the SommerLab.

Monosaccharide supplementation experiments

Monosaccharides were purchased from Sigma-Aldrich as follows: glucose (D-(+)-Dextrose, product number: D9434), galactose (D-(+)-Galactose, product number: G0750), fructose (D-(−)-Fructose, product number: F0127). All monosaccharides were dissolved in purified and autoclaved water to obtain 1M stock concentrations. Stock solutions were filtered through 0.22µm sterile filters before storage. Plate pouring was carried out under a sterile hood. Each monosaccharide stock was diluted into liquid NGM media to obtain desired final concentrations. Monosaccharide-added NGM was then stirred for approximately 1 minute to obtain a homogenous media. Media was then poured into 6cm plates using a sterile 10ml serological pipette, obtaining a volume of 9ml in each plate. Control plates with the same volume were poured from the main NGM media without monosaccharides. Plates were allowed to dry for 2 days and were then seeded with 300µl overnight grown *E. coli* OP50. After two days, three young hermaphrodites with eggs (from stocks grown under standard conditions) were added onto each plate for each strain, condition, and replicate. After adding worms, plates were incubated at 20°C until F1 adults emerged. Plates were then used for mouth-form phenotyping and ORO staining.

Fatty acid supplementation experiments

Fatty acid supplementation was carried out as described by Deline et al. (2013); however, cholesterol was added into the NGM media after autoclaving. Sodium oleate (Sigma-Aldrich, product number: O7501) and sodium linoleate (Sigma-Aldrich, product number: L8134) were used for oleic acid and linoleic acid supplementation, respectively. For each experiment, a 0.1M fatty acid working stock was prepared fresh in purified and autoclaved water. Once completely dissolved, the fatty acid working stocks were purged with nitrogen gas to prevent oxidation by air. Plate pouring was carried out as in monosaccharide supplementation except the volume of the NGM in each plate was 8ml. Both supplemented and control plates contained 0.1% Tergitol (NP-40, Sigma-Aldrich, product number: NP40S) to facilitate fatty acid absorption. Once poured, plates were covered with a box to avoid light oxidation. The same

steps as in monosaccharide supplementation were followed for seeding the plates with *E. coli* OP50, culturing nematodes, mouth-form phenotyping and ORO staining.

Mouth-form phenotyping

Mouth-form phenotypes were scored using Zeiss Discovery V.20 stereomicroscope based on mouth-width of individual worms as previously described (Ragsdale et al., 2013; Theska et al., 2020). Each plate was taken as a biological replicate and 30 worms were scored for their mouth-form, unless otherwise mentioned. The percentage of Eu animals of each plate was then calculated. Mouth-form percentage (Eu%) is graphed as a mean of percentages of all plates for each strain in a particular dietary condition.

Transgenerational experiments

For testing potential transgenerational effects, lines were established by picking three young (day 1) adult worms with eggs from standard condition onto oleic acid, glucose, and control plates, respectively. Ten lines were established for each condition and each line was propagated by transferring three young adult worms for five generations on the same condition. In case of contamination, the line was discontinued and not used for counting mouth form or reversal. At every generation, 6-7 lines (among the already established 10 lines) were selected and reverted back to standard culture conditions. For each condition, mouth-form phenotypes were scored for all plates at every generation and reversal.

Oil Red O staining

Post-fix ORO staining method was modified from Li et al. (2016). Oil Red O stock solution was prepared by dissolving 0.5g ORO (Sigma-Aldrich, product number: O0625) in 100ml isopropanol. This stock solution was shaken on a see-saw rocker for several days. To prepare ORO working solution, required amount was taken from the stock and centrifuged for 5 minutes at 4500rpm to pellet all ORO-related precipitates. Working ORO solution was prepared fresh by diluting precipitate-free ORO (supernatant) in purified water to obtain 60% ORO. This solution was also shaken on a see-saw rocker for 10 minutes and centrifuged for 5 minutes at 4500rpm before staining worms. For worm fixation, a 1% formaldehyde solution was used by diluting 1ml 16% formaldehyde solution (Thermo Scientific, catalogue number: 28906) in 15ml 1X PBS (phosphate buffered saline). For dehydrating worms, 60% isopropanol was prepared fresh by diluting isopropanol in purified water. First, worms were washed off the plates with 1X PBS and collected in a 15ml Falcon tube for each sample. Note that, after mouth-form phenotyping, several biological replicates (plates) were washed into one sample tube for each strain for a particular dietary condition. Once worms formed a sediment, extra volume was removed, leaving only 1ml. This volume containing worms was then transferred

to a new 1.5ml tube. To pellet worms after transfer or wash, samples were briefly centrifuged in a quick-spin centrifuge. Worms were fixed with 1% formaldehyde (1ml) for 30 minutes by shaking on a see-saw rocker. Then, samples were frozen in liquid nitrogen (~8 seconds) and thawed under running tap water, three times in total. Samples were washed with 1X PBS three times and were dehydrated in 60% isopropanol (1ml) for 2 minutes. Finally, samples were stained with 0.5ml 60% ORO working solution for 30 minutes on a see-saw rocker. The staining solution was applied through a 0.22μm sterile filter. After staining, samples were washed three times with 1X PBS. Supernatant of each sample was removed, leaving approximately 0.1ml sample of stained worms. Two drops of Vectashield (Biozol, catalogue number: H-1000) was then added into each sample. Samples were gently mixed by pipetting up and down before mounting onto microscope slides. Samples were then imaged for ORO quantification.

ORO quantification and analysis

Oil Red O quantification was performed as described in Feldhaus and Piskobulu (2024 preprint). Briefly, we imaged worms on an AxioImager.Z1 (Zeiss, Oberkochen, Germany) with a 10x/0.3 objective and an AxioCam 506 mono. For ORO quantification, images were obtained with filter sets for DAPI and Texas Red in transmission mode. The filter-based imaging with a monochrome camera was chosen to remove any potential bias from white balance settings of the camera. Oil Red O absorbance per worm was semi-automatically calculated by manually outlining the worm area and then obtaining the difference between the Texas Red channel (where ORO appears transparent) and the DAPI channel (where ORO absorbs) after correcting for differences in the spectral response for both channels. In total, 15 adult worms were quantified per strain and condition. For each worm, raw integrated density (RawIntDen), from channel 4 of the processed image, was taken as a measure of ORO absorbance. To compare individuals with different body size (e.g. mutants), ORO absorbance (RawIntDen) was normalised by the body area (μm^2) for each worm (RawIntDen/body area). For calculating body area, images of worms from the blue channel (DAPI) were saved in grayscale. The body area was then calculated for each worm by “Threshold” and “Analyze Particles” functions on Fiji (Schindelin et al., 2012). Therefore, the final ORO absorbance value was determined as RawIntDen divided by body area for each worm.

RNA extraction

For RNA extraction, three young adults with eggs, from standard condition, were placed on oleic acid, glucose, and control conditions, respectively. For each condition, we scanned through 5-9 plates and collected 150 F1 young adults without eggs in two separate tubes for RNA extraction (two technical replicates). RNA extraction was carried out using Direct-zol RNA

Microprep Kit (Zymo Research, R2060). Quality of RNA was assessed using NanoDrop. Concentration of RNA was assessed using Qubit 2.0 Fluorometer (Invitrogen). Samples were then diluted to required concentration and sent to NovoGene Co., Ltd for sequencing.

RNA-seq and KEGG enrichment analysis

Raw reads were aligned to the reference *P. pacificus* genome (El Paco) with STAR (version 2.7.1a) and quantified with featureCounts from the Subread R package (version 2.0.1) based on the latest gene annotations (Athanasouli et al., 2020; Dobin et al., 2013; Liao et al., 2014). The count matrix was filtered by removing genes with less than 10 reads, reducing the number of genes to be examined from 28896 to 18174 and 18049 for oleic acid and glucose respectively. Differential gene expression analysis was performed for each of the two conditions with DESeq2 (FDR-corrected $P<0.05$, version 1.42.0) (Love et al., 2014). The differentially expressed genes from each condition were tested for overrepresentation of KEGG pathways using Fisher's exact-test with multiple testing correction (Bonferroni corrected $P<0.05$) in R (version 4.3.1), based on the existing KEGG annotations for the *P. pacificus* genes (Athanasouli et al., 2023).

Phylogenetic tree

All *C. elegans* fatty acid desaturase protein sequences were obtained from WormBase (www.wormbase.org). Longer isoforms were picked for genes with different isoforms. Each *C. elegans* fatty acid desaturase protein was then blasted against *P. pacificus* on the protein database (www.pristionchus.org, El Paco annotation v3, 2020). From each blast result, all *P. pacificus* hits were collected and analysed on InterProScan (www.ebi.ac.uk) for conserved protein domains. We found two genes (PPA03289 and PPA05783) which did not contain desaturase domains, which were discarded. The rest of the protein sequences were aligned by ClustalW on MEGA11 software (Tamura et al., 2021). Based on this alignment, a maximum likelihood phylogenetic tree was constructed with LG model, and 100 bootstrap replications. The phylogenetic tree was visualised and edited on FigTree software (version 1.4.4).

CRISPR/Cas9 mutagenesis and identification of mutants

For the generation of mutants by CRISPR, the protocols of Witte et al. (2015) and Han et al. (2020) were followed with minor modifications. All components of the CRISPR/Cas9 complex were purchased from Integrated DNA Technologies (IDT). Each CRISPR RNA (crRNA) was designed to target 20bp upstream of the protospacer adjacent motif (PAM). To construct CRISPR/Cas9 complex, 5µl CRISPR RNA (crRNA) from 100µM stock was mixed with 5µl trans-activating CRISPR RNA (tracrRNA) obtained from 100µM stock (catalogue number 1072534). This mixture was first denatured at 95°C for 5 minutes, and then cooled down at

room temperature for annealing (5 minutes). Hybridized crRNA:tracrRNA product (5μl) was mixed with 1μl Cas9 protein from 62μM stock (catalogue number: 1081058) and incubated at room temperature for 5 minutes. Tris-EDTA buffer was then added into the CRISPR/Cas9 mix to obtain a final concentration of 18.1μM for crRNA:tracrRNA hybrid and 2.5μM for Cas9. Co-injection marker, with *Ppa-eft-3* promoter and TurboRFP sequences, was also integrated into the injection mix (Han et al., 2020). Injected worms were allowed to lay eggs for 24 hours. Fluorescent marker-positive plates were used to isolate emerging F1 progeny. From each positive plate, 8 -10 F1 worms were isolated. Each F1 worm was allowed to lay eggs and then was lysed in single worm lysis buffer (10mM Tris–HCl at pH 8.3, 50mM KCl, 2.5mM MgCl2, 0.45% NP-40, 0.45% Tween 20, 120μg/ml Proteinase K) in a thermal cycler program with 65°C for 1 hour; 95°C for 10 minutes. Lysates were used to carry out polymerase chain reactions to amplify the target region. Mutants in F1 generation (predominantly heterozygotes) were first identified through Sanger sequencing (GENEWIZ Germany GmbH). Subsequent re-isolation and sequencing of the following generation (F2) allowed capturing homozygous mutants. All crRNAs, and related primers for target genes are provided in Table S1.

Genetic repair of *Ppa-pddl-1(tu2028)*

To genetically revert *Ppa-pddl-1(tu2028)* back to wild type through CRISPR editing, a crRNA (5'- TACTATCCTCTATCCTATAG-3') was designed targeting 20bp upstream of the TGG PAM site at exon 4 based on the mutant sequence including 7bp insertion. A 100bp wild type repair template, covering the target region, was also used (synthesised by IDT). The repair template was as follows: 5'-CATGAATAGTTCTAACCACTTCTTCTCCTTCAGACATTACTATCCTCTAGTGGCCCTTCTC TGTTTCCTCATGCCAACTGTCGTTCCTGTCTATTATTGG-3'. The repair template was integrated into the CRISPR injection mix at a concentration of 10μM. Homozygous wildtype worms were isolated in F2 generation post injection, validated via sanger sequencing.

Fecundity experiment

To study fecundity of Eu and St morphs, young adult worms without eggs (F1) were isolated from glucose-supplemented diets (80mM concentration) based on their mouth morph and subsequently transferred to standard dietary conditions with 20μl *E. coli* OP50. Worms were passed onto new plates for 8 consecutive days and were removed at the end of 8th day. After five days from inoculation, viable progeny was counted on each plate.

Data visualisation and statistical analysis

All plots were generated in RStudio (version 1.2.5042) using ggplot2 and ggpubr packages. Illustrations were produced in Affinity Designer (version 1.9.2). Statistical significance testing

was carried out by comparing two independent groups, e.g., control vs. treatment, mutant vs. wildtype, Eu vs. St, for ORO absorbance and total fecundity. Briefly, data was first tested for normal distribution mostly via Shapiro-Wilk normality test with additional visual observations. In case of normal distribution, either two sample t-test (equal variance) or Welch two sample t-test (unequal variance) were used. In case of nonnormal distribution, Wilcoxon rank sum-test was used as the nonparametric equivalent. All statistics were performed in RStudio (version 1.2.5042).

Results

Monosaccharide and fatty acid supplementations render worms non-predatory in a concentration dependent manner

For all experiments, we studied hermaphrodites of the PS312 strain, the wild type of *P. pacificus* that had its genome being sequenced at chromosome level (Dieterich et al., 2008; Rödelsperger et al., 2017). This strain is highly Eu under standard lab conditions on an *E. coli* OP50 diet with more than 95% of animals expressing the Eu morph. First, we performed pilot experiments with monosaccharides ranging from 20mM to 300mM concentrations. Results revealed a concentration dependent effect on mouth-form plasticity. Specifically, the number of St worms increased with increasing concentrations of monosaccharides (Figure S1a). Among all, we prioritised glucose as the main St-inducing monosaccharide and performed further experiment by using 100mM and 150mM concentrations. Results indicated that 100mM glucose is sufficient to render worms highly St without exerting observable adverse effects (Figure 2a). Note that even under glucose-supplementation conditions, there is still an effect of population density on mouth-form expression as pilot experiments starting from 10 adult worms (Figure S1a) revealed higher Eu ratios than similar glucose concentration experiments with three adult inoculations (Figure 2a). For fatty acid supplementations, we used oleic acid (monounsaturated fatty acid) and linoleic acid (polyunsaturated fatty acid) to test 0.5mM and 1mM concentrations. Similar to monosaccharide supplementations, oleic acid effect on mouth-form plasticity was mediated in a concentration dependent manner (Figure 2a). Observations based on morphology and developmental pace of worms at both concentrations suggested that 0.5mM oleic acid would be more ideal for further experiments. Results also indicated that the same concentrations were detrimental for worms in the linoleic acid condition. While worms did not populate the plates at 1mM, they were developmentally slow and almost completely St at 0.5mM linoleic acid (Figure S1b). Therefore, we performed another experiment for linoleic acid with reduced concentrations ranging from 0.01mM to 0.2mM. We found that 0.2mM linoleic acid renders worms predominantly St, alleviating the detrimental effect observed at 0.5mM concentration (Figure S1c). Overall, these results show that monosaccharides and fatty acid supplementations induce changes of the mouth-form ratio towards St in *P. pacificus*. For further experiments, we selected oleic acid and glucose as main supplements at 0.5mM and 100mM concentrations, respectively.

Oleic acid and glucose diets do not induce a transgenerational effect on mouth-form plasticity

Some dietary and nutritional influences on nematodes are known to be transmitted transgenerationally (Beltran et al., 2020). Therefore, we propagated cultures on oleic acid and

glucose conditions for five generations to explore potential transgenerational effects of these diets on mouth-form plasticity (Figure S2a,b). We found that worms exhibit similar mouth-form ratios relative to their initial response (F1) through generations (Figure S2c,d). When we reverted worms from each generation back to the regular dietary condition without supplements, we did not observe any transgenerational memory of mouth form (Figure S2e,f). Note that over the course of these experiments, we observed more consistent and less variable mouth-form responses in glucose condition (highly St, on average below 20% Eu) relative to oleic acid across different batches.

Oleic acid and glucose supplementations induce fat storage

Next, we performed ORO staining on worm populations reared on oleic acid and glucose diets to visualise and quantify their lipid storage (Figure 1b). Relative to control groups, oleic acid- and glucose-fed worms showed an increase in ORO absorbance which signifies an increase in fat storage (Figure 2b). Whole-body images of stained animals also demonstrated the distribution of storage lipids. Anterior regions of worms clearly indicated lipid accumulation in both oleic acid and glucose conditions (Figure 2c). In particular the head region of the worm allows easy visualization of the influence of supplementation on fat storage (Figure 2c inlets). Thus, both tested supplements induce fat storage in addition to their effect on mouth-form expression.

Differential gene expression analysis points out pathways involved in lipogenesis and lipolysis

To understand which molecular factors are involved in mediating this nutrition-induced mouth-form change, we carried out RNA-seq on worm populations reared in supplemented and control conditions (Figure 3a). First, we performed differential gene expression analysis between each supplement and its control, and then a KEGG enrichment analysis of differentially expressed genes in both oleic acid and glucose conditions. KEGG enrichment analysis revealed pathways involved in lipogenesis (“biosynthesis of unsaturated fatty acids”, “fatty acid elongation”) and lipolysis (“peroxisome”, “fatty acid degradation”) (Figure 3b). This suggested that worms do not only store but also actively utilise lipids in supplemented conditions. Furthermore, we found genes which were upregulated and involved in peroxisomal beta-oxidation and biosynthesis of unsaturated fatty acids. Therefore, we selected genes in these pathways as candidates to investigate whether they mediate the effect of nutrient supplementation on mouth-form expression. We utilised CRISPR gene editing technology to introduce mutations in genetic components involved in both pathways as this method allows easy manipulation in *P. pacificus* (Han et al., 2020; Witte et al., 2015). For these experiments, we only used glucose-supplemented diet to induce fat storage and test responses of mutants

because of the higher reproducibility of results obtained through glucose supplementation and also the impracticality of fatty acid supplementations that require preventive measures from oxidation by light and air (see Materials and methods for more details).

Delta-9 desaturase activity is essential for a complete mouth-form response on glucose-supplemented diet

We targeted delta-9 fatty acid desaturases for studying the role of the biosynthesis of unsaturated fatty acids and *dhs-28* and *daf-22* for investigating the peroxisomal beta-oxidation pathway on nutritional plasticity of the *P. pacificus* mouth-form. Delta-9 desaturases facilitate lipid storage by acting early in the *de novo* fatty acid synthesis pathway, converting saturated fatty acids into monounsaturated fatty acids (Figure 4a). In *C. elegans,* for instance, simultaneous inhibition of the function of two delta-9 desaturases (*fat-6* and *fat-7*) results in the loss of endogenous unsaturated fatty acids and a significant reduction in lipid storage (Brock et al., 2007). Delta-9 desaturases are activated by the transcription factor *sbp-1* (sterol regulatory element binding protein). Like *C. elegans*, *P. pacificus* has a single *sbp-1* gene (PPA37968), which is a one-to-one ortholog of *Cel-sbp-1*. First, we introduced frameshift mutations in this gene to block the activity of delta-9 desaturases and explore its contribution to mouth-form plasticity. However, homozygous mutants were sterile; therefore, we could not further our investigation with this gene. Next, we specifically targeted individual delta-9 desaturases. Interestingly, this gene family has been amplified in *P. pacificus* relative to *C. elegans* (Markov et al., 2015). We constructed a phylogenetic tree based on protein sequences of all *P. pacificus* fatty acid desaturase domain-containing genes and related *C. elegans* desaturases. We found a total of 10 *P. pacificus* genes which contain a delta-9 fatty acid desaturase-like protein domain (Figure S3). Note that unlike for *sbp-1*, there is no one-to-one orthology relationship between delta-9 fatty acid desaturase genes between *P. pacificus* and *C. elegans*, therefore requiring a new nomenclature. Among these genes, ppa_stranded_DN27845_c0_g3_i3, PPA1053, PPA40514, and ppa_stranded_DN27845_c0_g2_i1 showed highest protein sequence similarity with *C. elegans* delta-9 desaturases (*fat-5*, *fat-6*, and *fat-7*). In this order, we named these genes as *Ppa-pddl-1* (*Pristionchus* delta-9 desaturase-like-1), *Ppa-pddl-2*, *Ppa-pddl*-3, and *Ppa-pddl-4* (Figure 4b, Figure S3). Among all, *Ppa-pddl-1*, *Ppa-pddl-3*, and *Ppa-pddl-4* are expressed throughout the development of *P. pacificus* based on previous gene expression analysis (Baskaran et al., 2015). Therefore, we prioritised these three genes and introduced mutations through CRISPR. We obtained three frameshift alleles for both *Ppa-pddl-3* (*tu2033*, *tu2034*, *tu2035*) and *Ppa-pddl-4* (*tu2030*, *tu2031*, *tu2032*), and one frame shift allele (*tu2028*) and a 3bp insertion allele (*tu2029*) for *Ppa-pddl-1* (Table S2). Among these mutants, *tu2028* and

tu2029 were developmentally slow and had significantly reduced fat storage until late adulthood under standard conditions.

Next, we assessed mouth-form responses of all mutants on glucose-supplemented diets. Results revealed that the response of delta-9 desaturase mutants to glucose is generally incomplete, that is, they have a higher Eu ratio than wild type animals (above 20% Eu) (Figure 4c). Compared to all other mutants, *Ppa-pddl-1* mutants were more consistent in their mouth-form response to the glucose-supplemented diet (greater than 40% Eu on average) (Figure 4c). Therefore, we further studied *Ppa-pddl-1* using *tu2028* as reference allele. First, we measured the lipid storage profile of *Ppa-pddl-1*(*tu2028)* on standard diet by ORO staining. We found that *Ppa-pddl-1*(*tu2028)* exhibits significantly lower levels of ORO absorbance relative to wild type (Figure S4a,b). To confirm that the response of *tu2028* to glucose-supplemented diet is due to its mutational background in *Ppa-pddl-1* and not to other mutations as a result of potential off target effects by CRISPR, we genetically reverted *Ppa-pddl-1*(*tu2028)* back to wild type using CRISPR editing via a repair template. Growing wild type, the repaired strain, and the original *Ppa-pddl-1*(*tu2028)* mutant on glucose-supplemented diets revealed highly similar levels of mouth-form ratios between wild type and repaired worms (Figure 4d). Specifically, both strains were highly St on glucose-supplemented diets. Repaired and wild type strains also showed a similar trajectory of ORO absorbance between control and glucose conditions (Figure 4e). Among all, *Ppa-pddl-1*(*tu2028)* had the weakest ORO absorbance in both conditions (Figure 4e). In addition, we observed that *Ppa-pddl-1*(*tu2028)* mutant animals restored their growth on glucose-supplemented diet and also regained their wild type physiological and morphological characteristics upon genetic repair (Figure 4f). Overall, these results highlight the significance of the function of *Ppa-pddl-1* in both lipid storage and mouth-form plasticity.

Peroxisomal beta-oxidation mutants fail to respond to glucose supplementation

Next, we studied mutants of two peroxisomal beta-oxidation genes, *dhs-28* and *daf-22*. *P. pacificus* has two copies of both genes in its genome (Markov et al., 2016) (Figure 5a). For *Ppa-dhs-28.1*, we induced mutation in the copy with the sterol carrier protein domain (PPA20393) and obtained two insertion and three deletion alleles (Table S2). All mutants exhibited the same observable characteristics and were morphologically smaller and developmentally slower relative to wild type animals. For *daf-22*, we used the available double mutant *Ppa-daf-22.1 Ppa-daf-22.2* (Markov et al., 2016). Note that mutants of *Ppa-dhs-28.1* were generated in the PS312 wild type background, whereas the *Ppa-daf-22.1 Ppa-daf-22.2* double mutant was generated in RS2333, which is a highly related derivative of PS312 that was used for several studies on dauer development (Falcke et al., 2018; Markov et al., 2016). Therefore, we included both highly similar wild type strains in our experiments as independent

controls. First, we assessed mouth-form responses of three *Ppa-dhs-28.1* mutant alleles (*tu1855*, *tu1856*, and *tu1858*) and the *Ppa-daf-22.1 Ppa-daf-22.2* double mutant on glucose-supplemented diets. Mutants of both genes had a consistently high Eu mouth-form ratio between control and glucose conditions. Next, we simultaneously assessed mouth-form plasticity and lipid storage of *Ppa-dhs-28.1(tu1855*) and the *Ppa-daf-22.1 Ppa-daf-22.2* double mutant on glucose-supplemented diets. Relative to both wild type strains, mutants remained highly predatory and showed no increase in ORO absorbance between control and glucose conditions (Figure 5b,c). We also observed that glucose-supplemented peroxisomal beta-oxidation mutants exhibit delayed development and reduced body size. Moreover, ORO-stained peroxisomal beta-oxidation mutants show a disruption in lipid storage integrity on glucose-supplemented diets (Figure 5d). They also accumulate larger lipid droplets relative to wild type, indicating reduced lipid oxidation. Taken together, these results indicate that the peroxisomal beta-oxidation pathway is required for the mouth-form response to glucose.

Mouth-form plasticity switch genes are required for phenotypic response in glucose diet

Next, we tested whether the influence of glucose-supplemented diets on mouth-form plasticity is mediated through the central gene regulatory network of mouth-form plasticity (Figure 6a). For that, we utilised mutants of plasticity switch genes which are St-form defective. Specifically, we used *sult-1(tu1061)* and the double mutant *nag-1(tu1142) nag-2(tu1143*) and assessed their mouth form and lipid storage responses on glucose-supplemented diets. Both mutants were unable to switch from Eu to St on glucose-supplemented diets, indicating that the nutritional effect acts upstream of the plasticity switch module (Figure 6b). However, plasticity switch mutants were able to increase their lipid storage when fed on glucose-supplemented diets (Figure 6c). This suggests that the plasticity switch genes are essential for mouth-form response but not for lipid storage.

Non-predatory worms exhibit a fitness advantage over predatory ones

Our final aim was to explore whether there is a fitness advantage of favouring the development of one morph over the other in a fat storage-inducing condition. For this, we chose fecundity as proxy for fitness to evaluate the hermaphrodites' reproductive success by counting viable progeny produced by isolated individuals. First, we reduced the concentration of glucose from 100mM to 80mM to increase the likelihood of obtaining Eu animals. Then, we isolated animals of both morphs grown in the glucose-supplemented diets and measured their fecundity on the standard diet (Figure 7a). Results revealed that St worms on average exhibit higher daily and total fecundity than Eu worms (Figure 7b,c). These findings suggest that promoting the

development of the St morph under such a dietary influence confers a fitness advantage to *P. pacificus*.

Discussion

This study highlights the significance of nutrition in mouth-form plasticity in *P. pacificus*. First, we showed that fat storage-promoting conditions induce the non-predatory morph. This suggests that growing in an "overly satiated" nutritional state increases the likelihood of worms developing as non-predatory. These findings are consistent with previous studies, which indicated that opposite dietary condition, *i.e.*, low nutrition, can lead to the predatory morph in *P. pacificus* (Bento et al., 2010); and even a cannibalistic novel predatory mouth-form in another diplogastrid nematode, *Allodiplogaster sudhausi* (Wighard et al., 2024). In addition, results obtained from transgenerational experiments suggest that worms must constantly be kept at overly satiated state during development to obtain and maintain a mouth-form response, at least in the wild type *P. pacificus* PS312 strain.

Supplementing NGM agar plates with glucose and oleic acids has been an effective method for delivering monosaccharides and fatty acids to worms via dietary ingestion (Alcántar-Fernández et al., 2018; Deline et al., 2013; Nomura et al., 2010). However, it is important to note that this method does not address any potential indirect effect caused by the interaction between the supplement and the bacteria (Kingsley et al., 2021), which can be a research interest on its own. In our experiments, we did not monitor fatty acid and glucose uptake or content in worms, instead we utilised ORO staining, to monitor the outcome of changes in lipid storage. Our results are concordant with previous findings. For instance, glucose supplementation to NGM plates does not only result in the accumulation of glucose but also triacylglycerols in *C. elegans* (Alcántar-Fernández et al., 2018). Oil Red O stains neutral lipids, mainly triacylglycerols. Studies in *C. elegans* have shown an increase in ORO levels in worms grown in oleic acid- and glucose-supplemented conditions (Choi et al., 2021; Lee et al., 2015). Therefore, ORO does not only provide a way of measuring the nutritional status of worms but also validate the uptake of supplementations. Moreover, our results showed that glucose supplementation consistently induces the St morph in higher proportions relative to oleic acid. Note that we could not rule out the cause of inconsistency and the high variability of mouth-form ratios in oleic acid supplementations. Regardless, we selected glucose-supplemented diet as the main fat storage-promoting condition to induce the non-predatory morph.

Further, mutant analyses revealed that genes involved in lipogenesis and lipolysis play a significant role in glucose supplementation-induced mouth-form plasticity. First, we showed that delta-9 desaturase activity, particularly *Ppa-pddl-1*, is required for a complete mouth-form

and lipid storage response. The spatial transcriptome data, published by Rödelsperger et al. (2021), indicates that *Ppa-pddl-1* is enriched in anatomical regions that suggests expression in the intestine. Inactivation of this gene causes a drastic reduction in lipid storage and developmental rate; such phenotypic effect is obtained through simultaneous inhibition of *fat-6* and *fat-7* in *C. elegans* (Brock et al., 2007). However, *Ppa-pddl-1* mutant still exhibits a partial ability to respond to glucose-supplemented diet. This suggests that remaining functional delta-9 desaturases may have compensated for the loss of *Ppa-pddl-1*. Therefore, further functional characterization of delta-9 desaturases is required to completely elucidate potential role of this gene family in mouth-form plasticity. Furthermore, growing peroxisomal beta-oxidation mutants on glucose-supplemented diet revealed that these mutants have disrupted lipid storage integrity, which resulted in a weaker ORO absorbance relative to wild type strains. Additional observations in these mutants, such as developmental delay and reduction in body size, suggest that they are not able to nutritionally benefit from the glucose-supplemented diet. Taken together, these findings suggest that pathways associated with storage and utilisation of lipids carry out essential metabolic processes, mediating mouth-form plasticity in a high nutrition environment.

Additionally, our findings suggest that the dietary effect of glucose supplementation acts upstream of the plasticity switch module. However, how this dietary effect is mediated to influence downstream components, affecting mouth-form decision remains subject of future research. Of note, *de novo* fatty acid synthesis and peroxisomal beta-oxidation pathways produce signalling molecules which have diverse functions (Artyukhin et al., 2018; Watts, 2016). For example, polyunsaturated fatty acids produced by the *de novo* fatty acid synthesis pathway can be integrated into storage lipids or processed further into signalling molecules such as eicosanoids, which can function as ligands for transcription factors, affecting gene expression. Cytochrome P450 enzymes are involved in the production of eicosanoids from polyunsaturated fatty acids (Kulas et al., 2008) and their activity have been associated with several biological functions such as development, dauer formation, and lipid metabolism in *C. elegans* (for review, see Larigot et al., 2022). Cytochrome P450-related pathways are enriched (KEGG) in our differential gene expression data for both oleic acid and glucose, suggesting a potential role. Moreover, peroxisomal beta-oxidation pathway produces ascarosides, extracellular signalling molecules which can influence dauer formation and mouth-form plasticity (Butcher, 2017; Butcher et al., 2007; Markov et al., 2016; Werner et al., 2018). Taken together, lipid-mediated signalling may have potential functions in nutrition-induced mouth-form plasticity and therefore further research is required to elucidate associated mechanisms.

Finally, we sought to explore whether there is an adaptive value of facilitating the development of the St morph in a high nutrition environment. In *P. pacificus*, while the Eu morph exhibits a fitness advantage associated with its predatory ability (Serobyan et al., 2014),

the St morph claims this advantage through a faster development and a higher fecundity (Dardiry et al., 2023; Serobyan et al., 2013). In addition, highly non-predatory natural isolates of *P. pacificus* have higher fecundity and exhibit faster development on standard dietary condition relative to highly predatory strains, which were isolated from similar localities in nature (Dardiry et al., 2023). Growing worms on glucose-supplemented diet allowed separation of morphs; then assessing their fecundity revealed a fitness disadvantage for the predatory mouth-form. Intriguingly, our findings suggest that glucose supplementation does not offset the fitness cost of producing the Eu phenotype or promote its development to begin with. Hence, the fitness advantage associated with the non-predatory morph suggests that favouring its development, under such dietary effect, can be beneficial for the population.

In summary, this study signifies the nutritional sensitivity of mouth-form polyphenism in *P. pacificus* and adds nutritional status as important environmental factor influencing mouth form. We introduced glucose-supplementation as a novel environmental condition to induce non-predatory morph. Our findings suggest that nutritional status of the worm can potentially dictate its mouth-form fate. We also found a strong association between lipid metabolism and mouth-form plasticity. Lastly, glucose-supplementation helped understand fitness consequences of the mouth-form determination, indicating an advantage for the non-predatory morph.

Acknowledgements

We would like to thank Dr Christian Rödelsperger for his guidance and support during this project. We also thank Heike Haussmann for freezing all the worm strains generated by this work. Additionally, we thank all the current and past members of the SommerLab for valuable discussions.

Competing interests

No competing interests declared.

Funding

This study was funded by the Max Planck Society. V.P. was supported by the International Max Planck Research School (IMPRS) “From Molecules to Organisms”.

Data availability

Raw Reads obtained from the RNA-seq were deposited in the European Nucleotide Archive (ENA) under the accession number PRJEB76787.

References

Akduman, N., Lightfoot, J. W., Röseler, W., Witte, H., Lo, W. S., Rödelsperger, C., & Sommer, R. J. (2020). Bacterial vitamin B12 production enhances nematode predatory behavior. *ISME Journal*, **14**, 1494–1507. doi:10.1038/s41396-020-0626-2

Alcántar-Fernández, J., Navarro, R. E., Salazar-Martínez, A. M., Pérez-Andrade, M. E., & Miranda-Ríos, J. (2018). *Caenorhabditis elegans* respond to high-glucose diets through a network of stress-responsive transcription factors. *PLoS ONE*, **13**. doi:10.1371/journal.pone.0199888

Artyukhin, A. B., Zhang, Y. K., Akagi, A. E., Panda, O., Sternberg, P. W., & Schroeder, F. C. (2018). Metabolomic “dark Matter” dependent on peroxisomal β-Oxidation in *Caenorhabditis elegans*. *Journal of the American Chemical Society*, **140**, 2841–2852. doi:10.1021/jacs.7b11811

Athanasouli, M., Akduman, N., Röseler, W., Theam, P., & Rödelsperger, C. (2023). Thousands of *Pristionchus pacificus* orphan genes were integrated into developmental networks that respond to diverse environmental microbiota. *PLOS Genetics*, **19**, e1010832. doi:10.1371/journal.pgen.1010832

Athanasouli, M., Witte, H., Weiler, C., Loschko, T., Eberhardt, G., Sommer, R. J., & Rödelsperger, C. (2020). Comparative genomics and community curation further improve gene annotations in the nematode *Pristionchus pacificus*. *BMC Genomics*, **21**, 708. doi:10.1186/s12864-020-07100-0

Baskaran, P., Rödelsperger, C., Prabh, N., Serobyan, V., Markov, G. V., Hirsekorn, A., & Dieterich, C. (2015). Ancient gene duplications have shaped developmental stage-specific expression in *Pristionchus pacificus*. *BMC Evolutionary Biology*, **15**. doi:10.1186/s12862-015-0466-2

Beltran, T., Shahrezaei, V., Katju, V., & Sarkies, P. (2020). Epimutations driven by small RNAs arise frequently but most have limited duration in *Caenorhabditis elegans*. *Nature Ecology & Evolution*, **4**, 1539–1548. doi:10.1038/s41559-020-01293-z

Bento, G., Ogawa, A., & Sommer, R. J. (2010). Co-option of the hormone-signalling module dafachronic acid-DAF-12 in nematode evolution. *Nature*, **466**, 494–497. doi:10.1038/nature09164

Bose, N., Ogawa, A., von Reuss, S. H., Yim, J. J., Ragsdale, E. J., Sommer, R. J., & Schroeder, F. C. (2012). Complex small-molecule architectures regulate phenotypic plasticity in a nematode. *Angewandte Chemie International Edition*, **51**, 12438–12443. doi:10.1002/anie.201206797

641 Brock, T. J., Browse, J., & Watts, J. L. (2007). Fatty acid desaturation and the regulation of
642 adiposity in *Caenorhabditis elegans*. *Genetics*, **176**, 865–875.
643 doi:10.1534/genetics.107.071860
644 Brown, A. L., Meiborg, A. B., Franz-Wachtel, M., Macek, B., Gordon, S., Rog, O., Weadick,
645 C. J., & Werner, M. S. (2024). Characterization of the *Pristionchus pacificus* "epigenetic
646 toolkit" reveals the evolutionary loss of the histone methyltransferase complex PRC2.
647 *Genetics*, **227**. doi:10.1093/genetics/iyae041
648 Bui, L. T., Ivers, N. A., & Ragsdale, E. J. (2018). A sulfotransferase dosage-dependently
649 regulates mouthpart polyphenism in the nematode *Pristionchus pacificus*. *Nature*
650 *Communications*, **9**, 4119. doi:10.1038/s41467-018-05612-8
651 Butcher, R. A. (2017). Small-molecule pheromones and hormones controlling nematode
652 development. *Nature Chemical Biology*, **13**, 577–586. doi:10.1038/nchembio.2356
653 Butcher, R. A., Fujita, M., Schroeder, F. C., & Clardy, J. (2007). Small-molecule pheromones
654 that control dauer development in *Caenorhabditis elegans*. *Nature Chemical Biology*, **3**,
655 420–422. doi:10.1038/nchembio.2007.3
656 Casasa, S., Katsougia, E., & Ragsdale, E. J. (2023). A Mediator subunit imparts robustness
657 to a polyphenism decision. *Proceedings of the National Academy of Sciences of the*
658 *United States of America*, **120**. doi:10.1073/pnas.2308816120
659 Casasa, S., Zattara, E. E., & Moczek, A. P. (2020). Nutrition-responsive gene expression and
660 the developmental evolution of insect polyphenism. *Nature Ecology and Evolution*, **4**,
661 970–978. doi:10.1038/s41559-020-1202-x
662 Chandra, V., Fetter-Pruneda, I., Oxley, P. R., Ritger, A. L., McKenzie, S. K., Libbrecht, R., &
663 Kronauer, D. J. C. (2018). Social regulation of insulin signaling and the evolution of
664 eusociality in ants. *Science*, **361**, 398–402. doi:10.1126/science.aar5723
665 Choi, L. S., Shi, C., Ashraf, J., Sohrabi, S., & Murphy, C. T. (2021). Oleic Acid protects
666 *Caenorhabditis* mothers from mating-induced death and the cost of reproduction.
667 *Frontiers in Cell and Developmental Biology*, **9**. doi:10.3389/fcell.2021.690373
668 Dardiry, M., Piskobulu, V., Kalirad, A., & Sommer, R. J. (2023). Experimental and theoretical
669 support for costs of plasticity and phenotype in a nematode cannibalistic trait. *Evolution*
670 *Letters*, **7**, 48–57. doi:10.1093/evlett/qrac001
671 Deline, M. L., Vrablik, T. L., & Watts, J. L. (2013). Dietary supplementation of
672 polyunsaturated fatty acids in *Caenorhabditis elegans*. *Journal of Visualized*
673 *Experiments*, **81**. doi:10.3791/50879
674 Dieterich, C., Clifton, S. W., Schuster, L. N., Chinwalla, A., Delehaunty, K., Dinkelacker, I.,
675 Fulton, L., Fulton, R., Godfrey, J., Minx, P., Mitreva, M., Roeseler, W., Tian, H., Witte,
676 H., Yang, S.-P., Wilson, R. K., & Sommer, R. J. (2008). The *Pristionchus pacificus*

677 genome provides a unique perspective on nematode lifestyle and parasitism. *Nature*
678 *Genetics*, **40**, 1193–1198. doi:10.1038/ng.227
679 Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson,
680 M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*,
681 **29**, 15–21. doi:10.1093/bioinformatics/bts635
682 Emlen, D. J. (1997). Diet alters male horn allometry in the beetle *Onthophagus acuminatus*
683 (Coleoptera: Scarabaeidae). *Proceedings of the Royal Society of London. Series B:*
684 *Biological Sciences*, **264**, 567–574. doi:10.1098/rspb.1997.0081
685 Falcke, J. M., Bose, N., Artyukhin, A. B., Rödelsperger, C., Markov, G. V., Yim, J. J., Grimm,
686 D., Claassen, M. H., Panda, O., Baccile, J. A., Zhang, Y. K., Le, H. H., Jolic, D.,
687 Schroeder, F. C., & Sommer, R. J. (2018). Linking genomic and metabolomic natural
688 variation uncovers nematode pheromone biosynthesis. *Cell Chemical Biology*, **25**, 787-
689 796.e12. doi:10.1016/j.chembiol.2018.04.004
690 Feldhaus, C., & Piskobulu, V. (2024). A low bias method for quantification of Oil Red O
691 staining. *BioRxiv*, 2024.06.18.599611. doi:10.1101/2024.06.18.599611
692 Ghalambor, C. K., McKay, J. K., Carroll, S. P., & Reznick, D. N. (2007). Adaptive versus non-
693 adaptive phenotypic plasticity and the potential for contemporary adaptation in new
694 environments. *Functional Ecology*, **21**, 394–407. doi:10.1111/j.1365-2435.2007.01283.x
695 Gotoh, H., Miyakawa, H., Ishikawa, A., Ishikawa, Y., Sugime, Y., Emlen, D. J., Lavine, L. C.,
696 & Miura, T. (2014). Developmental link between sex and nutrition; *doublesex* regulates
697 sex-specific mandible growth via juvenile hormone signaling in stag beetles. *PLoS*
698 *Genetics*, **10**. doi:10.1371/journal.pgen.1004098
699 Han, S., Schroeder, E. A., Silva-García, C. G., Hebestreit, K., Mair, W. B., & Brunet, A.
700 (2017). Mono-unsaturated fatty acids link H3K4me3 modifiers to *C. elegans* lifespan.
701 *Nature*, **544**, 185–190. doi:10.1038/nature21686
702 Han, Z., Lo, W.-S., Lightfoot, J. W., Witte, H., Sun, S., & Sommer, R. J. (2020). Improving
703 transgenesis efficiency and CRISPR-associated tools through codon optimization and
704 native intron addition in *Pristionchus* nematodes. *Genetics*, **216**, 947–956.
705 doi:10.1534/genetics.120.303785
706 Haydak, M. H. (1970). Honey bee nutrition. *Annual Review of Entomology*, **15**, 143–156.
707 doi:10.1146/annurev.en.15.010170.001043
708 Herrmann, M., Mayer, W. E., & Sommer, R. J. (2006). Nematodes of the genus *Pristionchus*
709 are closely associated with scarab beetles and the Colorado potato beetle in Western
710 Europe. *Zoology*, **109**, 96–108. doi:10.1016/j.zool.2006.03.001
711 Kamakura, M. (2011). Royalactin induces queen differentiation in honeybees. *Nature*, **473**,
712 478–483. doi:10.1038/nature10093

713 Kingsley, S. F., Seo, Y., Allen, C., Ghanta, K. S., Finkel, S., & Tissenbaum, H. A. (2021).
714 Bacterial processing of glucose modulates *C. elegans* lifespan and healthspan.
715 *Scientific Reports*, **11**, 5931. doi:10.1038/s41598-021-85046-3
716 Kucharski, R., Maleszka, J., Foret, S., & Maleszka, R. (2008). Nutritional control of
717 reproductive status in honeybees via DNA methylation. Science, **319**, 1827–1830.
718 doi:10.1126/science.1153069
719 Kulas, J., Schmidt, C., Rothe, M., Schunck, W.-H., & Menzel, R. (2008). Cytochrome P450-
720 dependent metabolism of eicosapentaenoic acid in the nematode *Caenorhabditis*
721 *elegans*. *Archives of Biochemistry and Biophysics*, **472**, 65–75.
722 doi:10.1016/j.abb.2008.02.002
723 Larigot, L., Mansuy, D., Borowski, I., Coumoul, X., & Dairou, J. (2022). Cytochromes P450 of
724 *Caenorhabditis elegans*: implication in biological functions and metabolism of
725 xenobiotics. *Biomolecules*, **12**, 342. doi:10.3390/biom12030342
726 Lee, D., Jeong, D.-E., Son, H. G., Yamaoka, Y., Kim, H., Seo, K., Khan, A. A., Roh, T.-Y.,
727 Moon, D. W., Lee, Y., & Lee, S.-J. V. (2015). SREBP and MDT-15 protect *C. elegans*
728 from glucose-induced accelerated aging by preventing accumulation of saturated fat.
729 *Genes & Development*, **29**, 2490–2503. doi:10.1101/gad.266304.115
730 Lenuzzi, M., Witte, H., Riebesell, M., Rödelsperger, C., Hong, R. L., & Sommer, R. J. (2023).
731 Influence of environmental temperature on mouth-form plasticity in *Pristionchus*
732 *pacificus* acts through daf-11-dependent cGMP signaling. *Journal of Experimental*
733 *Zoology Part B: Molecular and Developmental Evolution*, **340**, 214–224.
734 doi:10.1002/jez.b.23094
735 Levis, N. A., & Ragsdale, E. J. (2023). A histone demethylase links the loss of plasticity to
736 nongenetic inheritance and morphological change. *Nature Communications*, **14**.
737 doi:10.1038/s41467-023-44306-8
738 Li, S., Xu, S., Ma, Y., Wu, S., Feng, Y., Cui, Q., Chen, L., Zhou, S., Kong, Y., Zhang, X., Yu,
739 J., Wu, M., & Zhang, S. O. (2016). A genetic screen for mutants with supersized lipid
740 droplets in *Caenorhabditis elegans*. *G3: Genes, Genomes, Genetics*, **6**, 2407–2419.
741 doi:10.1534/g3.116.030866
742 Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program
743 for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
744 doi:10.1093/bioinformatics/btt656
745 Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and
746 dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**, 550.
747 doi:10.1186/s13059-014-0550-8

748 Markov, G. V., Baskaran, P., & Sommer, R. J. (2015). The same or not the same: lineage-
749 specific gene expansions and homology relationships in multigene families in
750 nematodes. *Journal of Molecular Evolution*, **80**, 18–36. doi:10.1007/s00239-014-9651-y
751 Markov, G. V., Meyer, J. M., Panda, O., Artyukhin, A. B., Claaßen, M., Witte, H., Schroeder,
752 F. C., & Sommer, R. J. (2016). Functional conservation and divergence of *daf-22*
753 paralogs in *Pristionchus pacificus* dauer development. *Molecular Biology and Evolution*,
754 **33**, 2506–2514. doi:10.1093/molbev/msw090
755 Meyer, J. M., Baskaran, P., Quast, C., Susoy, V., Rödelsperger, C., Glöckner, F. O., &
756 Sommer, R. J. (2017). Succession and dynamics of *Pristionchus* nematodes and their
757 microbiome during decomposition of *Oryctes borbonicus* on La Réunion Island.
758 *Environmental Microbiology*, **19**, 1476–1489. doi:10.1111/1462-2920.13697
759 Moczek, A. P. (1998). Horn polyphenism in the beetle *Onthophagus taurus*: larval diet quality
760 and plasticity in parental investment determine adult body size and male horn
761 morphology. *Behavioral Ecology*, **9**, 636–641. doi:10.1093/beheco/9.6.636
762 Namdeo, S., Moreno, E., Rödelsperger, C., Baskaran, P., Witte, H., & Sommer, R. J. (2018).
763 Two independent sulfation processes regulate mouth-form plasticity in the nematode
764 *Pristionchus pacificus*. *Development*, **145**. doi:10.1242/dev.166272
765 Nomura, T., Horikawa, M., Shimamura, S., Hashimoto, T., & Sakamoto, K. (2010). Fat
766 accumulation in *Caenorhabditis elegans* is mediated by SREBP homolog SBP-1.
767 *Genes and Nutrition*, **5**, 17–27. doi:10.1007/s12263-009-0157-y
768 O'Rourke, E. J., Soukas, A. A., Carr, C. E., & Ruvkun, G. (2009). *C. elegans* major fats are
769 stored in vesicles distinct from lysosome-related organelles. *Cell Metabolism*, **10**, 430–
770 435. doi:10.1016/j.cmet.2009.10.002
771 Pigliucci, M. (2001). *Phenotypic plasticity: beyond nature and nurture*. Johns Hopkins
772 University Press.
773 Ragsdale, E. J., Kanzaki, N., & Herrmann, M. (2015). Taxonomy and natural history: the
774 genus *Pristionchus*. In *Pristionchus pacificus* (pp. 77–120). BRILL.
775 doi:10.1163/9789004260306_005
776 Ragsdale, E. J., Müller, M. R., Rödelsperger, C., & Sommer, R. J. (2013). A developmental
777 switch coupled to the evolution of plasticity acts through a sulfatase. *Cell*, **155**, 922.
778 doi:10.1016/j.cell.2013.09.054
779 Renahan, T., Lo, W. S., Werner, M. S., Rochat, J., Herrmann, M., & Sommer, R. J. (2021).
780 Nematode biphasic 'boom and bust' dynamics are dependent on host bacterial load
781 while linking dauer and mouth-form polyphenisms. *Environmental Microbiology*, **23**,
782 5102–5113. doi:10.1111/1462-2920.15438

783 Renahan, T., & Sommer, R. J. (2022). Multidimensional competition of nematodes affects
784 plastic traits in a beetle ecosystem. *Frontiers in Cell and Developmental Biology*, **10**.
785 doi:10.3389/fcell.2022.985831
786 Rödelsperger, C., Ebbing, A., Sharma, D. R., Okumura, M., Sommer, R. J., & Korswagen, H.
787 C. (2021). Spatial transcriptomics of nematodes identifies sperm cells as a source of
788 genomic novelty and rapid evolution. *Molecular Biology and Evolution*, **38**, 229–243.
789 doi:10.1093/molbev/msaa207
790 Rödelsperger, C., Meyer, J. M., Prabh, N., Lanz, C., Bemm, F., & Sommer, R. J. (2017).
791 Single-molecule sequencing reveals the chromosome-scale genomic architecture of the
792 nematode model organism *Pristionchus pacificus*. *Cell Reports*, **21**, 834–844.
793 doi:10.1016/j.celrep.2017.09.077
794 Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch,
795 S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.-Y., White, D. J., Hartenstein, V.,
796 Eliceiri, K., Tomancak, P., & Cardona, A. (2012). Fiji: an open-source platform for
797 biological-image analysis. *Nature Methods*, **9**, 676–682. doi:10.1038/nmeth.2019
798 Schroeder, N. E. (2021). Introduction to *Pristionchus pacificus* anatomy. *Journal of*
799 *Nematology*, **53**. doi:10.21307/jofnem-2021-091
800 Serobyan, V., Ragsdale, E. J., Müller, M. R., & Sommer, R. J. (2013). Feeding plasticity in
801 the nematode *Pristionchus pacificus* is influenced by sex and social context and is
802 linked to developmental speed. *Evolution & Development*, **15**, 161–170.
803 doi:10.1111/ede.12030
804 Serobyan, V., Ragsdale, E. J., & Sommer, R. J. (2014). Adaptive value of a predatory mouth-
805 form in a dimorphic nematode. *Proceedings of the Royal Society B: Biological*
806 *Sciences*, **281**, 20141334. doi:10.1098/rspb.2014.1334
807 Serobyan, V., Xiao, H., Namdeo, S., Rödelsperger, C., Sieriebriennikov, B., Witte, H.,
808 Röseler, W., & Sommer, R. J. (2016). Chromatin remodelling and antisense-mediated
809 up-regulation of the developmental switch gene eud-1 control predatory feeding
810 plasticity. *Nature Communications*, **7**. doi:10.1038/ncomms12337
811 Sieriebriennikov, B., Prabh, N., Dardiry, M., Witte, H., Röseler, W., Kieninger, M. R.,
812 Rödelsperger, C., & Sommer, R. J. (2018). A developmental switch generating
813 phenotypic plasticity is part of a conserved multi-gene locus. *Cell Reports*, **23**, 2835-
814 2843.e4. doi:10.1016/j.celrep.2018.05.008
815 Sieriebriennikov, B., Sun, S., Lightfoot, J. W., Witte, H., Moreno, E., Rödelsperger, C., &
816 Sommer, R. J. (2020). Conserved nuclear hormone receptors controlling a novel plastic
817 trait target fast-evolving genes expressed in a single cell. *PLoS Genetics*, **16**(4).
818 doi:10.1371/journal.pgen.1008687

819 Simpson, S. J., Sword, G. A., & Lo, N. (2011). Polyphenism in insects. *Current Biology*, **21**,
820 R738–R749. doi:10.1016/j.cub.2011.06.006
821 Sommer, R. J. (2020). Phenotypic plasticity: from theory and genetics to current and future
822 challenges. *Genetics*, **215**, 1–13. doi:10.1534/genetics.120.303163
823 Sommer, R. J., Carta, L. K., Kim, S.-Y., & Sternberg, P. W. (1996). Morphological, genetic
824 and molecular description of *Pristionchus pacificus* sp. n. (Nematoda :
825 Neodiplogastridae). *Fundamental and Applied Nematology*, **19**, 511-521.
826 Tamura, K., Stecher, G., & Kumar, S. (2021). MEGA11: molecular evolutionary genetics
827 analysis version 11. *Molecular Biology and Evolution*, **38**, 3022–3027.
828 doi:10.1093/molbev/msab120
829 Theska, T., Sieriebriennikov, B., Wighard, S. S., Werner, M. S., & Sommer, R. J. (2020).
830 Geometric morphometrics of microscopic animals as exemplified by model nematodes.
831 *Nature Protocols*, **15**, 2611–2644. doi:10.1038/s41596-020-0347-z
832 Theska, T., & Sommer, R. J. (2024). Feeding-structure morphogenesis in “rhabditid” and
833 diplogastrid nematodes is not controlled by a conserved genetic module. *Evolution &*
834 *Development*, **26**. doi:10.1111/ede.12471
835 Wan, Q. L., Meng, X., Wang, C., Dai, W., Luo, Z., Yin, Z., Ju, Z., Fu, X., Yang, J., Ye, Q.,
836 Zhang, Z. H., & Zhou, Q. (2022). Histone H3K4me3 modification is a transgenerational
837 epigenetic signal for lipid metabolism in *Caenorhabditis elegans*. *Nature*
838 *Communications*, **13**. doi:10.1038/s41467-022-28469-4
839 Watts, J. L. (2016). Using *Caenorhabditis elegans* to uncover conserved functions of omega-
840 3 and omega-6 fatty acids. *Journal of Clinical Medicine*, **5**. doi:10.3390/jcm5020019
841 Werner, M. S., Claaßen, M. H., Renahan, T., Dardiry, M., & Sommer, R. J. (2018). Adult
842 influence on juvenile phenotypes by stage-specific pheromone production. *iScience*, **10**,
843 123–134. doi:10.1016/j.isci.2018.11.027
844 Werner, M. S., Loschko, T., King, T., Reich, S., Theska, T., Franz-Wachtel, M., Macek, B., &
845 Sommer, R. J. (2023). Histone 4 lysine 5/12 acetylation enables developmental
846 plasticity of *Pristionchus* mouth form. *Nature Communications*, **14**. doi:10.1038/s41467-
847 023-37734-z
848 Werner, M. S., Sieriebriennikov, B., Loschko, T., Namdeo, S., Lenuzzi, M., Dardiry, M.,
849 Renahan, T., Sharma, D. R., & Sommer, R. J. (2017). Environmental influence on
850 *Pristionchus pacificus* mouth form through different culture methods. *Scientific Reports*,
851 **7**. doi:10.1038/s41598-017-07455-7
852 West-Eberhard, M. J. (2003). *Developmental plasticity and evolution*. Oxford University
853 Press.

854 Wighard, S., Witte, H., & Sommer, R. J. (2024). Conserved switch genes that arose via
855 whole-genome duplication regulate a cannibalistic nematode morph. *Science*
856 *Advances*, **10**. doi:10.1126/sciadv.adk6062
857 Witte, H., Moreno, E., Rödelsperger, C., Kim, J., Kim, J.-S., Streit, A., & Sommer, R. J.
858 (2015). Gene inactivation using the CRISPR/Cas9 system in the nematode *Pristionchus*
859 *pacificus*. *Development Genes and Evolution*, **225**, 55–62. doi:10.1007/s00427-014-
860 0486-8
861 Yang, C. H., & Pospisilik, J. A. (2019). Polyphenism – A window into gene-environment
862 interactions and phenotypic plasticity. *Frontiers in Genetics*, **10**.
863 doi:10.3389/fgene.2019.00132

Figure Legends

Figure 1. Mouth-form polyphenism of *P. pacificus*, and the experimental design for supplementation studies. (a) Differential interference contrast (DIC) images of the eurystomatous (Eu) and the stenostomatous (St) mouth morphs. The Eu morph has a wide buccal cavity, a hook-like dorsal tooth (on the left, outlined), and a sub-ventral tooth. The narrow-mouthed stenostomatous (St) morph has a flint-shaped dorsal tooth (on the left, outlined). Scale bar is 5μm. (b) Illustration of the general experimental design using supplementation with monosaccharides and fatty acids.

Figure 2. Effect of oleic acid and glucose supplementations on mouth-form plasticity and lipid storage in *P. pacificus*. (a) Concentration-dependent effect of oleic acid and glucose on mouth-form plasticity. N ≥ 3 biological replicates per condition for each concentration. Each faint data point represents a replicate (plate), with 30 animals per replicate being scored for mouth-form percentage (Eu%). Error bars represent s.e.m. (b) Oil Red O absorbance (RawIntDen/body area) obtained from worms grown in oleic acid, glucose, and respective control conditions. N = 15 per condition. Each faint data point represents an individual worm. P values are obtained from Welch two sample t-test and two sample t-test for oleic acid and glucose comparisons, respectively. Bars represent mean values of all samples in each condition. Error bars represent s.d. (c) Representative images of ORO-quantified worms, indicating lipid storage profile. Images are acquired from the blue channel in grayscale. Lipid droplets appear dark. Scale bar is 50μm.

Figure 3. Transcriptome analysis. (a) An illustration of RNA extraction. (b) Overrepresented KEGG pathways of differentially expressed genes for oleic acid and glucose conditions. Statistical analysis by Fisher's exact-test with multiple testing correction (Bonferroni corrected $P<0.05$).

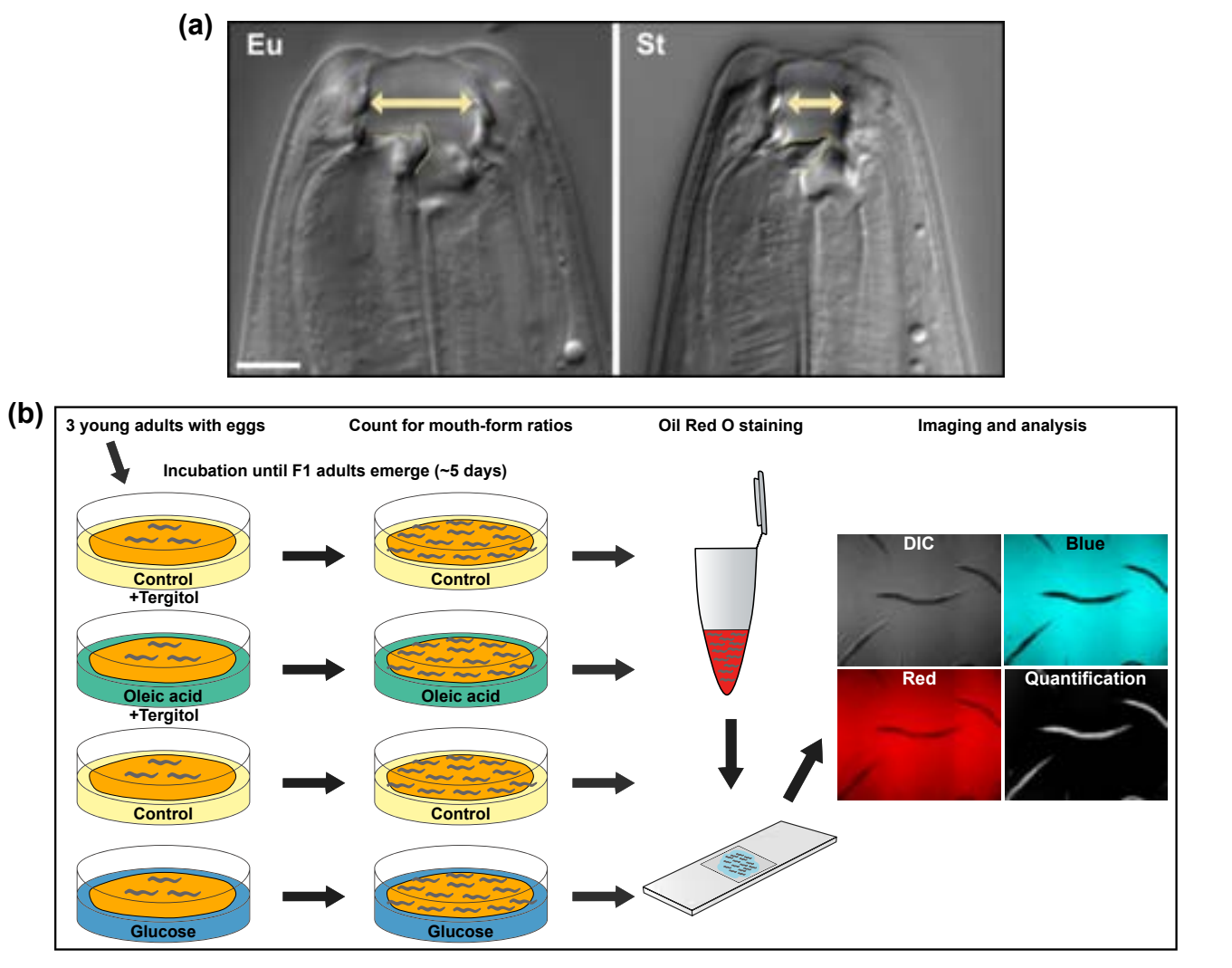
Figure 4. Significance of delta-9 desaturase activity in glucose supplementation-induced mouth-form plasticity. (a) Schematic representation of the essential activity of delta-9 desaturases, facilitating fat storage by converting saturated fatty acids into monounsaturated fatty acids. Sterol regulatory element binding protein 1 (SBP-1) activates delta-9 desaturases. (b) A simplified phylogenetic tree of *Pristionchus* delta-9 desaturase-like genes (*Ppa-pddl-1*, *Ppa-pddl-2*, *Ppa-pddl-3*, *Ppa-pddl-4*), and *C. elegans* delta-9 desaturases (*fat-5*, *fat-6*, and *fat-7*). Triangle denotes *P. pacificus* genes which are constitutively expressed during development (Baskaran et al., 2015). The phylogenetic tree is drawn based on the tree in Figure S3. (c) Eu percentages of CRISPR mutants of *Ppa-pddl-1*, *Ppa-pddl-3*, *Ppa-pddl-4*,

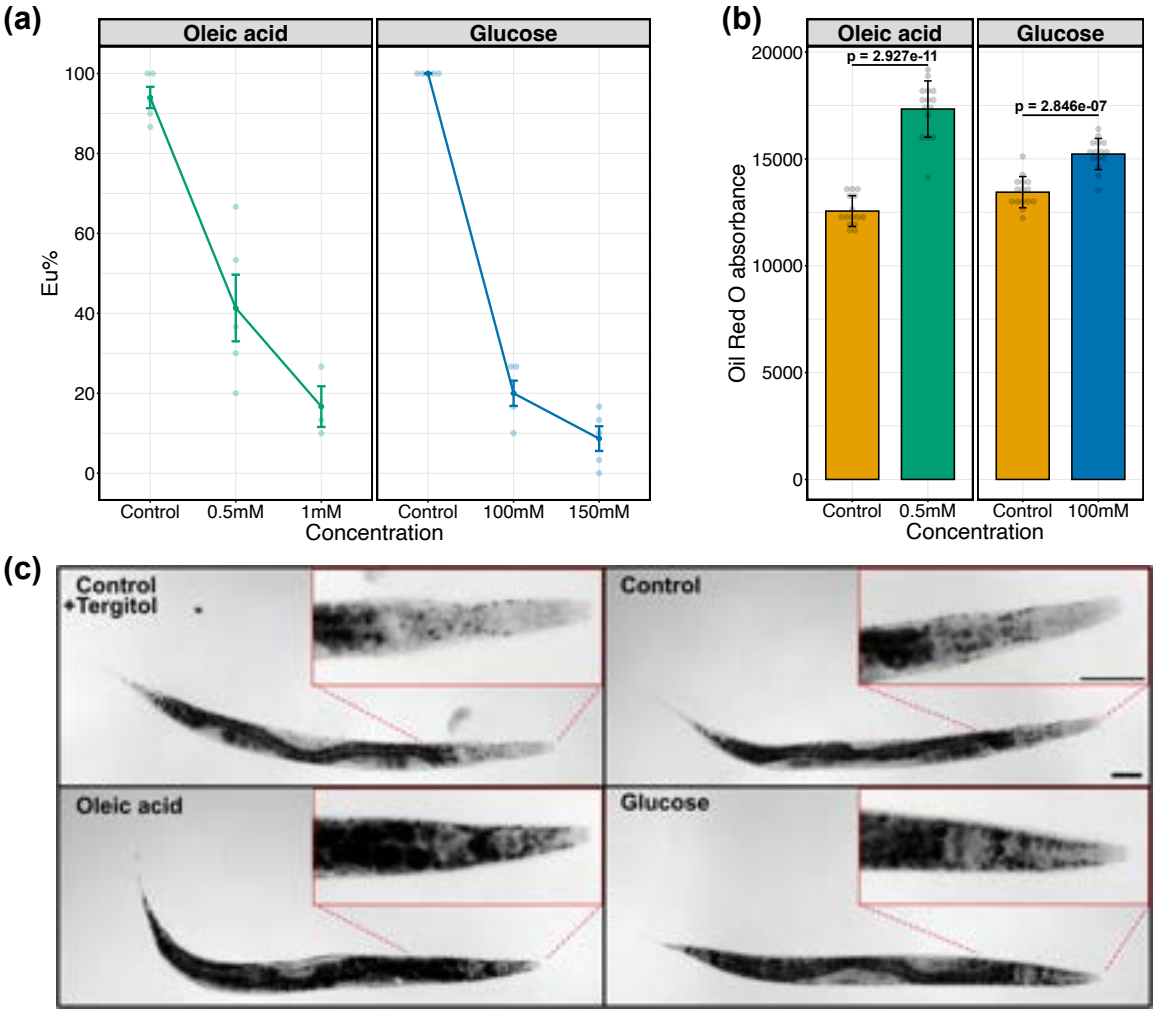
and wild type animals in control and glucose conditions. N ≥ 2 biological replicates per strain for each condition. Bars represent mean values of all replicates. Error bars represent s.d. (d) Eu percentages of wild type, *tu2028*(*Ppa-pddl-1*), and CRISPR repaired-*tu2028* strains in control and glucose conditions. N ≥ 3 biological replicates per strain for each condition. (c,d) Each faint data point represents a replicate (plate), with 30 animals per plate being scored for mouth-form percentage (Eu%). (e) ORO absorbance (RawIntDen/body area) obtained from wild type, *tu2028*(*Ppa-pddl-1*), and CRISPR repaired-*tu2028* strains grown on control and glucose conditions. N = 15 per strain for each condition. Each faint datapoint represents an individual worm. (d,e) Error bars represent s.e.m. (f) Representative images of ORO-quantified strains from control (-glucose) and glucose (+glucose) conditions. Images are acquired from the blue channel in grayscale. Lipid droplets appear dark. Scale bar is 100μm.

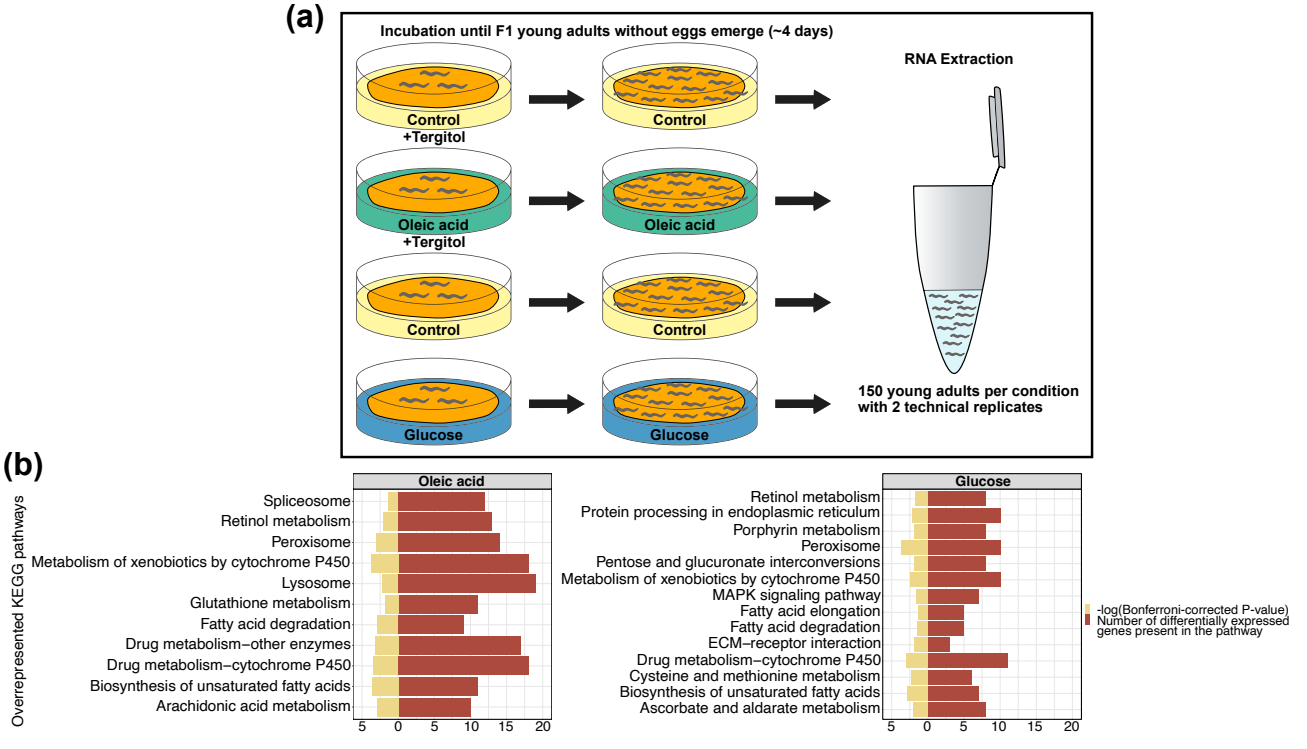
Figure 5. Response of peroxisomal beta-oxidation mutants to glucose-supplemented diet. (a) Schematic of peroxisomal beta-oxidation pathway, indicating duplicate copies in *P. pacificus* for *dhs-28* and *daf-22* genes relative to *C. elegans*. (b) Eu percentages of peroxisomal beta-oxidation mutants and wild type strains in control and glucose conditions. N ≥ 2 biological replicate per strain for each condition. Each faint data point represents a replicate (plate), with 30 animals per plate being scored for mouth-form percentage (Eu%). (c) ORO absorbance (RawIntDen/body area) obtained from peroxisomal beta-oxidation mutants and wild type strains grown on control and glucose conditions. N = 15 per strain for each condition. Each faint data point represents an individual worm. (b,c) Error bars represent s.e.m. (d) Representative images of ORO-quantified strains from control (-glucose) and glucose (+glucose) conditions. Images are acquired from the blue channel in grayscale. Lipid droplets appear dark. Scale bar is 100μm.

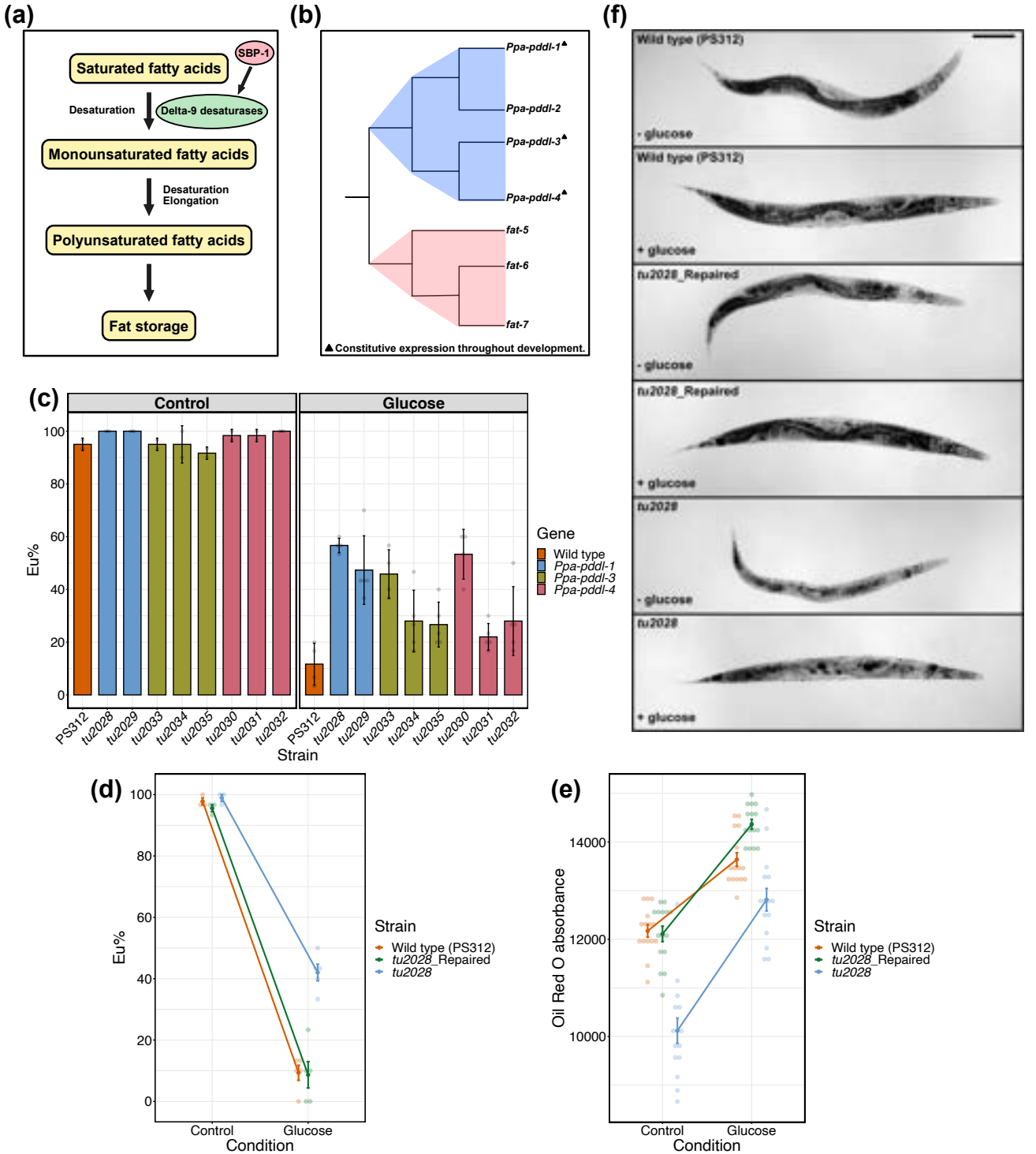
Figure 6. Role of plasticity switch genes in glucose supplementation-induced mouth-form plasticity. (a) Schematic of the mouth-form gene regulatory network, indicating different modules with genetic and epigenetic components. (b) Eu percentages of plasticity switch mutants and wild type strains in glucose and control conditions. N = 3 biological replicates per strain for each condition. Each faint data point represents a replicate (plate), with 30 animals per plate being scored for mouth-form percentage (Eu%). (c) ORO absorbance (RawIntDen/body area) measured in plasticity switch mutants and wild type strains grown on control and glucose conditions. N = 15 per strain for each condition. Each faint data point represents a worm. (b,c) Error bars represent s.e.m.

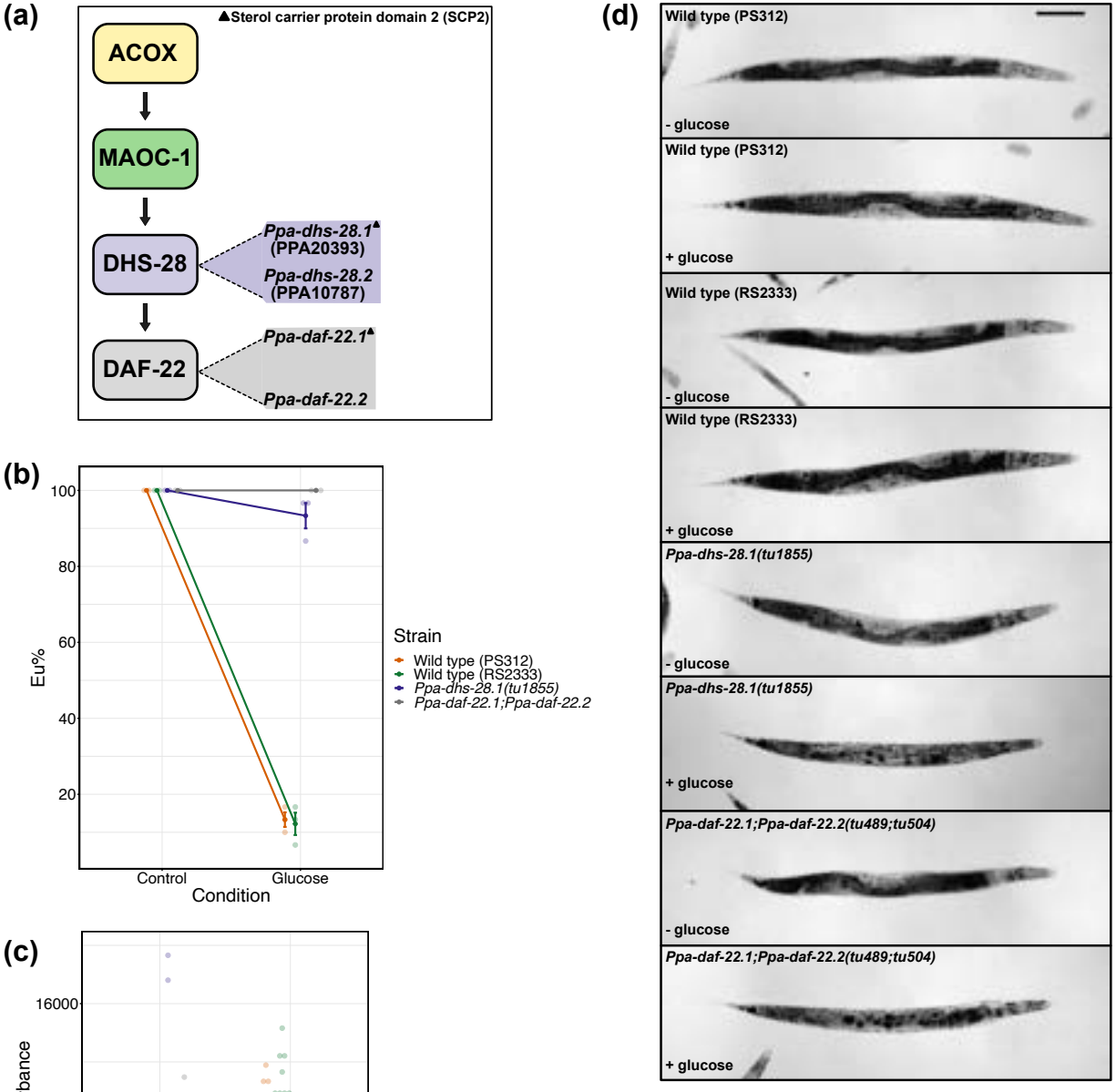
935 **Figure 7. Fecundity of different morphs after glucose supplementation.** (a) Illustration for
936 the experimental design to study fecundity of Eu and St animals obtained from glucose-
937 supplemented condition. (b) Daily fecundity of Eu and St animals. Error bars represent s.e.m.
938 (c) Overall fecundity of Eu and St animals. Bars represent mean values of all individuals for
939 each morph. Error bars represent s.d. P value is obtained from Wilcoxon rank sum test. (b,c)
940 N = 34 per morph. Each faint data point represents a worm.

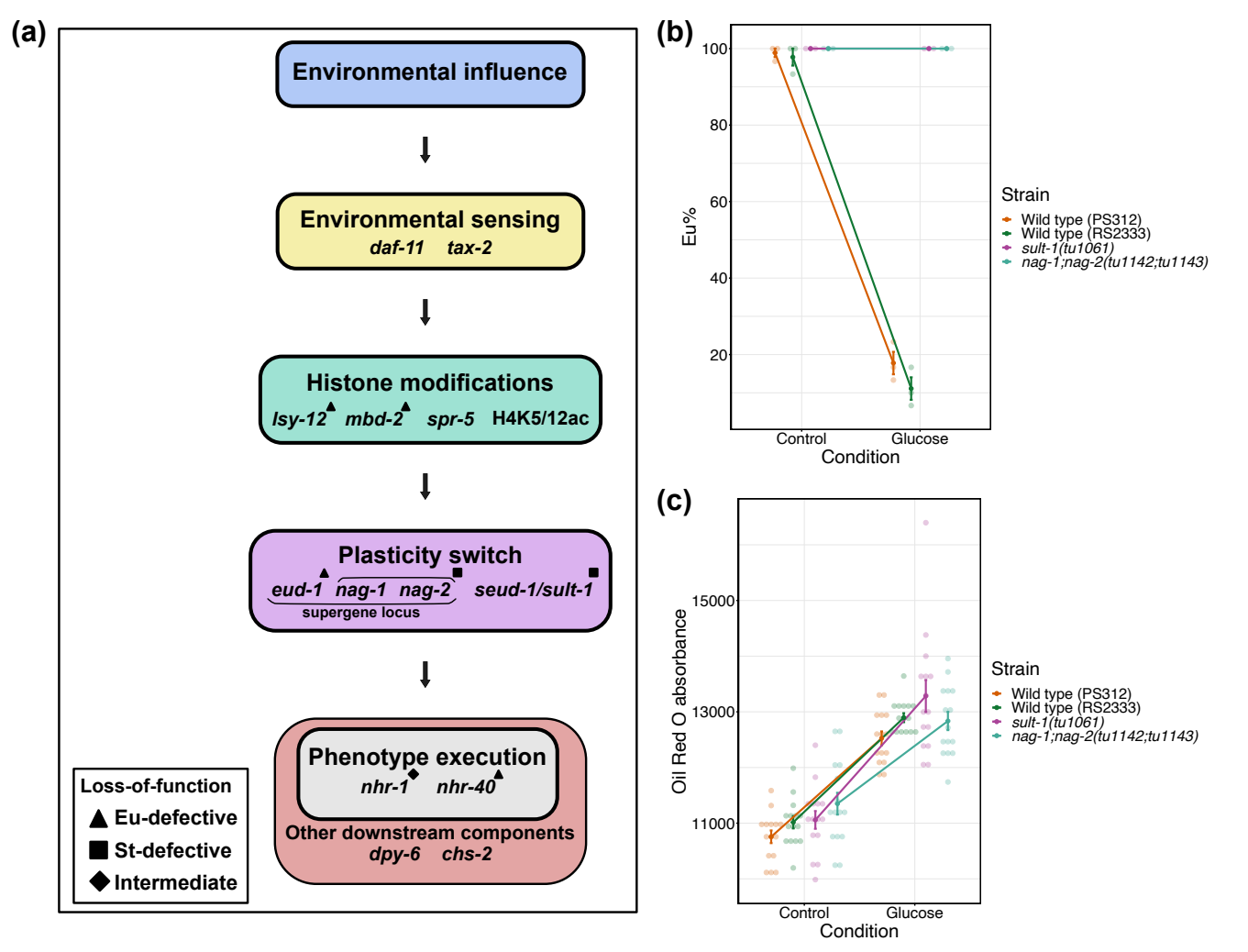

(a)
Eu
St
(b)
3 young adults with eggs
Count for mouth-form ratios
Oil Red O staining
Imaging and analysis
Incubation until F1 adults emerge (~5 days)
Control
+Tergitol
Oleic acid
+Tergitol
Control
Glucose
Control
Oleic acid
Control
Glucose
DIC
Blue
Red
Quantification

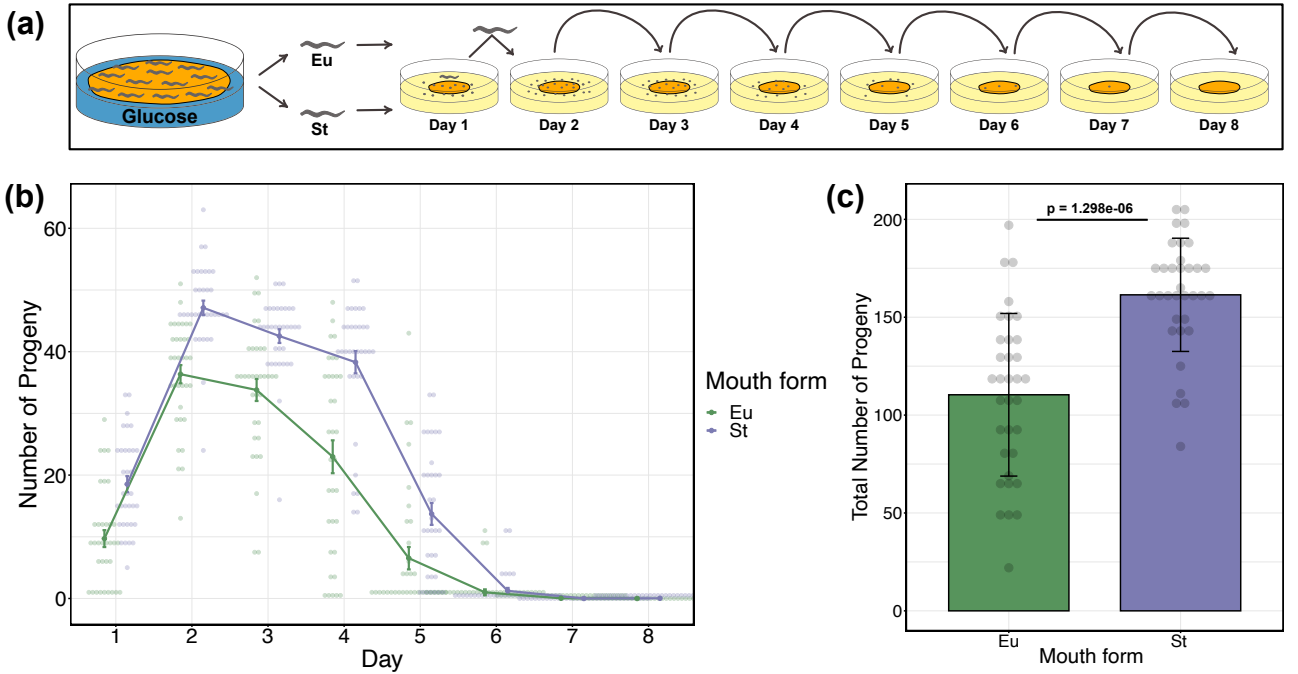

(a)
Oleic acid
Glucose
Eu%
100
80
60
40
20
0
Control
0.5mM
1mM
Control
100mM
150mM
Concentration
(b)
Oleic acid
Glucose
Oil Red O absorbance
20000
15000
10000
5000
0
p = 2.927e-11
p = 2.846e-07
Control
0.5mM
Control
100mM
Concentration
(c)
Control +Tergitol
Control
Oleic acid
Glucose


(a)
Incubation until F1 young adults without eggs emerge (~4 days)
Control +Tergitol
Control
Oleic acid +Tergitol
Oleic acid
Control
Control
Glucose
Glucose
RNA Extraction
150 young adults per condition with 2 technical replicates
(b)
Overrepresented KEGG pathways
Oleic acid
Spliceosome
Retinol metabolism
Peroxisome
Metabolism of xenobiotics by cytochrome P450
Lysosome
Glutathione metabolism
Fatty acid degradation
Drug metabolism−other enzymes
Drug metabolism−cytochrome P450
Biosynthesis of unsaturated fatty acids
Arachidonic acid metabolism
5 0 5 10 15 20
Glucose
Retinol metabolism
Protein processing in endoplasmic reticulum
Porphyrin metabolism
Peroxisome
Pentose and glucuronate interconversions
Metabolism of xenobiotics by cytochrome P450
MAPK signaling pathway
Fatty acid elongation
Fatty acid degradation
ECM−receptor interaction
Drug metabolism−cytochrome P450
Cysteine and methionine metabolism
Biosynthesis of unsaturated fatty acids
Ascorbate and aldarate metabolism
5 0 5 10 15
-log(Bonferroni-corrected P-value)
Number of differentially expressed genes present in the pathway


(a)
Saturated fatty acids
SBP-1
Desaturation
Delta-9 desaturases
Monounsaturated fatty acids
Desaturation
Elongation
Polyunsaturated fatty acids
Fat storage
(b)
Ppa-pddl-1▲
Ppa-pddl-2
Ppa-pddl-3▲
Ppa-pddl-4▲
fat-5
fat-6
fat-7
▲Constitutive expression throughout development.
(f)
Wild type (PS312)
- glucose
Wild type (PS312)
+ glucose
tu2028_Repaired
- glucose
tu2028_Repaired
+ glucose
tu2028
- glucose
tu2028
+ glucose
(c)
Control
Glucose
Eu%
Strain
PS312
tu2028
tu2029
tu2033
tu2034
tu2035
tu2030
tu2031
tu2032
Gene
Wild type
Ppa-pddl-1
Ppa-pddl-3
Ppa-pddl-4
(d)
Eu%
Condition
Control
Glucose
Strain
Wild type (PS312)
tu2028_Repaired
tu2028
(e)
Oil Red O absorbance
Condition
Control
Glucose
Strain
Wild type (PS312)
tu2028_Repaired
tu2028


(a)
▲Sterol carrier protein domain 2 (SCP2)
ACOX
MAOC-1
DHS-28
Ppa-dhs-28.1▲ (PPA20393)
Ppa-dhs-28.2 (PPA10787)
DAF-22
Ppa-daf-22.1▲
Ppa-daf-22.2
(b)
Eu%
100
80
60
40
20
Control
Glucose
Condition
Strain
Wild type (PS312)
Wild type (RS2333)
Ppa-dhs-28.1(tu1855)
Ppa-daf-22.1;Ppa-daf-22.2
(c)
Oil Red O absorbance
16000
14000
12000
Control
Glucose
Condition
Strain
Wild type (PS312)
Wild type (RS2333)
Ppa-dhs-28.1(tu1855)
Ppa-daf-22.1;Ppa-daf-22.2
(d)
Wild type (PS312)
- glucose
Wild type (PS312)
+ glucose
Wild type (RS2333)
- glucose
Wild type (RS2333)
+ glucose
Ppa-dhs-28.1(tu1855)
- glucose
Ppa-dhs-28.1(tu1855)
+ glucose
Ppa-daf-22.1;Ppa-daf-22.2(tu489;tu504)
- glucose
Ppa-daf-22.1;Ppa-daf-22.2(tu489;tu504)
+ glucose


(a)
Environmental influence
Environmental sensing
daf-11 tax-2
Histone modifications
lsy-12 mbd-2 spr-5 H4K5/12ac
Plasticity switch
eud-1 nag-1 nag-2 seud-1/sult-1
supergene locus
Phenotype execution
nhr-1 nhr-40
Other downstream components
dpy-6 chs-2
Loss-of-function
Eu-defective
St-defective
Intermediate
(b)
Eu%
100
80
60
40
20
Control
Glucose
Condition
Strain
Wild type (PS312)
Wild type (RS2333)
sult-1(tu1061)
nag-1;nag-2(tu1142;tu1143)
(c)
Oil Red O absorbance
15000
13000
11000
Control
Glucose
Condition
Strain
Wild type (PS312)
Wild type (RS2333)
sult-1(tu1061)
nag-1;nag-2(tu1142;tu1143)


(a)
Glucose
Eu
St
Day 1
Day 2
Day 3
Day 4
Day 5
Day 6
Day 7
Day 8
(b)
Number of Progeny
Day
Mouth form
Eu
St
(c)
p = 1.298e-06
Total Number of Progeny
Eu
St
Mouth form

High nutritional conditions influence feeding plasticity in *Pristionchus pacificus* and render worms non-predatory

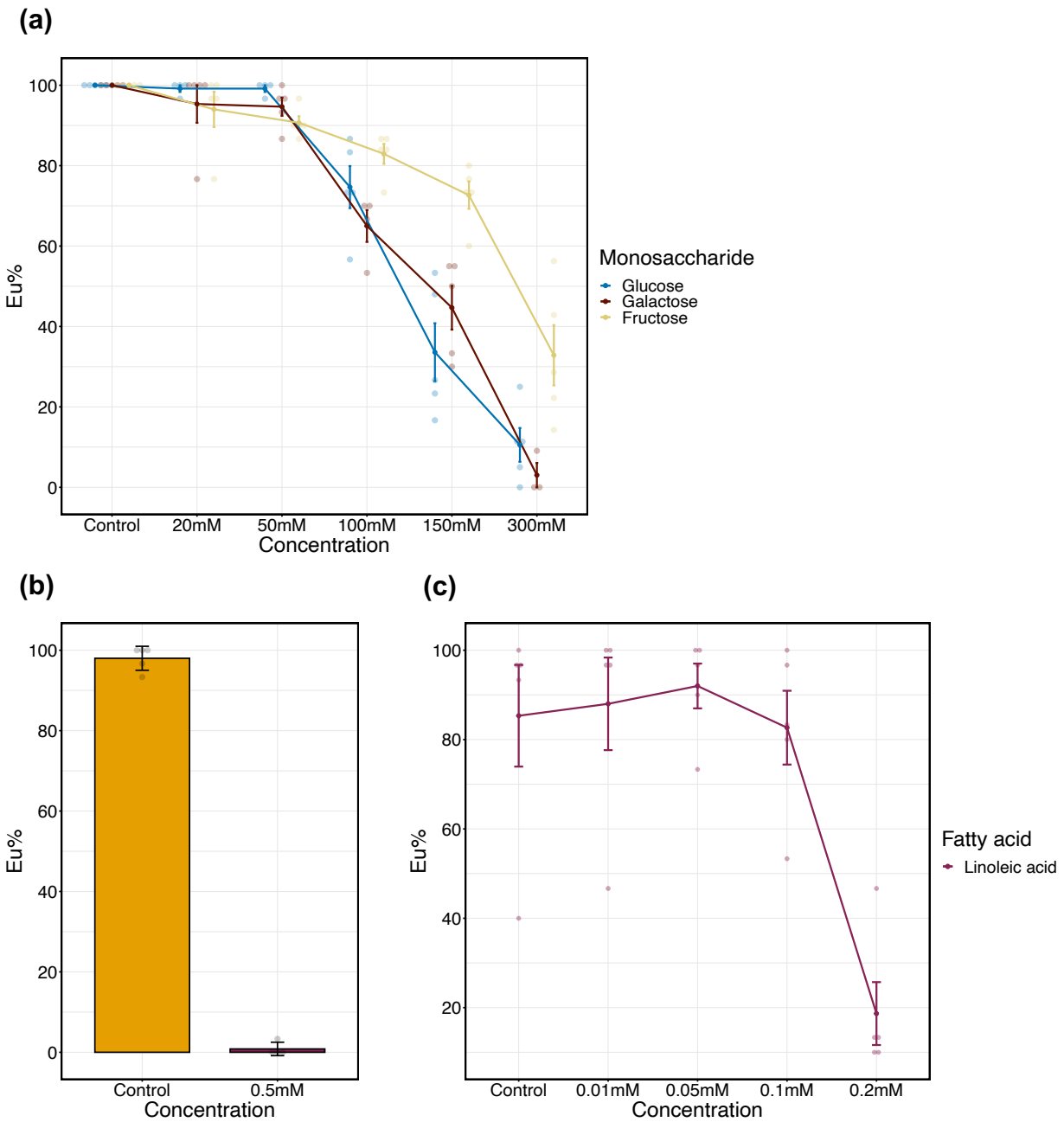
Supplementary Figures and Tables

Veysi Piskobulu[1], Marina Athanasouli[1], Hanh Witte[1], Christian Feldhaus[2], Adrian Streit[1] & Ralf J. Sommer[1,*]

[1] Max-Planck Institute for Biology Tübingen, Department for Integrative Evolutionary Biology, Max-Planck Ring 9, D-72076 Tübingen, Germany

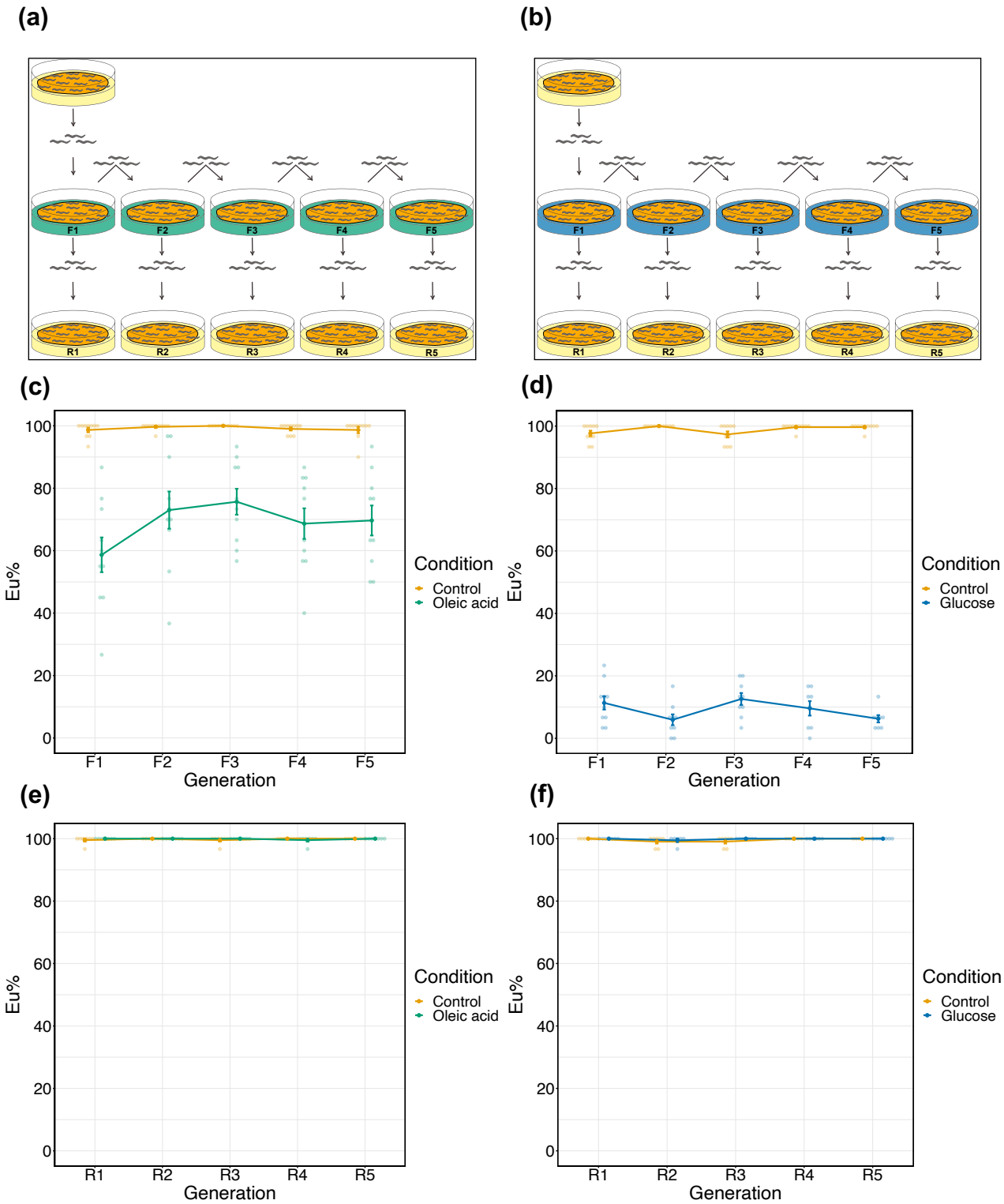
[2] Max-Planck Institute for Biology Tübingen, BioOptics Facility, Max-Planck Ring 5, D-72076 Tübingen, Germany

*author for correspondence at: ralf.sommer@tuebingen.mpg.de



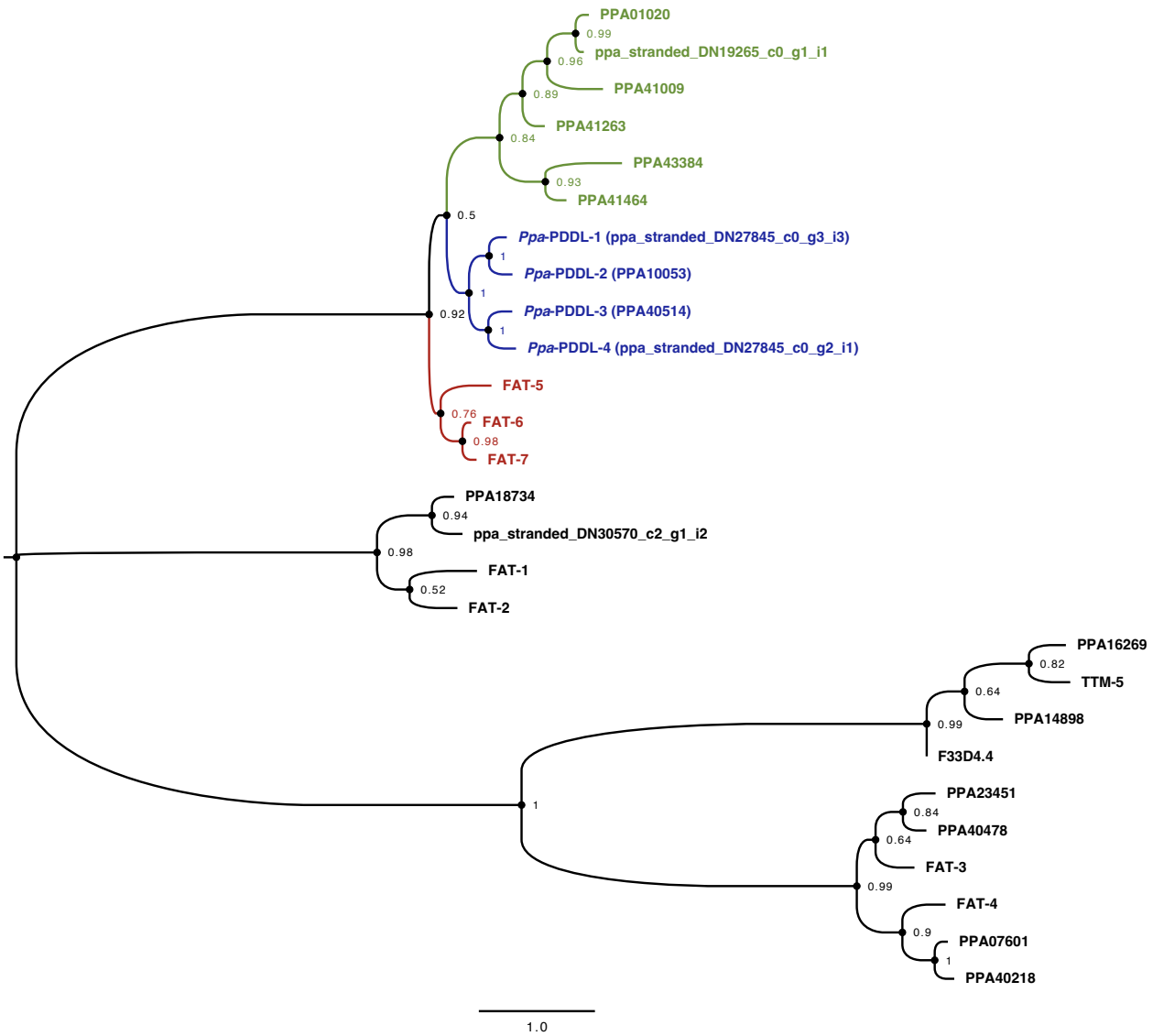
Supplementary Figure S1

Concentration dependent effect of supplements on mouth-form plasticity. (a) Eu percentages of worms grown on different concentrations of monosaccharides. In this pilot experiment, cultures were initiated by inoculating with 10 adult worms, allowing them to lay eggs for 2 hours. This method results in higher Eu percentages than inoculating with 3 worms. N ≥ 3 biological replicates per condition in each concentration. From each replicate (plate), 25-30 worms were scored, except for 300mM concentrations from which 4-28 worms were scored. Error bars represent s.e.m. (b) Effect of 0.5mM linoleic acid on mouth-form plasticity. N ≥ 4 biological replicates per concentration. From each replicate (plate), 28-30 worms were scored. Bars represent mean values of all replicates. Error bars represent s.d. (c) Linoleic acid concentration effect on mouth-form plasticity. N = 5 biological replicates per concentration. From each replicate (plate), 30 worms were scored. Error bars represent s.e.m. (a-c) Each faint data point represents a biological replicate (plate) scored for mouth-form ratio (Eu%).



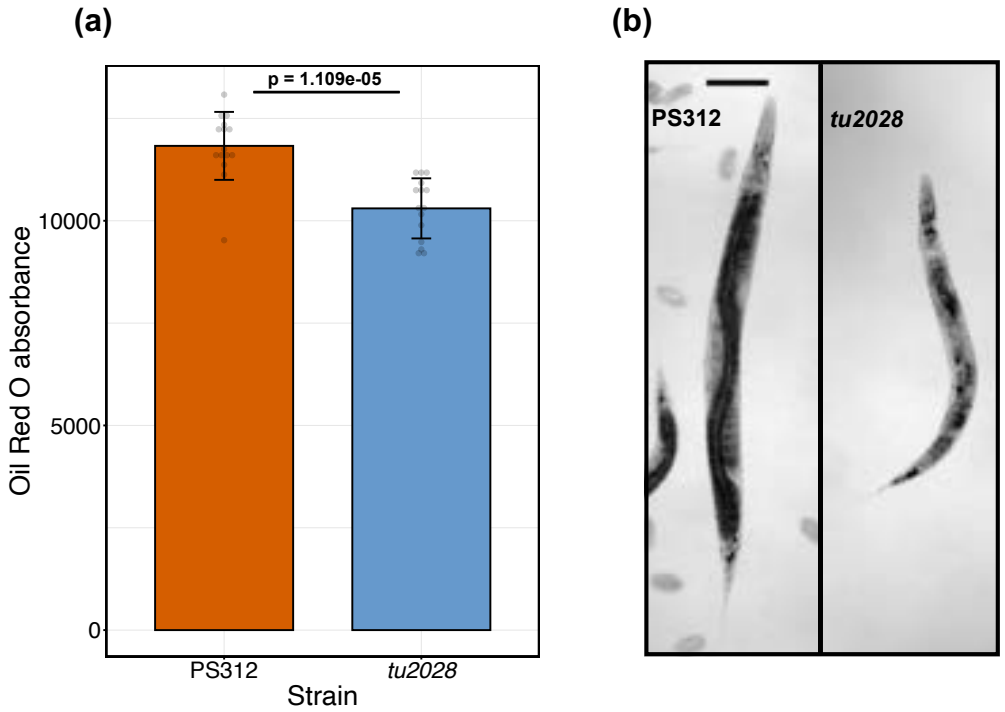
Supplementary Figure S2

Transgenerational effect of oleic acid and glucose supplementations on mouth-form plasticity. (a,b) Illustrations show the experimental design for studying the transgenerational effect of oleic acid (a) and glucose (b) supplementation on mouth-form plasticity. (a,b) The same experimental method was applied for respective control conditions. (c,d) Eu percentages of worms through generations (F1-F5) for oleic acid and control conditions (c); and for glucose and control conditions (d). (c) N = 10 biological replicates per condition for each generation. (d) N ≥ 8 biological replicates per condition for each generation. (e,f) Eu percentage of worms after reversal (F1-F5) for oleic acid and control conditions (e); and for glucose and control conditions (f). (e) N = 7 biological replicates per condition for each generation. (f) N ≥ 6 biological replicates per condition for each generation. (c-f) Each faint data point represents a replicate (plate), with 30 animals per plate being scored for mouth-form percentage (Eu%). Error bars represent s.e.m.



Supplementary Figure S3

Phylogenetic tree of all *P. pacificus* fatty acid desaturase domain-containing proteins with related *C. elegans* desaturases. A maximum likelihood phylogenetic tree, constructed with LG model, and 100 bootstrap replications. Bootstrap values are indicated next to branch nodes. Green and Blue colours denote *Pristionchus* delta-9 desaturase domain-containing proteins. Red colour denotes *C. elegans* delta-9 desaturases.



Supplementary Figure S4
Delta-9 desaturase mutant, *Ppa-pddl-1(tu2028)*, exhibits reduced lipid storage relative to wild type strain (PS312) on standard dietary condition. (a) ORO absorbance (RawIntDen/body area) obtained from wildtype (PS312) and *Ppa-pddl-1(tu2028)* strains. N = 15 per strain. P value is obtained from a two sample t-test. Each faint data point represents a worm. Bars represent mean values of all samples for each strain. Error bars represent s.d. (b) Representative images of ORO-quantified worms, indicating lipid storage profile. Images are acquired from the blue channel in grayscale. Lipid droplets appear dark. Scale bar is 100µm.

Supplementary Table S1. List of crRNA and primer sequences for genes that were studied for mutant analyses.

Gene name	Accession	crRNA	Forward Primer	Reverse Primer
Ppa-pddl-1	ppa_stranded_DN27845_c0_g3_i3	5′-CAGACATTACTATCCTCTAG-3′	5′-CCGTTAGAGTCTACTTCATGCTATGGAA-3′	5′-ATCAACCTGACCATATTTCAGTCTGACC-3′
Ppa-pddl-3	PPA40514	5′-CTATCTCCCCCTTGCGACTC-3′	5′-TTCTGATCTGTGGAACGACCCG-3′	5′-CACGGATTCGACGGGAGTGATG-3′
Ppa-pddl-4	ppa_stranded_DN27845_c0_g2_i1	5′-GTAATTCCCGTCTATTTCTG-3′	5′-TAACGATGTTTTCCTTCAGGTAATGGGC-3′	5′-CTGCTGCTTGTAGATTAGTCCATACAGA-3′
Ppa-dhs-28.1	PPA20393	5′-GGGGAGATCAAGGCAGCCGG-3′	5′-CGATATTGTTGCAGTGAACGAC-3′	5′-CTTCTAGTTACATCAGCTGTCTCG-3′

Supplementary Table S2. List of CRISPR mutants utilised in this study.

Gene Name	Gene Accession	Strain	Allele	Mutation Type	Mutation Location	Source
Ppa-pddl-1	ppa_stranded_DN27845_c0_g3_i3	RS4411	*tu2028*	7bp insertion	exon 4	This paper
Ppa-pddl-1	ppa_stranded_DN27845_c0_g3_i3	RS4412	*tu2029*	3bp insertion	exon 4	This paper
Ppa-pddl-3	PPA40514	RS4401	*tu2033*	5bp deletion	exon 5	This paper
Ppa-pddl-3	PPA40514	RS4402	*tu2034*	7bp deletion	exon 5	This paper
Ppa-pddl-3	PPA40514	RS4403	*tu2035*	7bp deletion	exon 5	This paper
Ppa-pddl-4	ppa_stranded_DN27845_c0_g2_i1	RS4398	*tu2030*	4bp deletion	exon 7	This paper
Ppa-pddl-4	ppa_stranded_DN27845_c0_g2_i1	RS4399	*tu2031*	10bp deletion	exon 7	This paper
Ppa-pddl-4	ppa_stranded_DN27845_c0_g2_i1	RS4400	*tu2032*	28bp insertion	exon 7	This paper
Ppa-dhs-28.1	PPA20393	RS4147	*tu1855*	4bp deletion	exon 3	This paper
Ppa-dhs-28.1	PPA20393	RS4138	*tu1856*	7bp deletion	exon 3	This paper
Ppa-dhs-28.1	PPA20393	RS4141	*tu1857*	8bp insertion	exon 3	This paper
Ppa-dhs-28.1	PPA20393	RS4140	*tu1858*	21bp deletion	exon 3	This paper
Ppa-dhs-28.1	PPA20393	RS4139	*tu1859*	19bp insertion	exon 3	This paper
Ppa-daf-22.1;Ppa-daf-22.2	ppa_stranded_DN16812_c0_g1_i1;PPA41516	RS2770	*tu489;tu504*	7bp deletion;7bp insertion	refer to Markov et al., 2016	Markov et al., 2016
sult-1	PPA12547	RS2974	*tu1061*	10bp deletion	exon 9	Namdeo et al., 2018
nag-1;nag-2	PPA06134;PPA34489	RS3195	*tu1142;tu1143*	1 SNP + 17bp insertion;2 SNPs + 9bp deletion	exon 5;exon 5	Sieriebriennikov et al., 2018

Interspecies systems biology links bacterial metabolic pathways to nematode gene expression, behavior, and survival

Marina Athanasouli[1], Tobias Loschko[1], Christian Rödelsperger[1,*]

[1]Department for Integrative Evolutionary Biology, Max Planck Institute for Biology, Max-Planck-Ring 9, 72076 Tübingen, Germany

[*] Corresponding author: christian.roedelsperger@tuebingen.mpg.de

Abstract

All animals live in tight association with complex microbial communities, yet studying the effects of individual bacteria remains challenging. Here, we exploit the system of the bacterial feeding nematode *Pristionchus pacificus*, where specific interactions can be studied in isolation. We sequenced the genomes of 84 *Pristionchus*-associated bacteria to investigate how differences in bacterial metabolic potential affect nematode traits. Specifically, we identified the Paerucumarin biosynthesis pathway as a strong candidate for nematode survival. Hypothesizing that bacterial production of coumarin derivatives contributed to increased lethality, we demonstrated nematicidal effects of two related compounds using supplementation assays. To explore metabolic interactions at a broader scale, we generated nematode transcriptomes for 38 bacterial diets and constructed a coexpression network that revealed distinct developmental and environmental signatures. Interspecies association studies resulted in a bipartite network with more than 2,800 interactions which may assist future studies to identify metabolites affecting various biological processes.

Keywords: Comparative genomics, *Pristionchus pacificus*, *Caenorhabditis elegans,* RNA-seq, microbe*,* microbiome, nematicide

Introduction

The human gut microbiome exemplifies the intricate associations between organisms and bacteria, highlighting the complexity of cross-kingdom interactions and their impact on host development and disease. These interactions are influenced by the heterogeneity of the gut microbiome, as well as variations in host diet and genetic background, making them challenging to study. In recent years, the nematode *Caenorhabditis elegans* has emerged as a powerful model for investigating host–microbiota interactions. This model allows for detailed examination of individual interactions through the use of monoxenic bacterial cultures, where worms are grown on single bacterial strains[1]. *C. elegans* engages with bacteria in multiple ways: it is a bacterial feeder, typically consuming the *Escherichia coli* OP50 strain in laboratory settings, and the bacterial diet significantly influences its development, as exemplified by the vitamin B_{12} production by *Comamonas* that impacts *C. elegans* development[2]. Moreover, *C. elegans* harbors a complex microbiome in its natural habitat, with certain bacteria enhancing the host's fitness under stress[3, 4], such as *Bacillus subtilis*, which provides resistance against Gram-positive pathogens when colonizing the nematode's gut[5].

The nematode *Pristionchus pacificus* was initially introduced as a satellite model organism for comparative studies with *C. elegans*. Having diverged approximately 130-310 million years ago[6], these species share several traits, including transparency, short generation times, and hermaphroditism, which make them valuable models for genetic research. More recently, *P. pacificus* has gained prominence as a model for studying the interplay between environment and genotype in producing diverse

phenotypes. A key example of this is the species' mouthform plasticity, where genetically identical individuals can develop either a narrow mouth suited for bacterial feeding or a wider mouth with an additional tooth for predation. This plasticity, regulated both genetically and environmentally, varies among natural isolates and is influenced by dietary factors. For instance, the reference *P. pacificus* strain PS312 predominantly produces the predatory morph when fed *E. coli* OP50 or other bacteria but shifts to the non-predatory morph when consuming a specific *Cryptococcus* yeast strain[7]. Additionally, factors such as specific pheromones and alternative culture conditions have been shown to affect mouthform plasticity[8, 9]. Despite the extensive characterization of the regulatory network controlling this developmental switch[10-14], little is known about how *P. pacificus* generally processes environmental signals. Our focus is on the nematode's interactions with environmental microbiota, as *P. pacificus* is often found in decaying scarab beetles, where it feeds on various microbes[15, 16]. Notably, one major difference between *C. elegans* and *P. pacificus* is that the latter lacks a specific morphological structure, the grinder, that is used by *C. elegans* to physically lyse the bacteria. This likely has widespread consequences on the way *P. pacificus* interacts with bacteria including pathogen susceptibility[17, 18]. Previously, we isolated and cultured 136 bacterial strains from *Pristionchus*-associated environments and demonstrated that *P. pacificus* exhibits differential development, behavior, and survival in response to these bacteria[19, 20]. Notably, we found that bacteria-derived vitamin B_{12} not only accelerates development but also enhances predatory behavior.

In this study, we apply a systems biology approach to investigate the interactions between *Pristionchus pacificus* and bacterial food sources derived from the nematode's

natural environment. Specifically, based on whole-genome sequencing, we explore the variation in metabolic potential among these bacteria and we test if this variation can be associated with the differential responses of the nematodes when exposed to these bacteria. Our main goal is to identify potential interactions between bacterial metabolic pathways and nematode phenotypes which include previously generated data on survival and chemotaxis behavior[19]. We perform supplementation experiments to further support that a specific group of metabolites could potentially contribute to the high toxicity of some bacteria. Further, to investigate metabolic interactions at a broader scale, we generated transcriptomic profiles of worms grown on 38 different bacteria and tested for associations between *P. pacificus* coexpression modules and bacterial metabolic pathways. This resulted in a large bipartite network involving more than 2,800 interactions, which highlights the complexity of the whole system and serves as a resource for future studies elucidating how specific metabolites control developmental and metabolic processes in *P. pacificus*.

Results

Sequencing of 84 bacteria establishes the genomic basis to study host microbe interactions

In order to link genomic variation in bacteria to differential responses in *P. pacificus* nematodes, we sequenced 84 strains from a large bacterial collection isolated previously from *Pristionchus*-associated environments[19]. This collection mostly comprised Proteobacteria as well as some members of the phyla Firmicutes, Actinobacteria, and Bacteroidetes. Among the Proteobacteria, the most deeply sampled families constitute Enterobacteriaceae, Pseudomonadaceae, and Moraxellaceae. 84 bacterial strains were selected for whole genome sequencing based on criteria such as fast growth, easiness of DNA extraction, low propensity of contamination. (Supplemental Table S1). The resulting genome assemblies were highly complete as indicated by a median single-copy BUSCO completeness of 99% with an interquartile range (IQR) of 98-100%. The largest parts of the bacterial genomes could be assembled into contigs spanning more than 100kb. Specifically, the N50 value, which indicates the minimum contig length in the set of largest contigs that account for at least half of the genome, was 324kb (IQR=107-498kb) and median genome assembly size was 4.9Mb (IQR=4.4-5.6Mb). Gene annotation yielded a median of 4,562 gene models per assembly (IQR=4,184-5,225) with a median single-copy BUSCO completeness level of 99% (IQR=98-100%). The complete overview about these different quality measures can be found in Supplemental Table S1. This set of bacterial genomes builds the basis to study host microbe interaction in the *Pristionchus* system.

Pristionchus-associated microbiota harbor previously uncharacterized bacterial strains

The predicted proteomes of all 84 genomes were taken to reconstruct a bacterial phylogeny (see Methods) (Fig. 1). The different lineages in the resulting species tree

were mostly consistent with previous genus assignments based on 16S sequencing[19]. To more accurately classify the strains, we performed homology searches against the non-redundant version of NCBI and determined the nomenclature of our bacteria by assigning each strain to the bacterial genus with the majority of best hits (Supplemental Table S2). During this process, we observed that the median percentage of identity of the strains *Sphingobacterium* L2, *Pseudomonas* L74 and *Erwinia* V71 to their respective best hits in NCBI was lower than 90%. Additionally, we were unable to definitively determine the genus for the closely related strains V69 and V91, despite a median identity of 90.6% and 96% respectively. We hypothesized that these bacterial strains could either represent novel isolates or have no associated whole genome sequencing data in the NCBI database. To confirm our hypothesis regarding the potentially novel bacteria, we searched the literature for recent, comprehensive phylogenies of the genera[21-24] and recreated these phylogenies including our strains and their best hits in the NCBI database. The reconstructed phylogenies for the three bacterial strains with low median identity (L2, L74, V71) revealed that our strains are outgroups to the clusters containing their best hits in their respective phylogenetic trees but were still more closely related to them compared to any other genus (Supplemental Fig. 1A-C). The reconstructed phylogeny including V69 and V91 exhibited the same patterns and indicated that both bacteria should be classified as *Phytobacter* strains (Supplemental Fig. 1D). These findings support that *Pristionchus*-associated microbiota constitute a previously undersampled environment that harbors novel bacterial strains and species.

Closely related bacteria exhibit substantial variation in metabolic potential

In order to characterize the variation of metabolic potential among members of the *Pristionchus*-associated microbiota, we reconstructed metabolic networks for the 84 bacterial genomes with the gapseq approach using existing pathway information from the MetaCyc database. MetaCyc was preferred to KEGG due to the higher detail it offers with regards to the available pathways. We then investigated the absence/presence patterns of metabolic pathways among the 84 bacteria and defined metabolic pathways groups (MPGs) based on identical patterns (Fig. 2A). This reduced the number of total metabolic pathways from 2,902 (Supplemental Table S3) to 715 MPGs. The two most abundant MPGs denote 1,832 and 45 individual pathways that are either absent or present in all strains (Fig. 2A). While most pathways that were not detected in our data correspond to pathways that are known only from eukaryotes, examples of the core metabolic pathways that are present in all strains relate to the TCA cycle, biosynthesis of several amino acids as well as fatty acid biosynthesis and elongation pathways, pointing towards necessary pathways for the survival of bacterial cells (Supplemental Table S2). The third most abundant MPG 3 denotes 13 pathways that are only missing in *Enterococcus* strain TSA3A which is one of only three Gram positive strains in our collection (Fig. 1). Its genome assembly has good contiguity (N50=234kb) and is highly complete (BUSCO completeness = 98%), which suggests that this pattern is unlikely due to a lower quality of this particular genome. Also, more detailed analysis of these 13 candidate pathways revealed that while some pathways are partially present other pathways are completely absent. Other MPGs show strong phylogenetic signatures such as MPG 7 and 13 that are restricted to the genera *Achromobacter* and *Pseudomonas*, respectively. We would therefore conclude that

most of the variation in metabolic potential rather reflects genome evolution than technical issues. In order to better quantify the metabolic variation in our bacterial collection, we counted the number of pathway differences for pairwise comparison within and across genera. This showed significantly fewer differences in the metabolic potential for strains of the same genus (Fig. 2B). However, even strains of the same genus can differ in dozens of metabolic pathways, which can potentially explain the previously observed variability in the molecular and phenotypic responses of *P. pacificus* nematodes[19, 25].

Interspecies association studies identify candidate pathways affecting nematode survival and behavior

During the initial culture-based characterization of *Pristionchus*-associated microbiota, Akduman et. al. performed assays for chemotaxis behavior and survival with over a hundred bacterial strains[19], many of which are part of our collection (Fig. 3A). In a pilot experiment, we employed our genomic data to screen for candidate pathways, and their associated metabolic products, that could possibly impact nematode survival and behavior. To this end, we tested for associations between the presence or absence of a MPG with the behavioral and survival phenotypic data mentioned above. With this strategy, we identified 24 MPGs exhibiting significant association with chemotaxis ($P < 0.05$, Wilcoxon-test, Supplemental Fig. S2 and Table S4) and 40 MPGs that are associated with survival (Fig. 3B, Supplemental Table S5). Among these candidate pathways, we noticed a dramatic difference in survival associated with between presence or absence of the paerucumarin and rhabduscin biosynthesis pathway.

Coumarin-derivatives and rhabduscin have frequently been reported to exhibit nematicidal activity[26-29], which suggests that similar bacteria-derived metabolites contribute to the reduced survival of *P. pacificus* nematodes. To complement this analysis with an additional screen for other virulence factors, we combined the predicted proteomes of the 84 bacteria with the virulence factor database (VFDB)[30] and performed a similar association study. This yielded 15 orthologous clusters with significant association (FDR-corrected P-value < 0.05), among which members of the paerucumarin biosynthesis gene cluster showed the strongest and most consistent effect (Supplemental Fig. S3). Closer examination of the presence of these genes in our data set revealed that these orthologs are exclusively present in strains of the genus *Serratia.* Since the paerucumarin gene cluster has been initially characterized in *Pseudomonas*[31], the homologs of those genes in *Serratia* will not necessarily produce the exactly same compound. Thus, hypothesizing that paerucumarin, rhabduscin, or similar metabolites could contribute to the pathogenicity of *Serratia* strains, we tested three coumarin-derivatives, 4-Hydroxycoumarin, 7-Methoxycoumarin, and Xanthotoxin, that were previously shown to possess some nematicidal activity and that are commercially available for their effect against *P. pacificus*[27-29]. Among those, 7-Methoxy coumarin and Xanthotoxin caused substantial lethality at 1mM concentrations ($P < 0.05$, t-test, Fig. 3C). This supports that some coumarin derivatives have nematicidal activity against *P. pacificus* and that such compounds could potentially contribute to the increased pathogenicity of *Serratia* strains. However, further experimental work will be needed to identify which metabolites are produced by the *Serratia* strains and whether they exhibit indeed nematicidal activity.

Closely related bacteria can induce vastly different transcriptomic responses in nematodes

Having demonstrated the utility of the bacterial genomes to uncover candidate pathways impacting nematode survival and behavior, we now focus on our main objective to identify interactions between bacterial metabolic pathways and regulatory and metabolic response networks in the worm. To this end, we generated nematode transcriptome profiles for a subset of 38 dietary bacteria that belong to the three classes, Alpha-, Beta- and Gammaproteobacteria, and represent 16 genera. These 38 bacteria were selected as they supported complete development from egg to adult for at least two generations, which was not the case for all the strains. For each bacterial strain, we generated two biological replicates by collecting F1 worms 72 hours after egg-laying (see *Methods*). While most animals should have reached adulthood at this time under standard *E. coli* OP50 diet[32], we cannot exclude that some bacterial diets either accelerate or slow down development leading to a shift of the transcriptomic profiles towards either older or younger developmental stages. We decided to sequence the transcriptomes of mixed-stage populations as we observed previously that even when adult worms are manually picked from plates, a developmental signature is still visible in the resulting RNA-seq data. This is because worms will still develop at different speeds and chronological age might not necessarily correspond to developmental age[25]. As a consequence, differences in stage composition likely explain the larger extent of transcriptomic variation in our study, i.e. quite a number of samples showed correlation coefficients $r < 0.9$ (Fig. 4A), which was not the case in our previous study

where we manually picked adult worms on 24 diverse bacteria[25]. We note that biological replicates do not always group together suggesting that the effect of experimental variation on the overall transcriptome can in individual cases be larger than the effect of different bacterial diets. However, generally, we observed that replicates of the same bacterial diet exhibit less variation and thus are more similar than transcriptomic profiles from different bacterial diets (Fig. 4B). With regard to the taxonomic grouping of the bacteria, we observe that bacterial strains of the same genus show more similar transcriptomic responses in the worms than members of different genera (Fig. 4C). However, even in comparisons within the same genus, transcriptomic responses can differ substantially with correlation coefficients < 0.9. This recapitulates our previous finding of substantial variation in bacteria from the same genus with regard to metabolic potential (Fig. 2B).

Specific network modules respond to signals from diverse bacteria

While the above mentioned transcriptomic analysis captures overall transcriptomic similarities, specific effects of individual bacteria might be overshadowed by global patterns. To investigate the effects of individual bacteria on smaller gene sets that may point towards specific regulatory or metabolic network modules, we constructed a gene coexpression network from the RNA-seq data following the protocol and parameters from our previous related work[25]. Overall, our coexpression network captures 24,375 (84%) of *P. pacificus* genes, of which 13,972 are found in the 60 largest modules with more than 20 genes. Visualization of expression values in these modules across the different microbiota demonstrated that individual modules appear to be induced by specific bacteria (Fig. 5). For example, the second largest module 2 shows high

expression on most *Pseudomonas* strains whereas module 7 exhibits highest expression only on the *Pseudomas protegens* strain LRB88. Note that some of these patterns may represent indirect effects of altered development on diverse bacterial diets. This is shown by distinct expression signatures of specific coexpression modules when overlaying them with the developmental transcriptomic data of *P. pacificus* (Supplemental Fig. 4)[32]. For example, module 1 peaks late in development close to adulthood and is slightly preceded by a peak of module 2. However, no matter if the expression of those genes is a direct or indirect consequence, it can be interpreted as a specific response of *P. pacificus* nematodes to a distinct microbiota[25].

Most coexpression modules can be functionally annotated

To functionally annotate the coexpression modules and to gain insight into the regulation of metabolic pathways, we analyzed the top 60 modules of the network that contained more than 20 genes and performed enrichment analyses, similar to previous work. This included overrepresentation of protein domains[33], regionally enriched genes as identified from spatial transcriptomics[34], KEGG pathways[35], and previous gene expression studies focusing on sex-biased genes[36], intestinal transcriptome[18], and developmental oscillations[32] (Supplemental Table S6). This allowed the labeling of 48 modules with biological terms based on the strong enrichment patterns (Fig. 6, Supplemental Table S7). For example, the largest module 1 (OOGEN_1) is overrepresented in hermaphrodite-biased and gonad-enriched genes (Fig 6). It also includes known germline expressed markers like *spo-11*, *dmc-1*, *hcp-4* (*cenp-c*), *syp-4*, and *hop-1*[37], indicating that it largely represents oogenesis. However, it also shows

some overrepresentation of neuropeptides and head-enriched genes suggesting that also a portion of neuronal genes are captured in this module. The second largest module 2 (SPERM_2) seems to be more homogenous, as all overrepresented terms point towards spermatogenesis (Fig. 6). The assignments of module 1 and 2 are also in line with their developmental signature (Supplemental Fig. 4) as sperm are generated during late larval development whereas oocytes are produced during adulthood. While many modules show less specific signals, individual modules show specific enrichments with individual pathways or anatomical regions. For example, module 36 (GLAND_36) is strongly associated with gland cell specific expression[13]. We used the microbial transcriptomes (Fig. 5), to label a handful of modules that did not show any enrichment in the above mentioned data set (Supplemental Table S7). Thus, module 13 (LRB7_13) is strongly upregulated on *Serratia quinivorans* strain LRB7 and module 27 (LRB111_27) exhibits highest expression on *Proteus terrae* strain LRB111. Altogether, the characterization of the 60 largest coexpression modules functionally annotates a large portion of the *P. pacificus* gene set which includes many lineage-specific genes for which no homologs are known in *C. elegans*[16, 25].

Interspecies association studies traces the response of coexpression modules to specific bacterial pathways

A main goal of this study was to trace back the transcriptomic response of individual coexpression modules in *P. pacificus* to candidate metabolic pathways in the bacteria. We hypothesized that some of the MPGs show strong associations with our coexpression modules. To this end, we combined the bacterial MPGs and the nematode

coexpression modules into a bipartite network and calculated the interaction of each MPG with the modules based on the significance of the overlap between the genes within a coexpression module and the genes that exhibit differential expression in response to the absence or presence of an MPG. Altogether, we observed more than 2,800 interactions that involved 620 MPGs and 42 coexpression modules (Supplemental Table S8). The module with the highest number of predicted interactions was intestinal module 14 (GUT_14). Furthermore, the majority of coexpression modules interacting significantly with at least 100 MPGs are intestinal, a finding highlighting the important role of the intestine and is consistent with recent studies[38]. To visualize the strongest interactions in our enriched dataset, we chose the 50 MPGs with the lowest Bonferroni-corrected p-values (Fig. 7). Surprisingly, the largest module of the network (OOGEN_1) exhibited interactions with only 46 MPGs (Supplemental Table S8), of which 21 are highly significant (Fig. 7). On the contrary, module 2 (SPERM_2) was found to interact with 26 MPGs overall, 22 are highly significant. Regarding the highly interacting MPGs, three showed the most interactions with our coexpression modules: formaldehyde biosynthesis from methylamine (PWY-6965), L-lysine degradation to glutaryl-CoA (PWY-6328) and putrescine degradation to 4-aminobutanoate (PUTDEG-PWY). Altogether, the bipartite network indicated the environmentally sensitive modules as well as their potential triggers.

Discussion

All animals live in tight associations with microbial communities and bacteria can have a significant impact on the biology of their host. In nematodes, bacteria have been demonstrated to affect several traits including development, pathogen resistance, and

behavior[2, 5, 20, 39, 40]. After observing such intriguing phenotypes, the key question arises what bacterial signal triggers this response. While one major advantage of the nematode system is that individual interactions can be studied in isolation, we combine data from many different bacteria to ask broader questions such as to what extent does the nematode's transcriptomic response vary across bacterial diets. Can these responses be traced back to specific metabolic pathways in the bacteria? These systems level questions distinguish our work from many other studies that focused on specific effects of individual bacteria[2, 5, 20, 39, 40]. By investigating a large number of bacterial diets, we aimed to have higher statistical power to also detect a greater number of metabolic interactions between dietary bacteria and *P. pacificus* nematodes. It should be noted that the capability to detect metabolic interaction will depend on the sampling of bacteria and the phylogenetic distribution of the pathway. For this reason, we have tried to maximize the number of sequenced samples, which resulted in a collection of 84 bacterial genomes that serve as a genomic basis for the current and future research on host-microbe interactions. At the same time, this data allowed us to investigate the extent of variation in metabolic potential within and between bacterial genera, which revealed considerable amounts of intra-genus level variation in bacterial metabolic potential as well as in the nematode transcriptomic response. Following a similar approach as Zimmermann et al.[41], who associated metabolic potential of members of the *C. elegans* microbiome with specific traits, we made use of previously generated phenotypic data to perform an inter-species association study to identify potential connections between bacterial pathways and nematode behavior and survival.

In particular, we could confirm the nematicidal activity of a certain coumarin-derivative on *P. pacificus*.

The generation of RNA-seq data of *P. pacificus* nematodes for 38 different bacterial diets represents to our knowledge the largest transcriptomic study in the field of interactions between nematodes and bacteria. This transcriptomic data placed us in a unique position to perform a large-scale interspecies association study, where we tested the absence/presence pattern of an MPG for a significant expression response of a given coexpression module. This identified more than 2,800 associations between bacterial metabolic pathways and specific coexpression modules. Each of these individual interactions implies a causal relationship between one or several metabolites in a bacterial metabolic pathway and the response of a specific coexpression module in the worm. The shear number of associations points towards a highly complex network of metabolic interactions between *P. pacificus* and its dietary bacteria. However, the actual number might be smaller, as similar absence/presence patterns among MPGs will likely predict the same interacting coexpression modules. At the same time, the number of interactions will be underestimated due to the lack of statistical power to detect interactions for MPGs that show very little variation among the bacterial genomes. Another uncertainty comes from the fact that some of the transcriptomic responses may represent indirect effects of altered development. Thus, future studies could employ additional transcriptomic data at a much higher temporal resolution to dissect immediate responses in response to bacteria from indirect effects that are mediated through altered development. This will simplify the selection of candidate metabolites for supplementation experiments in order to validate predicted interactions.

Knowing how different metabolites affect the transcriptional state of the worm, will ultimately allow us to manipulate the system in a controlled manner.

Methods

Key resources table

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
Bacterial strains		
Serratia marcescens	Akduman et al.[19]	LRB2
Acinetobacter courvalinii	Akduman et al.[19]	LRB3
Proteus terrae	Akduman et al.[19]	LRB4
Serratia quinivorans	Akduman et al.[19]	LRB7
Pseudomonas putida	Akduman et al.[19]	LRB8
Chryseobacterium sp. G0186	Akduman et al.[19]	LRB11
Acinetobacter calcoaceticus	Akduman et al.[19]	LRB14

Hafnia alvei	Akduman et al.[19]	LRB17
Stenotrophomonas lactitubi	Akduman et al.[19]	LRB19
Morganella morganii	Akduman et al.[19]	LRB20
Acinetobacter pittii	Akduman et al.[19]	LRB21
Pseudomonas fluorescens	Akduman et al.[19]	LRB26
Comamonas thiooxydans	Akduman et al.[19]	LRB28
Acinetobacter gerneri	Akduman et al.[19]	LRB29
Acinetobacter guillouiae	Akduman et al.[19]	LRB31
Acinetobacter guillouiae	Akduman et al.[19]	LRB32
Acinetobacter seifertii	Akduman et al.[19]	LRB33
Enterobacter asburiae	Akduman et al.[19]	LRB34
Serratia fonticola	Akduman et al.[19]	LRB40

Exiguobacterium sp. RIT341	Akduman et al.[19]	LRB41
Providencia rettgeri	Akduman et al.[19]	LRB44
Chryseobacterium sp. MEBOG06	Akduman et al.[19]	LRB45
Cronobacter dublinensis	Akduman et al.[19]	LRB46
Pseudomonas putida	Akduman et al.[19]	LRB53
Kurthia gibsonii	Akduman et al.[19]	LRB56
Acinetobacter sp. I-MWF	Akduman et al.[19]	LRB59
Pseudomonas putida	Akduman et al.[19]	LRB63
Serratia quinivorans	Akduman et al.[19]	LRB69
Serratia quinivorans	Akduman et al.[19]	LRB70
Serratia marcescens	Akduman et al.[19]	LRB72
Pseudomonas fluorescens	Akduman et al.[19]	LRB73

Acinetobacter guillouiae	Akduman et al.[19]	LRB74
Acinetobacter gerneri	Akduman et al.[19]	LRB80
Morganella morganii	Akduman et al.[19]	LRB81
Pseudomonas fluorescens	Akduman et al.[19]	LRB83
Pseudomonas protegens	Akduman et al.[19]	LRB88
Empedobacter falsenii	Akduman et al.[19]	LRB89
Acinetobacter sp. I-MWF	Akduman et al.[19]	LRB90
Enterobacter asburiae	Akduman et al.[19]	LRB98
Comamonas thiooxydans	Akduman et al.[19]	LRB103
Empedobacter falsenii	Akduman et al.[19]	LRB104
Empedobacter falsenii	Akduman et al.[19]	LRB108
Acinetobacter gerneri	Akduman et al.[19]	LRB109

Proteus terrae	Akduman et al.[19]	LRB111
Proteus terrae	Akduman et al.[19]	LRB112
Acinetobacter lwoffii	Akduman et al.[19]	LRB113
Serratia quinivorans	Akduman et al.[19]	LRB119
Chryseobacterium lactis	Akduman et al.[19]	LRB120
Acinetobacter calcoaceticus	Akduman et al.[19]	LRB97b
Sphingobacterium sp.	Akduman et al.[19]	L2
Achromobacter pestifer	Akduman et al.[19]	L23
Agrobacterium tumefaciens	Akduman et al.[19]	L27
Comamonas sp. BIGb0124	Akduman et al.[19]	L28
Agrobacterium tumefaciens	Akduman et al.[19]	L29
Stenotrophomonas rhizophila	Akduman et al.[19]	L31

Achromobacter pestifer	Akduman et al.[19]	L35
Agrobacterium tumefaciens	Akduman et al.[19]	L54
Achromobacter pestifer	Akduman et al.[19]	L55A
Pseudomonas sp.	Akduman et al.[19]	L74
Achromobacter pestifer	Akduman et al.[19]	L91
Achromobacter pestifer	Akduman et al.[19]	L93
Achromobacter kerstersii	Akduman et al.[19]	L100
Serratia marcescens	Akduman et al.[19]	V6
Citrobacter koseri	Akduman et al.[19]	V8
Klebsiella spallanzanii	Akduman et al.[19]	V14
Serratia marcescens	Akduman et al.[19]	V15
Citrobacter koseri	Akduman et al.[19]	V20

Klebsiella michiganensis	Akduman et al.[19]	V39
Klebsiella spallanzanii	Akduman et al.[19]	V46
Raoultella sp. 10-1	Akduman et al.[19]	V48
Citrobacter koseri	Akduman et al.[19]	V52
Citrobacter koseri	Akduman et al.[19]	V61
Serratia marcescens	Akduman et al.[19]	V64
Phytobacter sp.	Akduman et al.[19]	V69
Erwinia sp.	Akduman et al.[19]	V71
Citrobacter koseri	Akduman et al.[19]	V72
Citrobacter koseri	Akduman et al.[19]	V88
Phytobacter sp.	Akduman et al.[19]	V91
Enterobacter mori	Akduman et al.[19]	V94

Klebsiella spallanzanii	Akduman et al.[19]	V97
Erwinia aphidicola	Akduman et al.[19]	NA6
Enterobacter mori	Akduman et al.[19]	SFM3
Enterococcus casseliflavus	Akduman et al.[19]	TSA3A
Escherichia coli OP50		
Biological samples		
Pristionchus pacificus	Department for Integrative Biology, MPI Tübingen	PS312
Chemicals		
Xanthotoxin	TCI Deutschland	CAS: 298-81-7
4-Hydroxycoumarin	Sigma Aldrich	CAS: 1076-38-6
7-Methoxycoumarin	Sigma Aldrich	CAS: 531-59-9
Commercial assays		

MasterPure Complete DNA and RNA Purification Kit	Epicentre	Cat. No.: MCD85201
RNA Clean & Concentrator 25	Zymo Research	Cat. No.: R1018
Software and algorithms		
Affinity Designer	Affinity	
BUSCO	Manni et al.[42]	https://busco.ezlab.org/
DIAMOND	Buchfink et al.[43]	https://github.com/bbuchfink/diamond
DESeq2	Love et al.[44]	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
eggnog-Mapper	Cantalapiedra et al.[45]	https://github.com/eggnogdb/eggnog-mapper
featureCounts	Liao et al.[46]	https://subread.sourceforge.net/
gapseq	Zimmermann et al.[47]	https://github.com/jotech/gapseq
MCL	van Dongen[48]	https://github.com/micans/mcl
PROKKA	Seemann[49]	https://github.com/tseemann/prokka

Quast	Gurevich et al.[50]	https://quast.sourceforge.net/quast.html
Orthofinder	Emms and Kelly[51]	https://github.com/davidemms/OrthoFinder
SPAdes	Prjibelski et al.[52]	https://github.com/ablab/spades
STAR	Dobin[53]	https://github.com/alexdobin/STAR
Traitar3	Weimann et al.[54]	https://github.com/nick-youngblut/traitar3

Resource availability

Materials availability

This study did not generate new unique resources and reagents.

Data and code availability

The raw sequencing data for the bacterial genomes, the genome assemblies and the raw read counts of the transcriptomic profiles from this study have been deposited at the European Nucleotide Archive under Bioproject PRJEB80633.

Experimental model details

Bacterial culture conditions

All bacterial strains were initially restreaked from a glycerol stock on Lysogeny Broth (LB) plates and grown overnight in LB. Bacteria were grown at 30 °C or 37 °C depending on the species.

Nematode growth conditions

The wild type strain of *Pristionchus pacificus* (PS312) was maintained at 20 °C on nematode growth medium (NGM) seeded with *Escherichia coli* OP50 before use. From every generation, three adults were transferred to fresh NGM plates with a worm pick.

Method details

Whole genome sequencing of bacterial strains

All bacterial strains to be sequenced were grown overnight in LB in 15 ml falcon tubes and DNA samples were obtained using the Epicentre MasterPure Complete DNA and RNA Purification Kit (Illumina, San Diego, USA). DNA libraries were prepared using the Illumina DNA Prep kit according to the manufacturer's guidelines. The libraries were quantified using both a Qubit 2.0 Fluorometer (Thermo Fisher Scientific, Waltham, USA)

and a Bioanalyzer (Agilent Technologies, Santa Clara, California) and normalized to 2.5 nM. Samples were sequenced as 150bp single-end reads on multiplexed lanes of an Illumina HiSeq3000 in our in-house sequencing facility. Raw reads were deposited at the European Nucleotide Archive (ENA) under the study accession PRJEB80633.

Genome assembly and metabolic potential identification of bacterial strains

The bacterial genomes were assembled from the raw reads using SPAdes[52] (version 3.15.1, parameters: --careful -o -1 -2 -t 30 –cov-cutoff 'auto') and annotated with PROKKA[49] (version 1.14.6, parameters: --addgenes) and eggNOG-Mapper[45] (version 2.1.0-1, eggnog DB version: 5.0.2). The predicted bacterial protein sequences were used as input to detect orthogroups with OrthoFinder and reconstruct the phylogeny with the incorporated RAxML[51]. The bacterial genomes were used to predict the presence or absence of metabolic pathways with gapseq[47] (version 1.2, parameters: find -p all, MetaCyc[55] as the default database) as well as the presence of 67 bacterial phenotypes with the implementation of Traitar in python (traitar3[54], version 3.0.1, parameters: predict) . The bacterial metabolic pathways were grouped (MPGs) based on their presence/absence patterns across the bacterial strains to reduce redundancy. This decreased the number of bacterial metabolic pathways to be assessed from 2902 to 712. The nomenclature of the bacterial strains was determined by applying DIAMOND[43] (version 2.1.4.158, parameters: makedb and blastp) to compare the predicted bacterial protein sequences against the non-redundant version of the NCBI database (August 2023 version).

Association of MPGs with chemotaxis and survival

In the initial study, the chemotaxis index ranges from -1 (repulsion) to 1 (attraction) and the survival index from 0 to 100, the latter reflecting the percentage of worms that were alive on the 5th day post transfer[19]. For our analysis, we binarized the assay data by assigning the values below 40 percent survival and 0.2 chemotaxis as zero and the rest as one. The threshold for the binarization was determined based on the distribution of values. To test the association between each MPG and the nematode phenotypes, we divided the assay data to two vectors, representing the set of values when the nematode was maintained on bacterial strains with and without the MPG and performed a Mann-Whitney-U test, followed by Bonferroni-correction ($P < 0.05$). To associate bacterial virulence factors with survival, we downloaded the 4,236 core proteins from the Virulence Factor DataBase[30] (2024-07-31) and performed pairwise all-against-all BLASTP searches for all bacterial protein sets combined with the VFDB data (e-value < 10^{-5}). The widely used clustering algorithm mcl was used to cluster the data into 1605 orthogroups. These orthogroups were then tested for association with survival using a Wilcoxon-test (FDR-corrected $P < 0.1$).

Supplementation experiments

To check the nematicidal effect of coumarin derivatives *in vivo*, we supplemented 4-Hydroxycoumarin, 7-Methoxycoumarin and xanthotoxin to the nematode. The coumarins were added directly to liquid NGM to a final concentration of 1mM and the solution was subsequently poured to 6 cm plates (11 ml per plate). The plates were seeded with 300 µl of *E. coli* OP50 the next day and left to grow a bacterial lawn for two

days. After that time, 20 young adult hermaphrodites were transferred on each plate and their survival was tracked daily for five days. The surviving nematodes were transferred on new plates containing the coumarins on the third day after the initial transfer to avoid food depletion and misidentification from their offspring. Mortality was determined by lack of response to prodding with the pick. For the control and each coumarin, we performed five replicates.

Dietary experiments

Eggs from *P. pacificus* adults were obtained and spotted on NGM plates seeded with 75 µl bacterial overnight cultures to produce the parental generation (P0) on each diet. 25 gravid hermaphrodites from the P0 were transferred to a new plate seeded with the same bacterial strain and were left to lay eggs for five hours, after which point they were removed. The F1 worms were collected 72 hours after the initial transfer. Collection of the worms required washing the plate with autoclaved $M9^-$ buffer, centrifugation at 500g for a minute and discarding the supernatant in order to remove as much of the collected bacterial lawn as possible. The worm pellets were flash frozen in liquid N_2 and were used for RNA extraction or stored at -80^{o}C for later processing.

RNA sequencing

For total RNA extraction, the frozen worm pellets were treated with Trizole followed by purification with the Zymo RNA Clean & Concentrator 25 Kit according to the manufacturer's instructions. The extracted RNA was quantified and quality assessed with a NanoDrop ND1000 spectrometer (PeqLab, Erlangen, Germany) and a Qubit 2.0 Fluorometer. The samples were shipped to Novogene for library preparation and

messenger RNA (mRNA) sequencing. Libraries were sequenced as 150bp paired-end reads on an Illumina NovoSeq 6000 platform. The obtained raw reads have been deposited in the European Nucleotide Archive under the study accession number PRJEB80633.

Gene coexpression network

The raw reads were aligned to the reference *Pristionchus pacificus* genome with STAR[53] (version 2.7.1a) and quantified with featureCounts[46] from the Subread R package (version 2.0.1) based on the latest annotations. By filtering the count matrix to remove genes with less than 10 reads total, we reduced the number of *P. pacificus* genes to be assessed from 28896 to 24335. The read counts across different conditions and replicates were normalized with the DESeq2[44] counts function (option: normalized = TRUE) and were input in MCL[48] (version 22-282, r=0.7, I=2) to create a gene coexpression network. The network modules were tested for the enrichment of protein domain annotations (PFAM)[33], metabolic pathways (KEGG)[35] and gene expression sets from previous works[25] with Fisher's exact-test followed by Bonferroni correction ($P < 0.05$) in R (version 4.4.0). The network modules were assessed further for their interactions with the bacterial metabolic potential. Specifically, for each MPG, the normalized counts of the genes in each module were separated into two vectors based on whether the presence or absence of the MPG in the bacterial strain in order to represent the expression of a specific gene in a module in the presence or absence of a MPG. The presence and absence expression vectors for each gene were compared using Mann-Whitney-U test as well as to calculate the fold change in expression. To

determine the interaction of a module with a MPG, we retained the genes where the fold change was lower than 0.5 or higher than 2 and performed multiple hypothesis testing for the module genes that fulfilled the condition (Bonferroni < 0.1). The list of genes identified as interacting with the MPGs were tested for module enrichment using Fisher's exact test (Bonferroni-adjusted p-value<0.05) to determine the interaction between the MPGs and the coexpression modules. The bipartite network was implemented using Python (version 3.10.12).

References

1. Zhang J, Holdorf AD and Walhout AJ. C. elegans and its bacterial diet as a model for systems-level understanding of host-microbiota interactions. *Curr Opin Biotechnol* 2017; 46: 74-80. 2017/02/12. DOI: 10.1016/j.copbio.2017.01.008.
2. Watson E, MacNeil LT, Ritter AD, et al. Interspecies systems biology uncovers metabolites affecting C. elegans gene expression and life history traits. *Cell* 2014; 156: 759-770. 2014/02/18. DOI: 10.1016/j.cell.2014.01.047.
3. Dirksen P, Marsh SA, Braker I, et al. The native microbiome of the nematode Caenorhabditis elegans: gateway to a new host-microbiome model. *BMC Biol* 2016; 14: 38. 2016/05/11. DOI: 10.1186/s12915-016-0258-1.
4. Slowinski S, Ramirez I, Narayan V, et al. Interactions with a Complex Microbiota Mediate a Trade-Off between the Host Development Rate and Heat Stress Resistance. *Microorganisms* 2020; 8 2020/11/19. DOI: 10.3390/microorganisms8111781.
5. Iatsenko I, Yim JJ, Schroeder FC, et al. B. subtilis GS67 protects C. elegans from Gram-positive pathogens via fengycin-mediated microbial antagonism. *Curr Biol* 2014; 24: 2720-2727. 2014/12/03. DOI: 10.1016/j.cub.2014.09.055.
6. Howard RJ, Giacomelli M, Lozano-Fernandez J, et al. The Ediacaran origin of Ecdysozoa: integrating fossil and phylogenomic data. *Journal of the Geological Society* 2022; 179. DOI: 10.1144/jgs2021-107.
7. Sanghvi GV, Baskaran P, Röseler W, et al. Life History Responses and Gene Expression Profiles of the Nematode Pristionchus pacificus Cultured on Cryptococcus

Yeasts. *PLoS One* 2016; 11: e0164881. 2016/10/16. DOI: 10.1371/journal.pone.0164881.

8. Bose N, Ogawa A, von Reuss SH, et al. Complex small-molecule architectures regulate phenotypic plasticity in a nematode. *Angew Chem Int Ed Engl* 2012; 51: 12438-12443. 2012/11/20. DOI: 10.1002/anie.201206797.

9. Werner MS, Sieriebriennikov B, Loschko T, et al. Environmental influence on Pristionchus pacificus mouth form through different culture methods. *Sci Rep* 2017; 7: 7207. 2017/08/05. DOI: 10.1038/s41598-017-07455-7.

10. Kieninger MR, Ivers NA, Rödelsperger C, et al. The Nuclear Hormone Receptor NHR-40 Acts Downstream of the Sulfatase EUD-1 as Part of a Developmental Plasticity Switch in Pristionchus. *Curr Biol* 2016; 26: 2174-2179. 2016/07/28. DOI: 10.1016/j.cub.2016.06.018.

11. Ragsdale EJ, Muller MR, Rödelsperger C, et al. A developmental switch coupled to the evolution of plasticity acts through a sulfatase. *Cell* 2013; 155: 922-933. 2013/11/12. DOI: 10.1016/j.cell.2013.09.054.

12. Namdeo S, Moreno E, Rödelsperger C, et al. Two independent sulfation processes regulate mouth-form plasticity in the nematode Pristionchus pacificus. *Development* 2018; 145 2018/07/04. DOI: 10.1242/dev.166272.

13. Sieriebriennikov B, Sun S, Lightfoot JW, et al. Conserved nuclear hormone receptors controlling a novel plastic trait target fast-evolving genes expressed in a single cell. *PLoS Genet* 2020; 16:e1008687. DOI: 10.1371/journal.pgen.1008687.

14. Lenuzzi M, Witte H, Riebesell M, et al. Influence of environmental temperature on mouth-form plasticity in Pristionchus pacificus acts through daf-11-dependent cGMP signaling. *J Exp Zool B Mol Dev Evol* 2023; 340: 214-224. 2021/08/12. DOI: 10.1002/jez.b.23094.

15. Meyer JM, Baskaran P, Quast C, et al. Succession and dynamics of Pristionchus nematodes and their microbiome during decomposition of Oryctes borbonicus on La Reunion Island. *Environ Microbiol* 2017; 19: 1476-1489. 2017/02/16. DOI: 10.1111/1462-2920.13697.

16. Lo WS, Sommer RJ and Han Z. Microbiota succession influences nematode physiology in a beetle microcosm ecosystem. *Nat Commun* 2024; 15: 5137. 2024/06/16. DOI: 10.1038/s41467-024-49513-5.

17. Sinha A, Rae R, Iatsenko I, et al. System wide analysis of the evolution of innate immunity in the nematode model species Caenorhabditis elegans and Pristionchus pacificus. *PLoS One* 2012; 7: e44255. 2012/10/03. DOI: 10.1371/journal.pone.0044255.

18. Lightfoot JW, Chauhan VM, Aylott JW, et al. Comparative transcriptomics of the nematode gut identifies global shifts in feeding mode and pathogen susceptibility. *BMC Res Notes* 2016; 9: 142. 2016/03/06. DOI: 10.1186/s13104-016-1886-9.

19. Akduman N, Rödelsperger C and Sommer RJ. Culture-based analysis of Pristionchus-associated microbiota from beetles and figs for studying

nematode-bacterial interactions. *PLoS One* 2018; 13: e0198018. 2018/06/05. DOI: 10.1371/journal.pone.0198018.
20. Akduman N, Lightfoot JW, Röseler W, et al. Bacterial vitamin B(12) production enhances nematode predatory behavior. *ISME J* 2020; 14: 1494-1507. 2020/03/11. DOI: 10.1038/s41396-020-0626-2.
21. Brady C, Kaur S, Crampton B, et al. Transfer of Erwinia toletana and Erwinia iniecta to a novel genus Winslowiella gen. nov. as Winslowiella toletana comb. nov. and Winslowiella iniecta comb. nov. and description of Winslowiella arboricola sp. nov., isolated from bleeding cankers on broadleaf hosts. *Front Microbiol* 2022; 13: 1063107. 2022/12/06. DOI: 10.3389/fmicb.2022.1063107.
22. Girard L, Lood C, Hofte M, et al. The Ever-Expanding Pseudomonas Genus: Description of 43 New Species and Partition of the Pseudomonas putida Group. *Microorganisms* 2021; 9 2021/08/28. DOI: 10.3390/microorganisms9081766.
23. Kakumanu ML, Marayati BF, Wada-Katsumata A, et al. Sphingobacterium phlebotomi sp. nov., a new member of family Sphingobacteriaceae isolated from sand fly rearing substrate. *Int J Syst Evol Microbiol* 2021; 71 2021/05/07. DOI: 10.1099/ijsem.0.004809.
24. Smits THM, Arend L, Cardew S, et al. Resolving taxonomic confusion: establishing the genus Phytobacter on the list of clinically relevant Enterobacteriaceae. *Eur J Clin Microbiol Infect Dis* 2022; 41: 547-558. 2022/02/17. DOI: 10.1007/s10096-022-04413-8.
25. Athanasouli M, Akduman N, Röseler W, et al. Thousands of Pristionchus pacificus orphan genes were integrated into developmental networks that respond to diverse environmental microbiota. *PLoS Genet* 2023; 19: e1010832. 2023/07/03. DOI: 10.1371/journal.pgen.1010832.
26. Abebew D, Sayedain FS, Bode E, et al. Uncovering Nematicidal Natural Products from Xenorhabdus Bacteria. *J Agric Food Chem* 2022; 70: 498-506. 2022/01/05. DOI: 10.1021/acs.jafc.1c05454.
27. Guo QQ, Du GC, Li YX, et al. Nematotoxic coumarins from Angelica pubescens Maxim. f. biserrata Shan et Yuan roots and their physiological effects on Bursaphelenchus xylophilus. *J Nematol* 2018; 50: 559-568. 2018/01/01. DOI: 10.21307/jofnem-2018-045.
28. Mahajan R, Kaur DJ and Bajaj KL. Nematicidal activity of some phenolic compounds against Meloidogyne incognita *Nematologia Mediterranea* 1992; 20: 217-219.
29. Molnar M, Komar M, Brahmbhatt H, et al. Deep Eutectic Solvents as Convenient Media for Synthesis of Novel Coumarinyl Schiff Bases and Their QSAR Studies. *Molecules* 2017; 22 2017/09/06. DOI: 10.3390/molecules22091482.

30. Liu B, Zheng D, Zhou S, et al. VFDB 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Res* 2022; 50: D912-D917. 2021/12/02. DOI: 10.1093/nar/gkab1107.
31. Clarke-Pearson MF and Brady SF. Paerucumarin, a new metabolite produced by the pvc gene cluster from Pseudomonas aeruginosa. *J Bacteriol* 2008; 190: 6927-6930. 2008/08/12. DOI: 10.1128/JB.00801-08.
32. Sun S, Rödelsperger C and Sommer RJ. Single worm transcriptomics identifies a developmental core network of oscillating genes with deep conservation across nematodes. *Genome Res* 2021; 31: 1590-1601. 2021/07/25. DOI: 10.1101/gr.275303.121.
33. Mistry J, Chuguransky S, Williams L, et al. Pfam: The protein families database in 2021. *Nucleic Acids Res* 2021; 49: D412-D419. 2020/10/31. DOI: 10.1093/nar/gkaa913.
34. Rödelsperger C, Ebbing A, Sharma DR, et al. Spatial Transcriptomics of Nematodes Identifies Sperm Cells as a Source of Genomic Novelty and Rapid Evolution. *Mol Biol Evol* 2021; 38: 229-243. 2020/08/14. DOI: 10.1093/molbev/msaa207.
35. Kanehisa M and Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000; 28: 27-30. 1999/12/11. DOI: 10.1093/nar/28.1.27.
36. Rödelsperger C, Röseler W, Prabh N, et al. Phylotranscriptomics of Pristionchus Nematodes Reveals Parallel Gene Loss in Six Hermaphroditic Lineages. *Curr Biol* 2018; 28: 3123-3127 e3125. 2018/09/25. DOI: 10.1016/j.cub.2018.07.041.
37. Zhang L, Kohler S, Rillo-Bohn R, et al. A compartmentalized signaling network mediates crossover control in meiosis. *Elife* 2018; 7 2018/03/10. DOI: 10.7554/eLife.30789.
38. Hodge F, Bajuszova V and van Oosten-Hawle P. The Intestine as a Lifespan- and Proteostasis-Promoting Signaling Tissue. *Front Aging* 2022; 3: 897741. 2022/07/14. DOI: 10.3389/fragi.2022.897741.
39. O'Donnell MP, Fox BW, Chao PH, et al. A neurotransmitter produced by gut bacteria modulates host sensory behaviour. *Nature* 2020; 583: 415-420. 2020/06/20. DOI: 10.1038/s41586-020-2395-5.
40. Lo WS, Han Z, Witte H, et al. Synergistic interaction of gut microbiota enhances the growth of nematode through neuroendocrine signaling. *Curr Biol* 2022; 32: 2037-2050 e2034. 2022/04/10. DOI: 10.1016/j.cub.2022.03.056.
41. Zimmermann J, Obeng N, Yang W, et al. The functional repertoire contained within the native microbiota of the model nematode Caenorhabditis elegans. *ISME J* 2020; 14: 26-38. 2019/09/06. DOI: 10.1038/s41396-019-0504-y.
42. Manni M, Berkeley MR, Seppey M, et al. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of

Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol* 2021; 38: 4647-4654. 2021/07/29. DOI: 10.1093/molbev/msab199.

43. Buchfink B, Reuter K and Drost HG. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* 2021; 18: 366-368. 2021/04/09. DOI: 10.1038/s41592-021-01101-x.

44. Love MI, Huber W and Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014; 15: 550. 2014/12/18. DOI: 10.1186/s13059-014-0550-8.

45. Cantalapiedra CP, Hernández-Plaza A, Letunic I, et al. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution* 2021; 38: 5825-5829. DOI: 10.1093/molbev/msab293.

46. Liao Y, Smyth GK and Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014; 30: 923-930. 2013/11/15. DOI: 10.1093/bioinformatics/btt656.

47. Zimmermann J, Kaleta C and Waschina S. gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. *Genome Biol* 2021; 22: 81. 2021/03/12. DOI: 10.1186/s13059-021-02295-1.

48. Van Dongen S. Graph Clustering Via a Discrete Uncoupling Process. *SIAM Journal on Matrix Analysis and Applications* 2008; 30: 121-141. DOI: 10.1137/040608635.

49. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014; 30: 2068-2069. 2014/03/20. DOI: 10.1093/bioinformatics/btu153.

50. Gurevich A, Saveliev V, Vyahhi N, et al. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013; 29: 1072-1075. 2013/02/21. DOI: 10.1093/bioinformatics/btt086.

51. Emms DM and Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* 2015; 16: 157. 2015/08/06. DOI: 10.1186/s13059-015-0721-2.

52. Prjibelski A, Antipov D, Meleshko D, et al. Using SPAdes De Novo Assembler. *Curr Protoc Bioinformatics* 2020; 70: e102. 2020/06/20. DOI: 10.1002/cpbi.102.

53. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013; 29: 15-21. 2012/10/30. DOI: 10.1093/bioinformatics/bts635.

54. Weimann A, Mooren K, Frank J, et al. From Genomes to Phenotypes: Traitar, the Microbial Trait Analyzer. *mSystems* 2016; 1 2017/01/10. DOI: 10.1128/mSystems.00101-16.

55. Caspi R, Altman T, Billington R, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* 2014; 42: D459-471. 2013/11/15. DOI: 10.1093/nar/gkt1103.

Figure Legends

Figure 1. Phylogenomics of 84 bacterial genomes. The tree depicts the bacterial phylogeny acquired using RAxML after whole genome sequencing, genome assembly and orthogroup detection. Phenotype prediction with traitar3 revealed that the overwhelming majority of bacterial strains sequenced are anaerobic with the exception of TSA3A and LRB41 who are facultative anaerobic as well as the complete lack of anaerobes. Our collection contains only three Gram positive strains while the rest are gram negative. Genome sizes may vary even within a genus as is the case with *Acinetobacter*, where we observe genomes ranging between 3 and 5 Mb.

Figure 2. Variation in bacterial metabolic potential. (A) Grouping of the predicted MPGs depending on their presence or absence across the bacterial phylogeny revealed the metabolic potential of the individual strains. The majority of MPGs detected were absent from all strains with a core of 45 bacterial pathways predicted always present. The rest of the MPGs exhibit distinct patterns even among the same genus. (B) Pairwise comparisons of the MPGs present in strains of the same genus versus strains of different genera indicated that closely related bacteria share a higher number of MPGs and thus, might exhibit a genus-specific metabolic signature.

Figure 3. Association of nematode survival with the bacterial pathways. (A) An association study between the MPGs and mean survival data narrowed down the MPGs

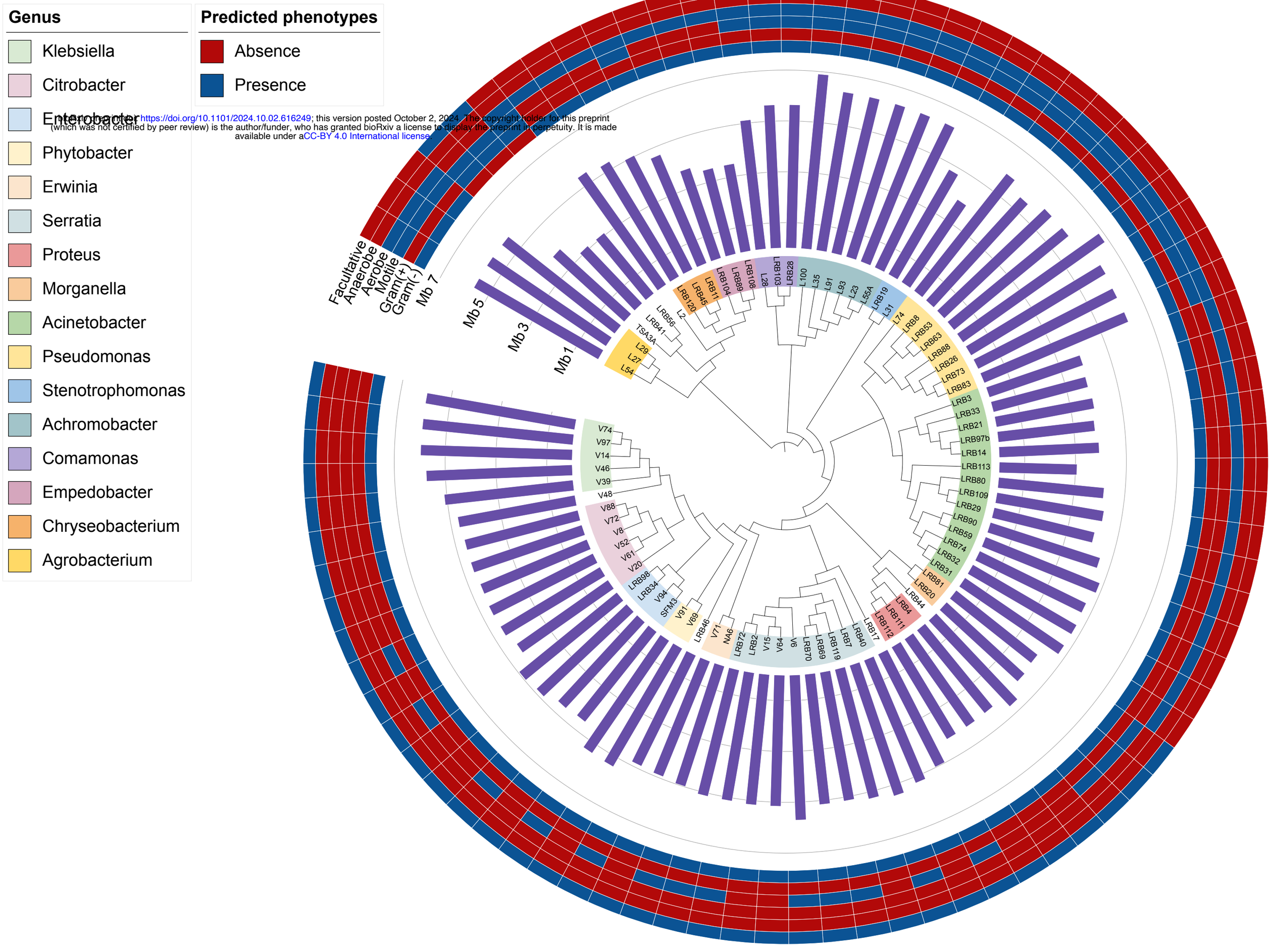
with a significant effect on the trait from hundreds to 41, a subsample of which is shown here. Generally, the presence of an MPG appears to result in lower survival, especially when bacteria produce chitinases and compounds such as paerucumarin and rhabduscin. (B) Supplementation tests with three types of coumarins showed that 7-methoxy coumarin and xanthotoxin affect nematode survival ($P<0.01$, t-test).

Figure 4. Variation in the transcriptomic response of the nematode. (A) Correlation analysis and clustering of the transcriptomes from *P. pacificus* nematodes in response to 38 different bacteria. The overall transcriptomes are highly similar with most pairwise comparisons showing correlation coefficients >0.9. (B) Biological replicates exhibited higher correlation coefficients compared to transcriptomes from different bacterial diets. (C) Transcriptomes of bacteria from the same genus showed significantly higher correlation coefficients compared to dietary samples from different genera.

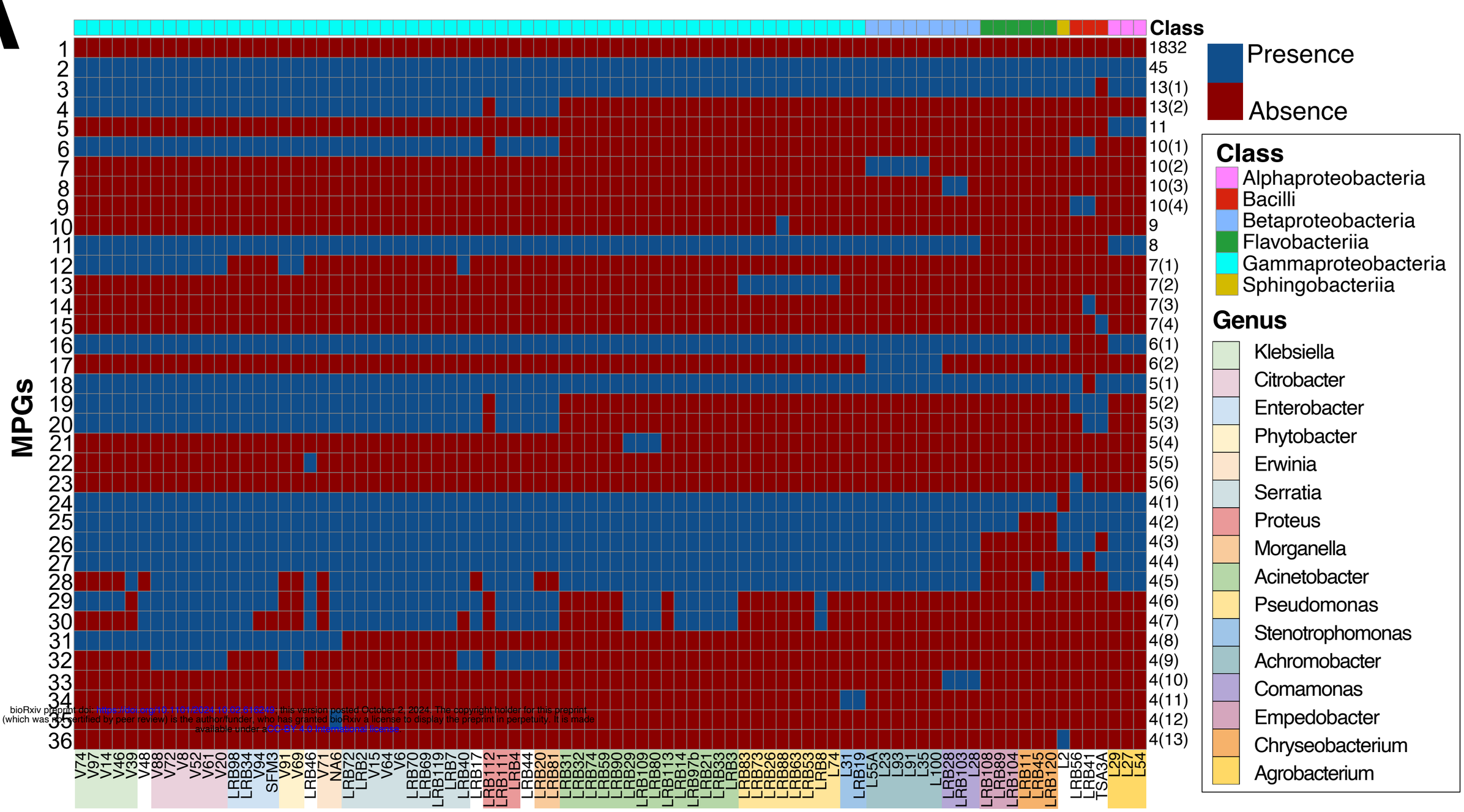
Figure 5. Expression of modules across the different bacterial diets. The z-score normalized expression of the top 60 coexpression modules across the different diets was visualized as a heatmap. The expression of the biological replicates was averaged in order to produce a single expression profile for each of the 38 diets. Modules with more than 50 genes were randomly downsampled to aid the visualization of smaller modules and the comparisons between them. Expression patterns were unique for each module and depended on the bacterial diet, as was the case with modules 2, 7 and 15 on LRB88.

Figure 6. Enrichment analysis of the gene coexpression modules using existing transcriptomic datasets and functional annotation. The networks show selected results of the overrepresentation analysis for specific coexpression modules. See Supplemental Table S6 for the full data set. Based on the strongest enrichment patterns, we connected 48 of the processed 60 modules with biological processes and protein domains, allowing the functional labeling of the majority of the modules (Supplemental Table S7).

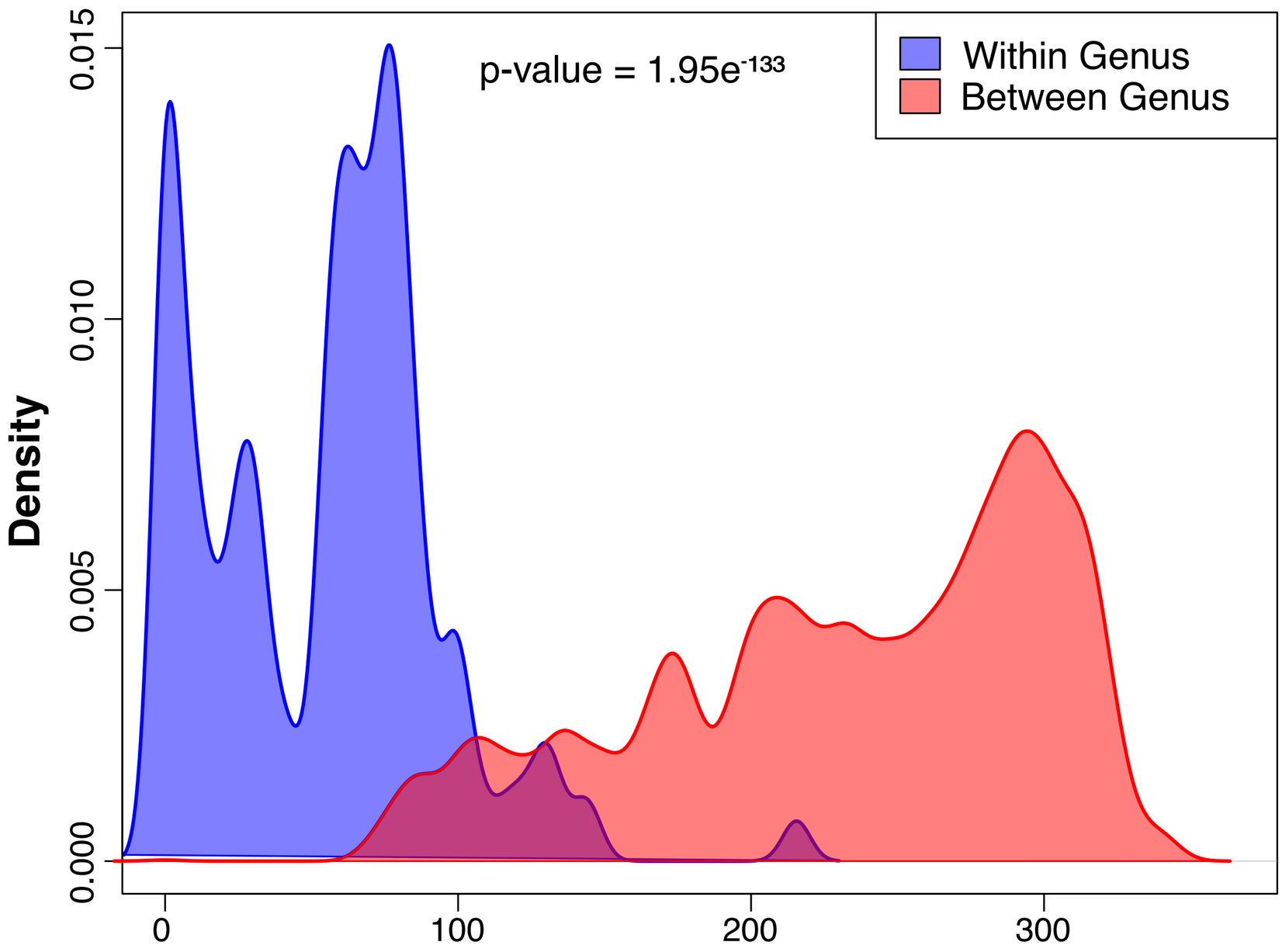
Figure 7. Interactions between bacterial pathways and coexpression modules. (A) A bipartite network was built for MPGs and coexpression modules by defining interactions based on the significance of the gene set overlap (lowest Bonferroni-adjusted p-value<0.05). The plot shows all interactions between selected MPGs that were defined based on the 50 most significant interactions and the top 60 modules. The p-value scale depicts the negative logarithm of the corrected p-values.



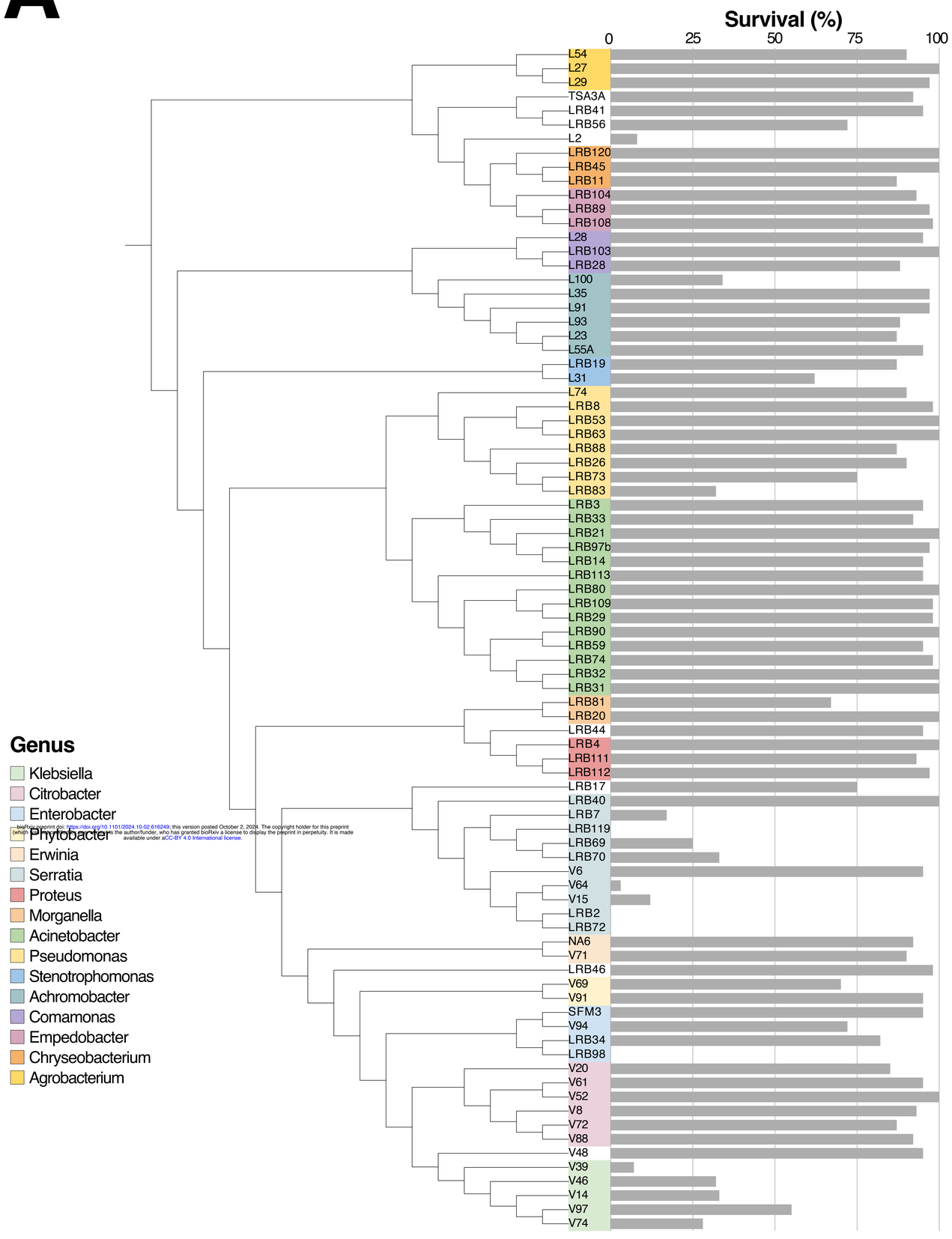
A



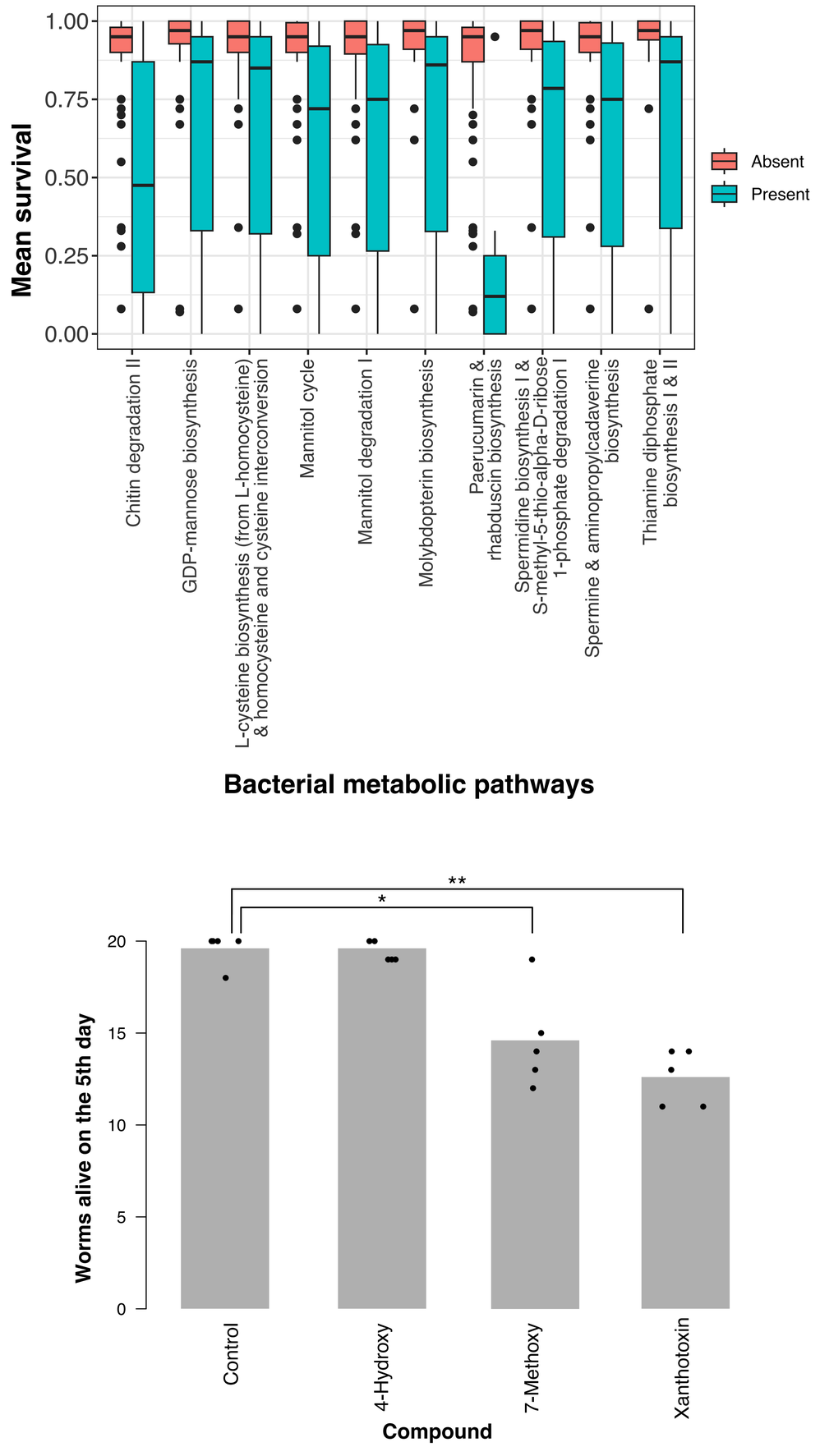
B



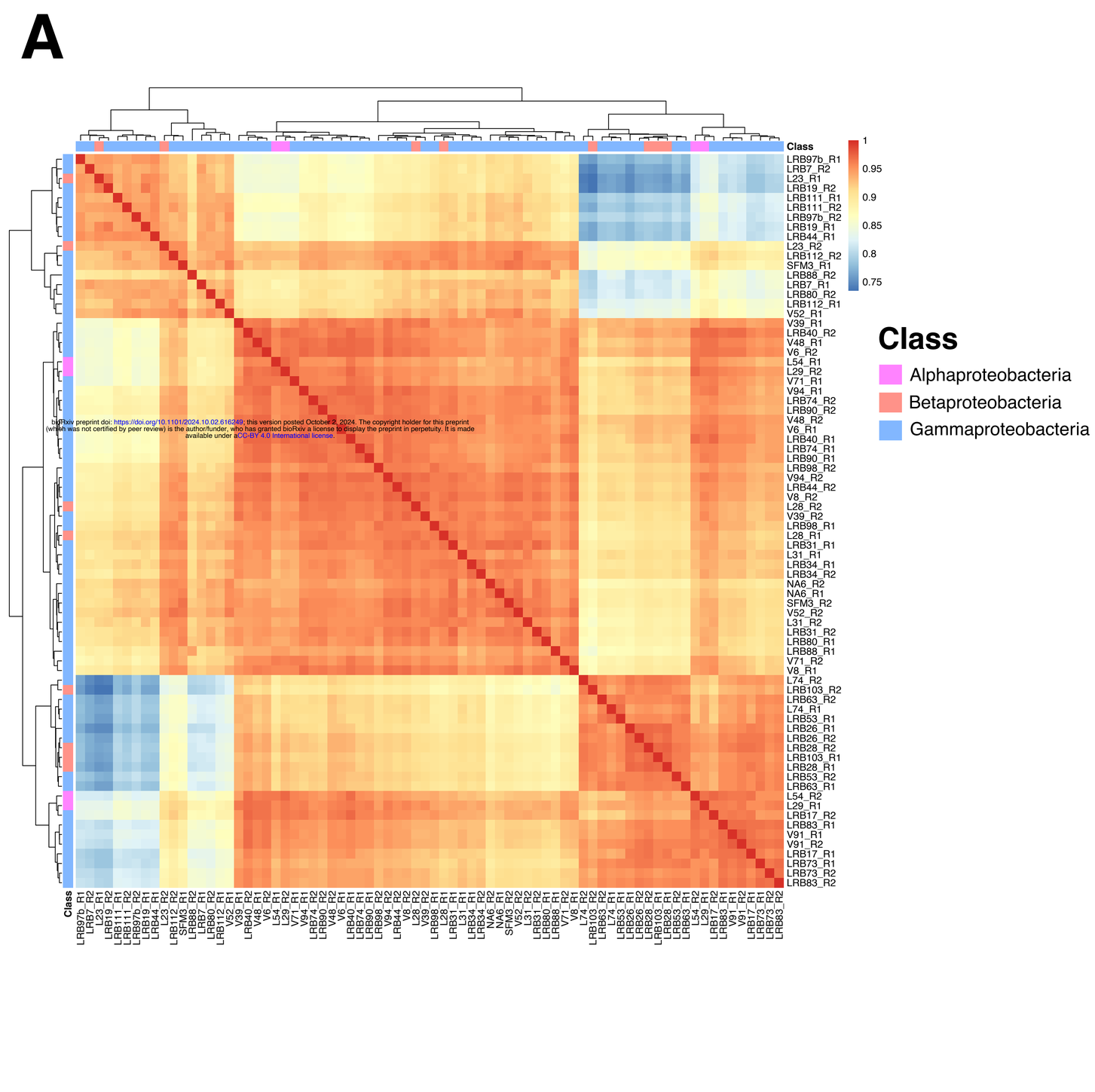
A



B



C


A
Class
LRB97b_R1
LRB7_R2
L23_R1
LRB19_R2
LRB111_R1
LRB111_R2
LRB97b_R2
LRB19_R1
LRB44_R1
L23_R2
LRB112_R2
SFM3_R1
LRB88_R2
LRB7_R1
LRB80_R2
LRB112_R1
V52_R1
V39_R1
LRB40_R2
V48_R1
V6_R2
L54_R1
L29_R2
V71_R1
V94_R1
LRB74_R2
LRB90_R2
V48_R2
V6_R1
LRB40_R1
LRB74_R1
LRB90_R1
LRB98_R2
V94_R2
LRB44_R2
V8_R2
L28_R2
V39_R2
LRB98_R1
L28_R1
LRB31_R1
L31_R1
LRB34_R1
LRB34_R2
NA6_R2
NA6_R1
SFM3_R2
V52_R2
L31_R2
LRB31_R2
LRB80_R1
LRB88_R1
V71_R2
V8_R1
L74_R2
LRB103_R2
LRB63_R2
L74_R1
LRB53_R1
LRB26_R1
LRB26_R2
LRB28_R2
LRB103_R1
LRB28_R1
LRB53_R2
LRB63_R1
L54_R2
L29_R1
LRB17_R2
LRB83_R1
V91_R1
V91_R2
LRB17_R1
LRB73_R1
LRB73_R2
LRB83_R2
1
0.95
0.9
0.85
0.8
0.75
Class
Alphaproteobacteria
Betaproteobacteria
Gammaproteobacteria
bioRxiv preprint doi: https://doi.org/10.1101/2024.10.02.616346; this version posted October 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.
B
Replicates
Diets
p-value = 2.46e-09
Density
0
10
20
30
40
50
0.6
0.7
0.8
0.9
1.0
C
Within genus
Between genera
p-value = 3.5e-10
Density
0
5
10
15
20
25
0.6
0.7
0.8
0.9
1.0

Bacterial diets

Genes

Module

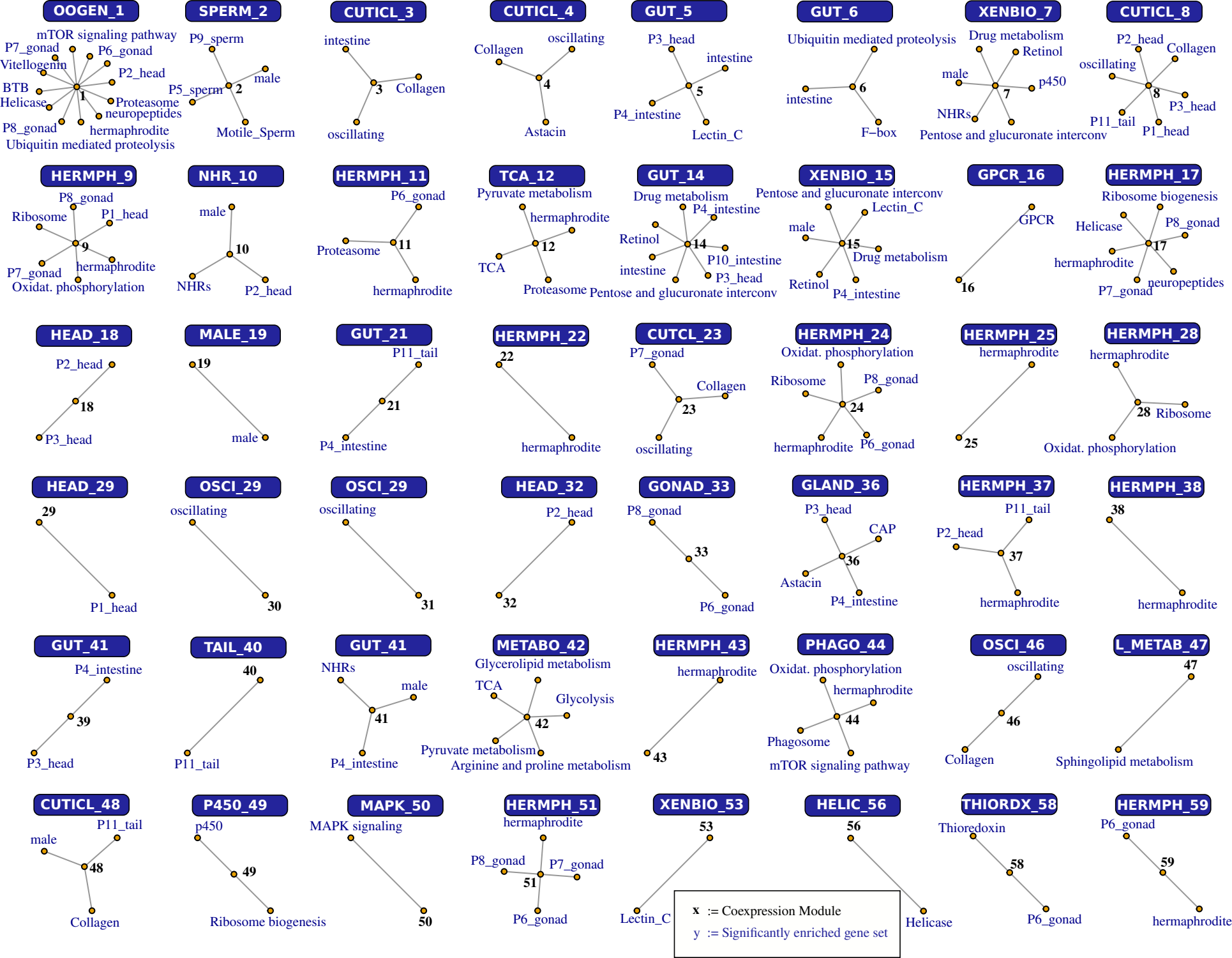
V39
V48
V8
V52
LRB98
LRB34
V94
SFM3
V91
V71
NA6
V6
LRB7
LRB40
LRB17
LRB112
LRB111
LRB44
LRB31
LRB74
LRB90
LRB80
LRB97b
LRB83
LRB73
LRB26
LRB88
LRB63
LRB53
L74
L31
LRB19
L23
LRB28
LRB103
L28
L29
L54

Z-score

6
4
2
0
−2

Genus

- Klebsiella
- Citrobacter
- Enterobacter
- Phytobacter
- Erwinia
- Serratia
- Proteus
- Acinetobacter
- Pseudomonas
- Stenotrophomonas
- Achromobacter
- Comamonas
- Agrobacterium


OOGEN_1
mTOR signaling pathway
P7_gonad
P6_gonad
Vitellogenin
P2_head
BTB
Helicase
1
Proteasome
neuropeptides
P8_gonad
hermaphrodite
Ubiquitin mediated proteolysis
SPERM_2
P9_sperm
male
P5_sperm
2
Motile_Sperm
CUTICL_3
intestine
3
Collagen
oscillating
CUTICL_4
Collagen
oscillating
4
Astacin
GUT_5
P3_head
intestine
5
P4_intestine
Lectin_C
GUT_6
Ubiquitin mediated proteolysis
6
intestine
F−box
XENBIO_7
Drug metabolism
Retinol
male
p450
7
NHRs
Pentose and glucuronate interconv
CUTICL_8
P2_head
Collagen
oscillating
8
P3_head
P11_tail
P1_head
HERMPH_9
P8_gonad
Ribosome
P1_head
9
P7_gonad
hermaphrodite
Oxidat. phosphorylation
NHR_10
male
10
NHRs
P2_head
HERMPH_11
P6_gonad
Proteasome
11
hermaphrodite
TCA_12
Pyruvate metabolism
hermaphrodite
12
TCA
Proteasome
GUT_14
Drug metabolism
P4_intestine
Retinol
14
P10_intestine
intestine
P3_head
Pentose and glucuronate interconv
XENBIO_15
Pentose and glucuronate interconv
Lectin_C
male
15
Drug metabolism
Retinol
P4_intestine
GPCR_16
GPCR
16
HERMPH_17
Ribosome biogenesis
Helicase
P8_gonad
17
hermaphrodite
P7_gonad
neuropeptides
HEAD_18
P2_head
18
P3_head
MALE_19
19
male
GUT_21
P11_tail
21
P4_intestine
HERMPH_22
22
hermaphrodite
CUTCL_23
P7_gonad
Collagen
23
oscillating
HERMPH_24
Oxidat. phosphorylation
Ribosome
P8_gonad
24
hermaphrodite
P6_gonad
HERMPH_25
hermaphrodite
25
HERMPH_28
hermaphrodite
28
Ribosome
Oxidat. phosphorylation
HEAD_29
29
P1_head
OSCI_29
oscillating
30
OSCI_29
oscillating
31
HEAD_32
P2_head
32
GONAD_33
P8_gonad
33
P6_gonad
GLAND_36
P3_head
CAP
36
Astacin
P4_intestine
HERMPH_37
P11_tail
P2_head
37
hermaphrodite
HERMPH_38
38
hermaphrodite
GUT_41
P4_intestine
39
P3_head
TAIL_40
40
P11_tail
GUT_41
NHRs
male
41
P4_intestine
METABO_42
Glycerolipid metabolism
TCA
Glycolysis
42
Pyruvate metabolism
Arginine and proline metabolism
HERMPH_43
hermaphrodite
43
PHAGO_44
Oxidat. phosphorylation
hermaphrodite
44
Phagosome
mTOR signaling pathway
OSCI_46
oscillating
46
Collagen
L_METAB_47
47
Sphingolipid metabolism
CUTICL_48
P11_tail
male
48
Collagen
P450_49
p450
49
Ribosome biogenesis
MAPK_50
MAPK signaling
50
HERMPH_51
hermaphrodite
P8_gonad
P7_gonad
51
P6_gonad
XENBIO_53
53
Lectin_C
HELIC_56
56
Helicase
x := Coexpression Module
y := Significantly enriched gene set
THIORDX_58
Thioredoxin
58
P6_gonad
HERMPH_59
P6_gonad
59
hermaphrodite

MPGs

L-tryptophan degradation II
Andamide degradation
Ethanol degradation IV
Ethanol degradation III, salidroside biosynthesis, L-tyrosine degradation to 4-hydroxyphenylactate
Mycothiol oxidation
L-lysine degradation I
NADH to cytochrome bo oxidase and bd oxidase electron transfer II
D-lactate to cytochrome bo oxidase electron transfer, methylglyoxal degradation I
Ethanol degradation I
beta-Alanine degradation III
L-methionine biosynthesis IV
Adenosylcobinamide-GDP biosynthesis from cobyrinate a,c-diamide
(5Z)-dodecenoate biosynthesis II
Tetracycline resistance
L-malate degradation
L-lyxonate degradation
Linuron degradation
UDP-N-acetyl-alpha-D-fucosamine biosynthesis
Alkylnitronates degradation
Group 34
Androstenedione degradation I (aerobic)
2-aminoethylphosphonate degradation II
Oxalate degradation III
Guanine and guanosine salvage
D-glucarate degradation II
Shikimate degradation II
Yersiniabactin biosynthesis
Phosphatidylethanolamine biosynthesis III, phosphatidylserine biosynthesis II, cob(II)yrinate diamide biosynthesis II
Alginate biosynthesis I (algal)
Dimethylsulfoniopropanoate degradation I
di-trans,polycis-undecaprenyl phosphate biosynthesis
Group 13
L-tryptophan degradation to 2-amino-3-carboxymuconate semialdehyde
L-lysine degradation V
2-hydroxypenta-2,4-dienoate degradation
L-arginine biosynthesis III
Uracil degradation I, thymine degradation
Glycine betaine degradation I
Catechol degradation to 2-hydroxypentadienoate I and II, 3-chlorocatechol degradation III
Atrazine degradation I (aerobic)
Aerobic respiration I and II, L-ornithine degradation I
Group 10
Xylitol degradation
L-isoleucine biosynthesis I
L-ornithine biosynthesis I
GDP-D-rhamnose biosynthesis
Gallate degradation II
D-galactonate degradation
L-arginine degradation IV, putrescine biosynthesis II
L-arginine degradation VI

1 2 3 4 5 7 8 10 14 15 19 20 21 22 23 30 31 36 39 41 42 46 47 49 53 55

Modules

-log10(Bonferroni-adjusted p<0.05)
300
200
100

Gene ratio
0.2
0.4
0.6
0.8

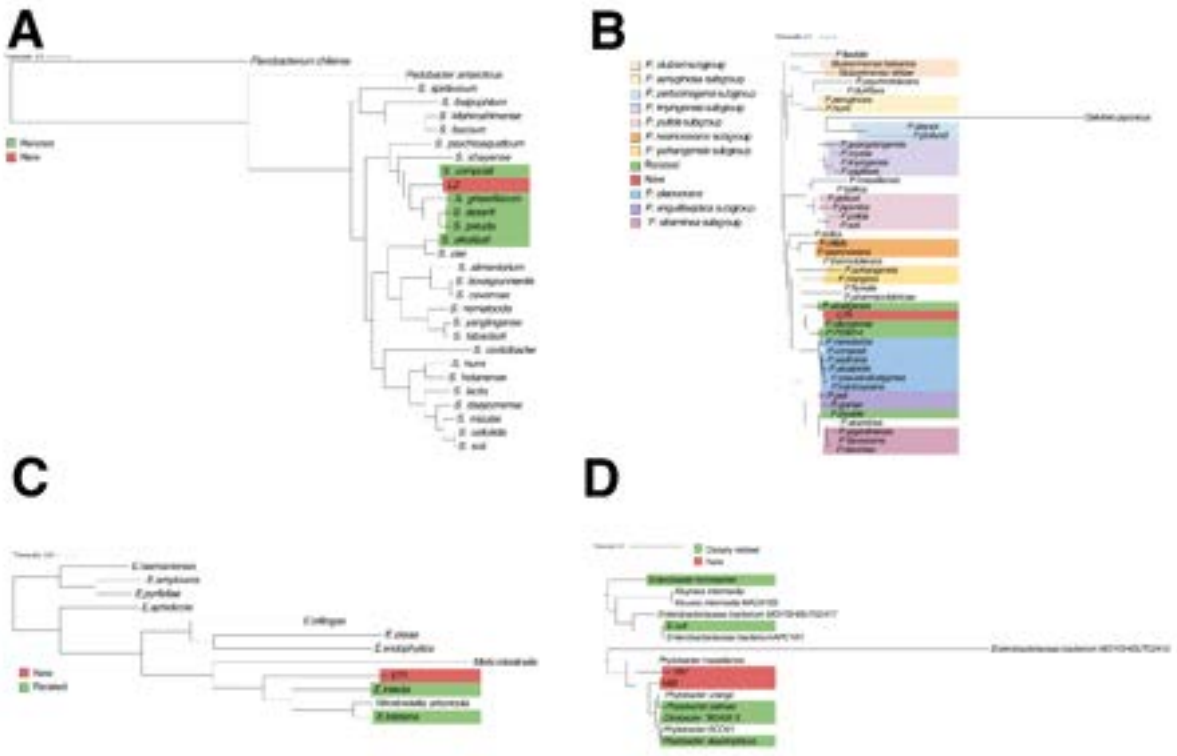
Supplementary Material for

Interspecies systems biology links bacterial metabolic pathways to nematode gene expression, behavior, and survival

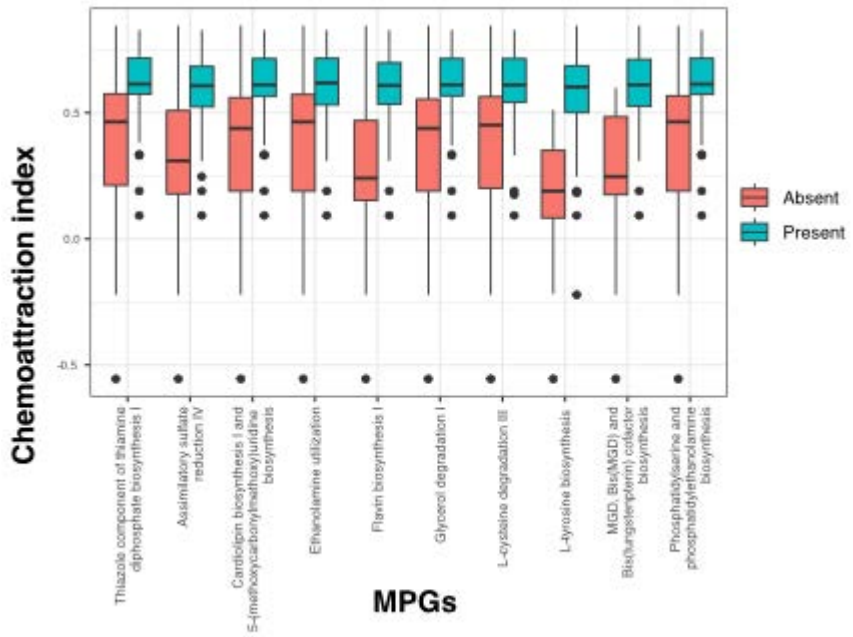
Marina Athanasouli[1], Tobias Loschko[1], Christian Rödelsperger[1,*]

[1] Department for Integrative Evolutionary Biology, Max Planck Institute for Biology, Max-Planck-Ring 9, 72076 Tübingen, Germany

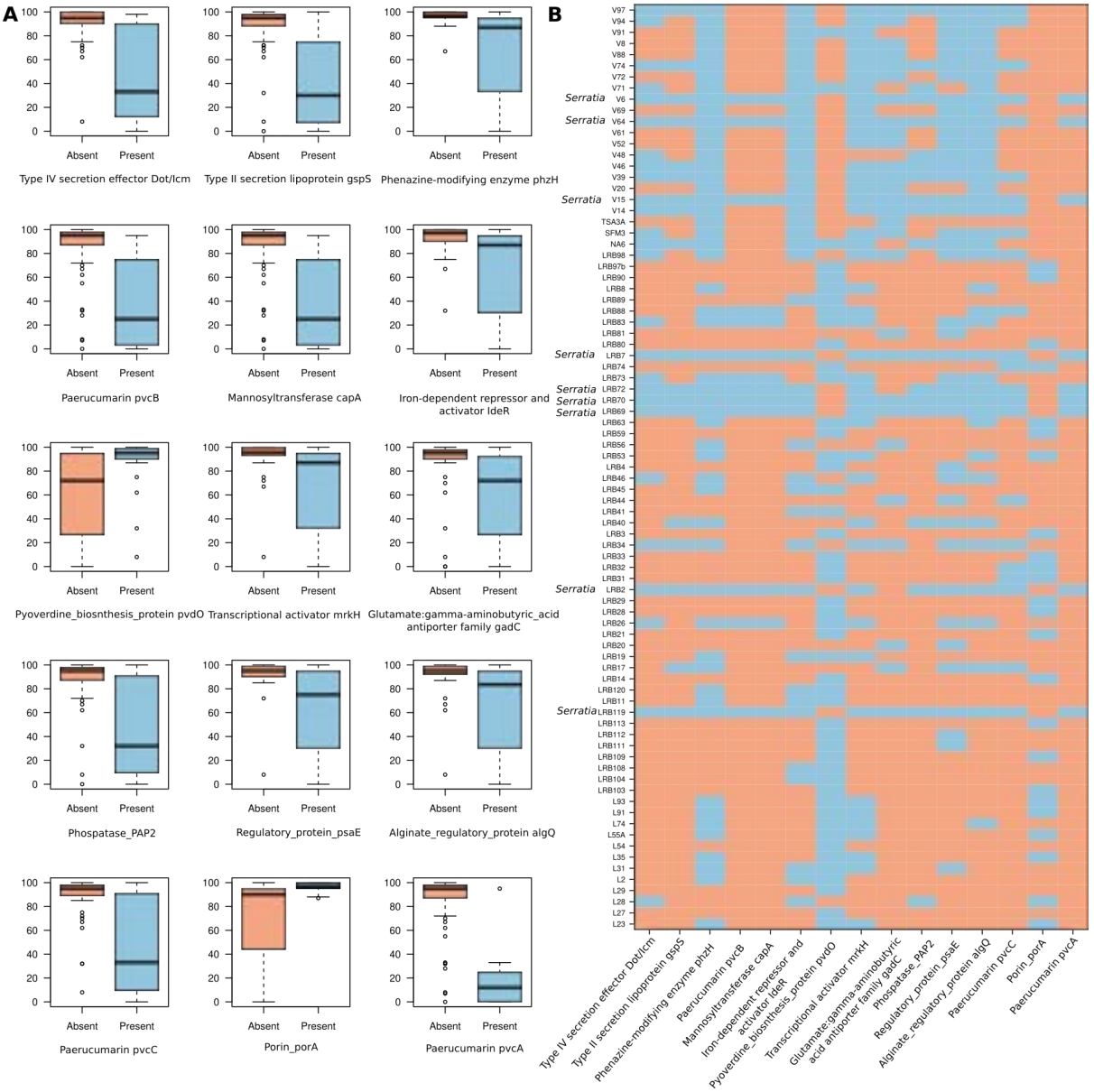
[*] Corresponding author: christian.roedelsperger@tuebingen.mpg.de



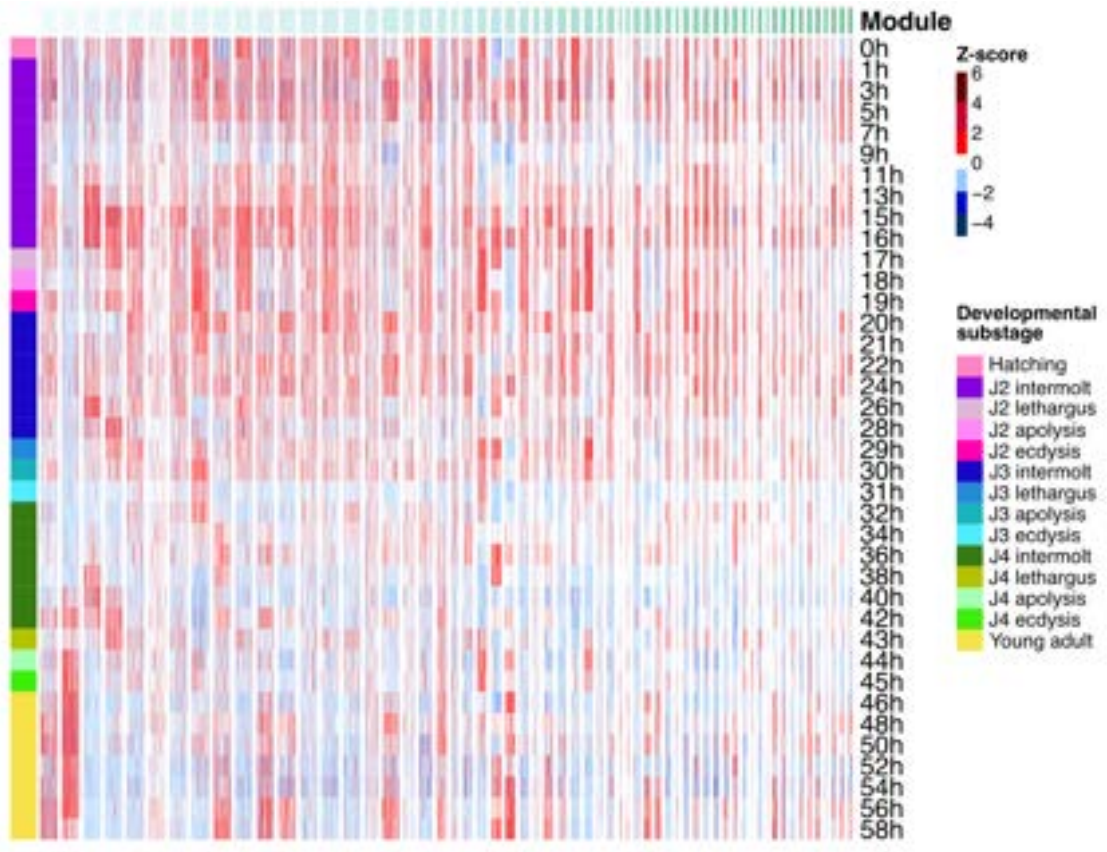
Supplemental Figure S1. Reconstruction of bacterial phylogenies to characterize novel strains. Potentially novel bacterial strains in our collection were integrated into previously published phylogenies by reconstructing species trees that include the candidate genome together with their best hits in the non-redundant version of NCBI. (A) The phylogeny for *Sphingobacterium* L2 (B) The phylogeny for *Pseudomonas* L74. (C) The phylogeny for *Erwinia* V71. (D) The phylogeny for *Phytobacter* V69 and V91.



Supplemental Figure S2. Association of MPGs with chemoattraction in the nematode. Association analysis with the MPGs against the chemoattraction data from Akduman et al (2018) identified 25 MPGs with a predicted significant impact on chemotaxis behavior, a subset of which is depicted here.



Supplemental Figure S3. Association of bacterial virulence factors with nematode survival. A) The core protein set of the Virulence Factor DataBase (VFDB) analysis was clustered together with the bacterial protein sets into orthogroups which were then used to test for associations with survival data. The boxplots show the most significantly associated orthogroups, labeled with a representative VFDB member. B) The heatmap shows the presence/absence pattern of the candidate orthogroups across the bacterial genomes.



Supplemental Figure S4. Visualization of the co-expression modules with developmental time course data from Sun et al (2021). The top modules of the co-expression network were visualized with developmental time-course data. This revealed developmental signatures in various modules.