INVESTIGATING THE EVOLUTION AND ECOLOGY OF OBSCURE BACTERIAL SYMBIOSES FOUND IN INVERTEBRATES, CILIATES AND ALGAE.

Thesis submitted to the University of Liverpool for the degree of Doctor of Philosophy

Helen Rebecca Davison

January 2023

Acknowledgements

I am deeply indebted to Greg Hurst for his support, advice, and encouragement throughout this PhD and beyond. I would not have half the confidence in my own abilities without him pushing me out of my comfort zone on a regular basis.

This project would not have been possible without the guidance or patience of Stefanos Siozios and his seemingly bottomless well of knowledge of bioinformatics. Thank you for being my somewhat involuntary debug duck.

My gratitude to Craig MacAdam for allowing me to experience working for a conservation charity and giving me the opportunity to meet more mad bug people outside of academia. Your enthusiasm is infectious.

David Atkinson and Stephen Cornell for sanity checking my ideas early on in my projects when it came to experimental design and statistics.

Jordan Jones, Ewa Chrostek, Emily Hornett, and the rest of the postgrad offices for being a wonderful and helpful community of people to be a part of.

The Greyhawk crew for providing an escape into many other worlds and sharing a delightfully cursed sense of humour. May you always find the real fake doors to lead you forward and accidentally make pets of cosmic horrors. Fuck you all and I love you.

My gratitude to Bree Foster for suffering through a PhD before me and inadvertently showing me how not to tackle one. I'm glad you've found a job that is basically a PhD student rehabilitation scheme and I wish you luck. Our weekly anime nights have kept me sane in the last few years.

To my dogs, Harry and Rosey, for keeping my feet warm as I write and reminding me to take breaks.

And last but not least, my parents for all the tea and biscuits, and being suitable impressed even by small achievements that I would not even think to celebrate.

Table of Contents

Table of Figures

Table of Tables

Abstract

Bacterial symbioses form a fundamental part of the biology of most eukaryotic lifeforms, influencing their evolution, ecology, and behaviour. Their importance has been increasingly recognised in the last few decades, aided by advances in genomic and bioinformatic methods and analyses. As with most emerging fields, most of our knowledge comes from selected 'model' case studies, leaving the breadth of possible symbioses poorly explored. In this thesis I utilise a combination of bioinformatics, genomics, fieldwork, and microscopy to explore obscure symbioses across invertebrates, algae, and ciliates. First, I broaden the scope of available genomic and metabolic data available for rarer symbionts in invertebrates, a group that are often studied for their heritable symbionts. I argue that the group previously called Torix *Rickettsia* is distinct and diverse and should be regarded as a genus with at least three species, which I name '*Candidatus* Tisiphia'. I also report the first genome for the genus '*Candidatus* Megaira', a widely recorded but poorly understood symbiont of microeukaryotes. I then explored the distribution of various symbiotic bacteria found in ciliates and algae, two host groups that are known to have strong links to symbiotic bacteria and the origins of symbioses but are rarely examined. I show the genus '*Ca*. Megaira' appears as a deeply diverse, multi-species group of symbionts that is deserving of family status. I find '*Ca*. Megaira' in both algae and ciliate species and infer that they have the potential to form protective symbioses. Likewise, I find diverse Parachlamydiales in algae and ciliates and propose three new species groups to aid taxonomic clarification of these bacteria. I provide potential microeukaryotic hosts for a group that are often divorced from host species when described and propose the possibility of nutritional and protective symbioses. Lastly, I develop a potential host-symbiont study system for future functional studies. Here, I demonstrate the existence of a likely heritable *Ca.* Tisiphia symbiont in the mosquito *Anopheles plumbeus*. It represents a potentially important system for onward application in manipulation of anopheline vector populations, which are currently restricted to a single symbiont. Finally, I synthesize these findings and argue future research should focus on the phenotypes of real-world symbioses discovered within this research.

Chapter 1. Introduction

1.1 The importance of symbiotic bacteria: What do we know and what are the gaps?

Interest in the role of microbial communities that live on and in organisms is continually growing, with special relevance to health and medicine and, increasingly, in studies of animal behaviour and ecology. This growing fascination in microbial symbionts has generated pop-sci books like "I am multitudes" by Ed Yong (Yong, 2016). However, as with any new field, our knowledge of endosymbioses is limited to a few well documented examples. In this chapter and the rest of this thesis, I will focus on the diversity, ecology and evolution of symbionts of invertebrates, algae, and ciliates.

Box 1.1 Terminology

Symbiont – an organism that lives with another organism.

Endosymbiont – an organism that lives inside a host organism

Parasite – an organism that leaches off the host without causing out right death

Obligate (endo)symbiont – the organism cannot live as anything other than a symbiont, OR the host cannot live without a particular symbiont.

Facultative (endo)symbiont – the host or symbiont are not wholly dependent on the other for survival.

Mutualistic (endo)symbiont – both parties benefit from the relationship

Commensal – host and symbiont live together without strongly impacting the other

The word symbiont is often mistakenly used to describe only beneficial associates (Box 1.1). Symbioses are energetically costly to maintain and may be quickly lost in the absence of manipulation by one or both parties, or without external pressure. Some symbionts produce toxins to which the host becomes irreversibly addicted and unable to reproduce without (Zchori-Fein, Borad and Harari, 2006; Kremer *et al.*, 2009). Others kill male progeny so only infected individuals are viable (Montenegro *et al.*, 2005; Duplouy and O'Neill, 2010; Hayashi *et al.*, 2016). The line between beneficial symbiosis and parasitism is thin and dynamic, and parasitic bacteria are, in many conceptualizations, additionally referred to as symbionts. This is because, in a literal definition, Symbiosis simply describes the fact that two organisms live in close

association. Within symbiosis there is a sliding scale from beneficial to costly (the parasitism-mutualism continuum) and from obligate to facultative (Figure 1.1 and Box 1.2).

Symbiotic bacteria are present across all kingdoms of eukaryotic life. They are important facets of whole organism biology and can shape how their hosts interact with the environment and how they evolve. In many cases these interactions are transient, they change quickly with the environment and between generations. For instance, mammalian gut microbiota aid nutrient processing but species composition changes with environment, diet and stress (Turnbaugh *et al.*, 2009; Kau *et al.*, 2011; Hanski *et al.*, 2012). However, symbioses with obligate intracellular bacteria like *Wolbachia* tend to be more persistent, and in many cases, they become intrinsically linked with the host's reproduction and natural history (Schardl, Leuchtmann and Spiering, 2004; Moran, McCutcheon and Nakabachi, 2008; Hurst, 2017).

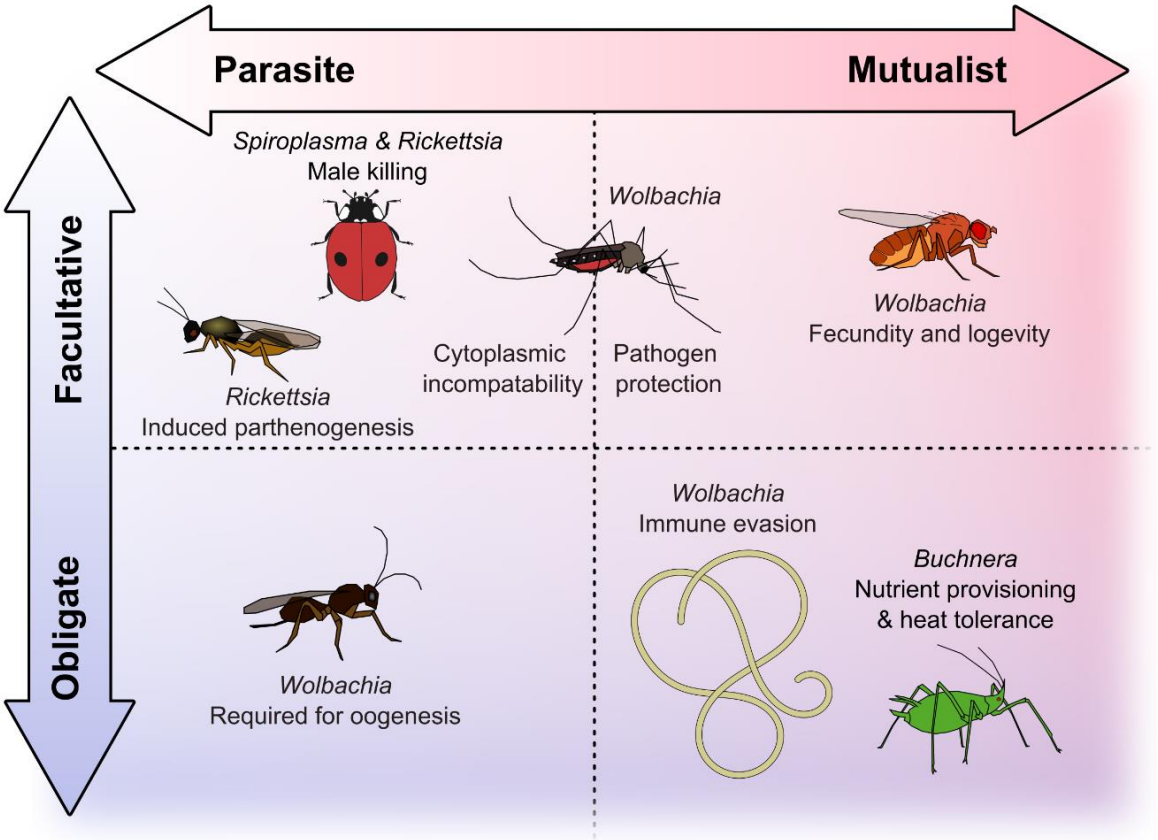


Figure 1.1. The sliding scale of symbiotic interactions. Examples of effects and organisms include: male killing in *Adalia bipunctata* (Hurst et al., 1999), induced parthenogenesis in *Pnigalio soemius* wasps (Giorgini et al., 2010), oogenesis in *Asobara tabida* wasps (Kremer et al., 2009), cytoplasmic incompatibility and pathogen protection in Culicine mosquitoes (Zélé et al., 2012), effects on fecundity and longevity in *Drosophila melanogaster* flies (Fry, Palmer and Rand, 2004), nutrient provisioning and heat tolerance in *Aphidicola*, and immune evasion in the filarial parasite *Onchocerca ocheng* (Hansen et al., 2011). Diagram based on Figure 1 in (Gill, Darby and Makepeace, 2014).

1.1.1 Evolution and ecology of bacterial symbionts

The existence of endosymbiotic bacteria has been recognised since at least 1896 (Dangeard, 1896), but their effects and evolution have only been extensively studied from the 1960s since Buchner's works on symbioses in animals and plants (Buchner, 1965; Sapp, 2002). Even then, the full breadth and prevalence of endosymbiotic bacteria has not been fully appreciated until the last 30 years, which coincides with advances first in molecular genetics (PCR based detection methods and DNA sequence-based taxonomy) and then genomic technology (Next generation sequencing). Many of these symbionts live in the cells of their hosts and are closely linked to their host's evolution and can be inherited between generations (Moran, McCutcheon and Nakabachi, 2008). We know that symbiotic bacteria interact with their hosts in 'give and take' relationships that are subject to evolution (Oliver *et al.*, 2003; Xie, Vilchez and Mateos, 2010; Xie *et al.*, 2011).

We have observed that symbionts can be considered like genes; as they are passed from one generation to the next, and their dynamics is indeed commonly modelled in population genetic frameworks (Hurst, 2017). They also allow horizontal transfer of traits between eukaryotes (Oliver *et al.*, 2010), which might enable spontaneous resistance to natural enemies to occur or altering vector competency for diseases like Dengue, Mayar, and Leaf Curl Virus (Walker *et al.*, 2011; Kliot *et al.*, 2014; Pereira *et al.*, 2018). Symbioses can impact their host in several ways that can strongly impact their evolution (Box

Box 1.2 – The broad types of symbiosis

Nutritional symbioses – host and symbiont exchange nutrients that one or the other cannot produce alone.

Protective symbioses – The host benefits from protection against natural enemies or environmental stress. This covers parasitoids, pathogenic microbes, viruses, thermal stress, and abiotic toxins like insecticides or heavy metals.

Reproductive manipulation – The reproductive cycle is altered in a way that is more beneficial to one party. The male line becomes an evolutionary dead end for the bacteria if it is only inherited through female eggs, so it can lead to female biased population and may alter the reproductive system of the host. Changes in reproductive ability can also transform reproductive behaviour.

1.2,Table 1.1), but the broad ecology of most endosymbiotic bacteria in the environment is unknown, and most research has focussed on terrestrial systems. Some have theorised that symbiosis first occurred in aquatic habitats (Schrallhammer *et al.*, 2013), and if we go as far back as the inception of the eukaryotic cell as a fusion between Archaea and Alphaproteobacteria, this is very likely to be true. It is far easier to obtain a microbial symbiont when submerged in water because the organism is surrounded by microbe rich broth. For instance, Bobtail squid take advantage of this by gathering specific *Vibrio fischerii* bacteria from sea water which provide bioluminescence for counter illumination (Boettcher, Ruby and McFall-Ngai, 1996). Further, for microeukaryotes, an intracellular symbiont is highly likely also be a heritable symbiont, as cytoplasm is shared between daughter cells on asexual fission.

The method of inheritance and acquisition greatly influences how a symbiont evolves with their host. Because endosymbionts are often vertically transmitted, they tend not to be outright pathogenic, even those related to typically pathogenic species, like *Burkholderia*, *Rickettsia* and *Parachlamydia* (Weinert *et al.*, 2009; Flórez *et al.*, 2017; Arthofer *et al.*, 2022). However, if a symbiont is capable of biparental transmission through both eggs and sperm, or is picked up from the environment, it may retain pathogenic characteristics (Horn, 2008). Vertical transmission has two evolutionary impacts. First, the bottleneck on host reproduction increases within host relatedness and thus minimizes conflict amongst bacteria within a host. Second, the dependence on the host for transmission selects for the microbe to nurture the host for onward transmission. For example, the D strain of *Amoeba proteus* became infected with a virulent bacterial infection, later identified as '*Candidatus* Legionella jeonii', which negatively infected growth and reproduction of the host (Jeon and Lorch, 1967). Over time this infection became less virulent and within a few years the amoeba (now known as the xD strain) had become entirely dependent on the bacteria to be viable (Jeon and Ahn, 1978; Choi *et al.*, 1997). Alternatively, vectored pathogens might become symbionts of the vector species. This pattern has been observed across insect species with *Arsenophonus*, where the bacteria are either vertically transmitted (Bressan, 2014) are seasonally dependant on their host for horizontal transmission (Drew *et al.*, 2021).

Another route for the origin of a new symbiosis is sweeping rapidly into populations from a horizontal transmission 'host shift' event. In 6 years between 2000 and 2006, a Belli group *Rickettsia* swept through North American *Bemisia tabaci* populations (Himler *et al.*, 2011). It provided several potential benefits to the white fly including increased growth rates, pathogen resistance and fecundity (Himler *et al.*, 2011; Hendry, Hunter and Baltrus, 2014). However, as quickly as it swept into the population since 2012, it has begun to decline (Bockoven *et al.*, 2020). Similar waves of infection by endosymbionts have been commonly observed in other insect species, indicating that associations are spatially and temporally dynamic (Drew *et al.*, 2021).

Although symbionts tend to be non-pathogenic, this does not mean that they are not in competition with their host or will not harm their host. The effects of partnerships between bacteria and invertebrates are context dependant. In many cases there is a metabolic cost to maintaining populations of bacteria in the body such that the host only benefits when other aspects of infection phenotype outweigh the costs (Hrček, McLean and Godfray, 2016). For example, *Spiroplasma* provides protection in the presence of parasitoid wasps but reduces fecundity so, in the absence of external pressure from wasps, it becomes a negative trait that is quickly lost from the population (Watts *et al.*, 2009; Xie *et al.*, 2011, 2014). In other cases, environmental conditions can affect only the symbiont with knock on effects to the host. Some *Buchnera* (an obligate symbiont of aphids) carry a point mutation which improves reproduction rates at low temperature but prevent the production of heat shock proteins (Dunbar *et al.*, 2007). These heat sensitive *Buchnera* die off at high temperatures, ablating also their host's ability to reproduce.

It has been argued that heritable symbionts are to eukaryotes what plasmids are to bacteria. Like plasmids, symbionts are an important component of adaptive variation (Gogarten and Townsend, 2005; Moran, 2007; Hurst, 2017), can transfer to novel host species, and can transfer segments of new genetic material to their host's genome (Kondo *et al.*, 2002). In addition, a host can have multiple symbionts that might interact to produce different phenotypes (Leclair *et al.*, 2017; McLean *et al.*, 2018; Weldon, Russell and Oliver, 2020; Smee, Raines and Ferrari, 2021). They also become fixed or segregate out of populations in non-random ways, therefore, they cannot be impact free (Jaenike, 2007).

Table 1.1. Examples of the various types of symbiosis

Type of Symbiosis	Symbiont	Host species	Effect	Reference
Nutritional symbiosis	*Buchnera aphidicola*	Aphids and leaf hoppers	B vitamin production supplements the lack of nutrients found in a diet of plant sap	(Blow *et al.*, 2020)
	Stammera sp.	Cassidinae (Tortoise beetles)	Pectinase diversity of symbionts allows host to consume a broader variety of plants	(Salem *et al.*, 2020)
	Methylococcales	*Laminatubus* and *Bispira* deep sea worms	Enable chemoautotrophic energy production via methanogenesis in deep sea vents	(Goffredi *et al.*, 2020)
Protective Symbiosis	*Spiroplasma*	*Drosophila* sp. flies	Protection against parasitoid wasps	(Xie, Vilchez and Mateos, 2010; Xie *et al.*, 2014)
	Rickettsia	*Bemisia tabaci* (white fly)	Protection against entomopathogenic fungi	(Hendry, Hunter and Baltrus, 2014)
	Parachlamydia acanthamoebae	*Acanthamoeba castellanii*	Protection against giant viruses by disrupting the production of viral factories	(Arthofer *et al.*, 2022)
	Buchnera aphidicola	*Acyrthosiphon pisum* (Pea aphid)	Thermal tolerance is reduced. *Buchnera* dies at high temperature, rendering their host unable to reproduce	(Dunbar *et al.*, 2007)
	Rickettsia	*Bemisia tabaci* (white fly)	Reduces insecticide resistance	(Kontsedalov *et al.*, 2008)
Reproductive Manipulation	*Rickettsia*	*Adalia* sp. (Ladybirds)	Male killing	(Hurst *et al.*, 1994)
	Spiroplasma	*Drosophila* sp., *Mallada desjardinsi* (green lacewing)	Male killing	(Montenegro *et al.*, 2005; Hayashi *et al.*, 2016)
	Wolbachia	*Asobara tabida* (parasitoid wasp)	Toxin-anti toxin competition leading to dependence on the bacteria for ovarioles to function	(Kremer *et al.*, 2009)
	Wolbachia	Culex pipiens, Aedes aegypti (Mosquitoes)	Cytoplasmic incompatability	(Altinli *et al.*, 2018)
	Rickettsia, *Cardinium*	*Pnigalio soemius, Encarsia hispida* (parasitoid wasps)	Induces parthenogenesis by suppressing male production in species with haplodiploid females	(Giorgini et al., 2009, 2010)

The relationship between a host and its symbiont can be tracked through time via their shared phylogeny. Some symbiont-host partnerships are ancient and obligate, and their evolution strongly linked to one another, as is seen with *Blattabacterium* in termites and cockroaches (Lo, 2003) or *Buchnera* in aphids (Munson, Baumann and Kinsey, 1991). Some ancient facultative symbionts like *Rickettsia* and *Wolbachia* have long standing relationships with several distant hosts and exhibit transfers between them over evolutionary long periods of time (Lefoulon *et al.*, 2016; Pilgrim *et al.*, 2021). Other more recent symbionts like *Hamiltonella* in aphids and whitefly occur commonly but in narrow host ranges and may also transfer between hosts (Kaech and Vorburger, 2021; Wu *et al.*, 2022).

Unfortunately, most of the phenotypes that have been studied are restricted to a few species or have only been studied in laboratory settings that can never truly reflect their ecology. The vast majority of symbionts and endosymbionts do not have much information besides 16S rRNA or other conserved marker sequences. For instance, there are over 700 instances of Torix group *Rickettsia*) infection in invertebrates alone, but only two available genomes prior to the work in this thesis (Pilgrim *et al.*, 2017; Wang *et al.*, 2020). Similarly, the phylum Chlamydiota has eight disputed official families, but has been described as having anywhere from 181 to over 1000 undescribed families (Lagkouvardos *et al.*, 2014; Collingro, Köstlbacher and Horn, 2020). The family Rhabdochlamydiaceae alone is considered to have a putative 388 unnamed genera and 14,051 species (Halter *et al.*, 2022).

1.1.2 Symbiotic bacteria in invertebrates

Invertebrates are infected by a wide diversity of symbiotic bacteria. They are found in everything from sponges and annelids to arthropods and molluscs (Weinert *et al.*, 2015; Goffredi *et al.*, 2020; Pilgrim *et al.*, 2021; Carrier *et al.*, 2022). These symbioses have wide ranging effects on nutrition, defence, growth, and reproduction, with some of the best studied examples found in terrestrial arthropods. Scientists have estimated that around 70% of terrestrial arthropods form close symbioses with bacteria, including *Wolbachia* which could infect around 50% of arthropod species (Weinert *et al.*, 2015).

Where symbionts are only inherited though the female line bacteria sometimes evolve the ability to kill males, transform them into females, or prevent uninfected males from

reproducing with infected females (cytoplasmic incompatibility). In some cases, this phenotype establishes in an evolutionary arms race to between the host's reproductive success and the bacteria maximising it passage into the next generation (Jiggins, Hurst and Majerus, 2000; Duplouy and O'Neill, 2010; Reynolds *et al.*, 2019). Further, reproductive manipulation can change reproductive behaviour (Jiggins, Hurst and Majerus, 2000), or fundamentally alter how a species reproduces, for instance by inducing parthenogenesis (Giorgini *et al.*, 2010).

In other cases, the host may not be able to reproduce without the symbiont as a by-product of escalating competition. For instance, the wasp *Asobara tabida* has become so dependent on *Wolbachia* that it destroys its own ovaries in the absence of the symbiont (Pannebakker *et al.*, 2007). Meanwhile in white fly, the symbiont *Rickettsia* provides reproductive advantage to only females, by increasing growth rates and fecundity (Himler *et al.*, 2011).

Pea aphids, *Acyrthosiphon pisum*, and their primary symbiont *Buchnera aphidicola* represent arguably the most established system of an insect and its community of symbionts (Buchner, 1965). They have obligate nutritional symbionts like *Buchnera aphidicola* and sometimes *Erwinia*, that supplement the aphid's all sap diet with essential amino acids and B vitamins (Feng *et al.*, 2019; Blow *et al.*, 2020; Manzano-Marín *et al.*, 2020). In addition, aphids are often infected with a host of other facultative symbionts like *Hamiltonella defensa, Rickettsia* and *Spiroplasma* that provide protection against parasitoid wasps and fungal infection (Oliver *et al.*, 2003, 2008; Łukasik *et al.*, 2013). The efficacy and costs of these communities varies with the make-up of the symbionts within the aphid as well as the environment they are in (Oliver *et al.*, 2008; Leclair *et al.*, 2017; Smee, Raines and Ferrari, 2021).

The aphid's ecology and evolution are intrinsically linked to their community of symbionts (Smee, Raines and Ferrari, 2021). They would not be able to persist in their ecological niche without *Buchnera* to supplement their nutritionally poor diet, at the cost of being vulnerable to certain stresses that impact the symbiont. For instance, the presence of heat intolerant obligate symbionts effectively restricts the host to specific temperature ranges. A point mutation in *Buchnera* reduces the bacteria's heat tolerance (Dunbar *et al.*, 2007)

and aphids cannot reproduce without *Buchnera* because they are unable to make the necessary amino acids by themselves.

1.1.3 Symbiotic bacteria in microeukaryotes

Endosymbionts in microeukaryotes have similar effects to those in invertebrates and some of the earliest symbionts described were recognised first in microeukaryotes (Penard, 1902). For instance, killer particles (i.e. *Caedibacter*) were some of the first known instances of reproductive manipulation by bacteria (Sonneborn, 1943; Schrallhammer, 2010). Some *Caedibacter* produce a toxin that can only be counteracted by its own antitoxin systems. If a *Paramecium* does not have *Caedibacter* in an infected population, the uninfected individuals will die. It also allows infected individuals to reach higher densities, providing a fitness advantage (Grosser *et al.*, 2018). More recently, examples of protective symbioses have been observed between *Parachlamydia acanthamoeba* and amoeba, where *P. acanthamoeba* is able to prevent giant viruses from forming viral factories in its host (Arthofer *et al.*, 2022).

One interesting example of extreme nutritional symbiosis can be found in *Pelomyxa palautris*. This amoeba appears to have no mitochondria and is missing several other organelles (Daniels and Breyer, 1967; Gutiérrez *et al.*, 2017). Instead, it has several endosymbiotic bacteria, including methanogens, which seem to replace the function of these missing organelles (Gutiérrez *et al.*, 2017).

Little is known about the phenotypes and genotypes of most microeukaryote-bacteria relationships, and less is known about their ecology and distribution. For the most part, the diversity of infections in microeukaryotes has been underreported outside of 'model' host clades. However, there is a growing body of contemporary information that expands on old research and supports the idea of broad and unexplored symbioses in microeukaryotes. For some groups such as Chlamydiota and Rickettsiales, theory suggests that they began as symbionts of microeukaryotes before evolving into insect symbionts or mammalian pathogens (Driscoll *et al.*, 2013; Schrallhammer *et al.*, 2013; Kang *et al.*, 2014). For example, several putatively "ancient" relatives of the pathogenic and insect symbiotic bacteria *Rickettsia* have only been described in ciliates and algae (e.g '*Ca*. Trichorickettsia', and '*Ca*. Megaira') (Schrallhammer *et al.*, 2013; Sabaneyeva *et al.*, 2018; Castelli *et al.*, 2019). This pattern of evolution makes sense given the ancient nature of

microeukaryotes compared to multicellular animals and plants. Diverse endosymbionts circulate amongst microeukaryotes, and a subset of these have established lineages within animals and plant hosts.

1.2 The benefits of bioinformatics for studying obscure endosymbiotic bacteria

Bioinformatics is the art of using data science to explore large and complex biological information. It provides insightful exploratory and analytical tools that can be used to glean information on the evolution and potential function of organisms based on proteome or genome or other biological data. Many bacteria are only known from mass sequencing efforts of environmental DNA (Lagkouvardos *et al.*, 2014; Pilgrim *et al.*, 2021). Bioinformatics provides an entry point for studying these unculturable or obscure bacteria that would otherwise remain unexplored.

Bioinformatics also allows us to find novel endosymbioses. When the genome of an organism is sequenced, all bacterial DNA is sequenced alongside it. The raw sequence can then be treated in the same way as an environmental metagenome to extract bacterial genomes from it. This mix of biological entities in a single DNA pool allows us to run the aligned raw DNA reads though binning algorithms that can find clusters of similar genomic sequence that belong to unique organisms (Lagkouvardos *et al.*, 2014; Breitwieser, Lu and Salzberg, 2017; Scholz *et al.*, 2020). Those bins are then quality checked and any potential endosymbiotic bacteria are identified. Once novel endosymbiont genomes have been assembled, we can then apply phylogenomic techniques to explore its relatedness to other strains, and homology inference to assess potential function.

1.2.1 A brief summary of Phylogenomics, Comparative genomics and Metabolic Prediction

Phylogenomics is the combined study of genomics and evolution. It covers a broad range of techniques used to examine the evolution of organisms based only on their genomic information (Box 1.3). When a new microbial genome sequence is obtained, it can be placed in the phylogenetic context of relatives. For instance, singly copy orthologous genes can be aligned and used to estimate phylogenetic relatedness of core genomes. Within this, genes that are not present in all genomes can be mapped to the phylogeny,

such that genome dynamics over time can be estimated. Within the comparative genomic context, particular attention may be applied to overall metabolic competences (rather than individual gene presence/absence). For instance, in symbiosis we can examine competence for key pathways such as B vitamin or Essential amino acid synthesis – and examine where and when these have evolved. Putative defensive systems can likewise be predicted and mapped onto phylogenies.

I will be using a combination of all the above techniques to clarify the evolutionary relationships and potential functions of obscure endosymbiotic bacteria that lack this key information.

Box 1.3 Phylogenomics terminology

Pangenomics – examines the patterns in genome information and statistics across multiple related genomes. Pangenomics can be used to assess the rate of evolution within genera as well as extract core sequences shared between large numbers of genomes.

Phylogeny – examines the distances and relations between aligned amino acid or nucleotide sequences.

Average Amino Acid Identity (ANI) and Average Nucleotide Identity (ANI) – pairwise comparisons of genomes to assess how much of their amino acid or nucleotide sequence is shared. ANI and AAI can be used to assess species and genus boundaries, respectively. This is particularly important for clades where the number of type reference genome sequences are scarce.

1.3 Finding meaningful, practical study systems

Ideally, research would be carried out in an unlimited number of systems, but that is not feasible nor practical. The August Krogh principle dictates that "For many problems there is an animal on which it can be most conveniently studied" (Krebs, 1975; Jørgensen, 2001). This guideline and the general human need for taking the road of least resistance has resulted in most of our knowledge deriving from a few well-known study systems. However, researchers are beginning to experiment with more unusual systems as genomic techniques become more affordable and genetic techniques (such as CRISPR) more available in non-models (Matz, 2018; Dietrich *et al.*, 2020).

The main questions that need to be considered when choosing a potential study system are:

1. *How easy is the organism to sample and are there any ethical or legal barriers to using it?*

   If you can't buy or gather more than a few individuals, then setting up a laboratory system will be a challenge. If experiments require ethical approval, there may be limitations on sample size that can be obtained to minimize welfare concerns.
2. *How easy is the organism to look after?*

   An organism's capacity for laboratory maintenance will affect how quickly experiments can run, how easy they are to standardise, and ultimately how reproducible your results are.
3. *What resources will you need, and can they be adapted from pre-existing techniques?*

   Resources might include husbandry facilities and experimental materials, training requirements, and the available literature. If your organism or something similar has been studied before the barrier for entry will be lower.
4. *How expensive is your organism going to be to study and will you have the support of a community or institution?*

   The long-term viability of your system will depend on your ability to gather institutional and financial support. You must also consider what communities

exist around your study system and if your plans will be of interest to them. Cost needs to be weighed up with the potential interest in the new system.

5. *What makes your system interesting?*

   Whether it has medical or commercial applications or how novel your system is will affect how "useful" your system is perceived to be. Medical applications attract more funding, but is should not be the only reason you have chosen a system. Perhaps it lets you investigate fundamental biology of systems that are otherwise black boxes; these may have natural environment significance or develop into biomedical models.

1.3.1 Why Rickettsiales

The main group of bacteria I will focus on in this thesis is the Order Rickettsiales, and within that, the family Rickettsiaceae (Figure 1.2). Rickettsiales cover a wide variety of bacteria from pathogens to beneficial bacteria (Weinert *et al.*, 2009). It also contains the family thought to be the closest relative to eukaryotic mitochondria, the Midichloriaceae (Andersson *et al.*, 1998; Giannotti *et al.*, 2022). Most Rickettsiales species are obligate intracellular bacteria, meaning they cannot replicate outside cells of a host organism. To date there is a singular example of an extracellular Rickettsiales called *Deianiraea*, which is an ectoparasite of *Paramecium* (Castelli *et al.*, 2019).

The evolution and phenotypes of Rickettsiales are generally poorly defined outside of pathogenic species or the prolific endosymbiont *Wolbachia.* For the most part, Rickettsiales is made up of disparate groups of loosely related Families with poor within-clade definition. For instance, within the Rickettsiaceae family, most known examples belong to the pathogenic Spotted Fever Group in the Genus *Rickettsia* (see Chapter 2). The remaining *Rickettsia* lack the information to properly define within family evolution. For instance, the former *Rickettsia,* '*Ca*. Megaira' was known as Hydra Group *Rickettsia* for several years following its discovery (Weinert *et al.*, 2009; Kawafune *et al.*, 2015; Pilgrim *et al.*, 2017). It is in fact so genomically different from the main *Rickettsia* clades that it could be classed as its own family (see Chapter 2 and 3). As it is, '*Ca*. Megaira' was only given its own genus name in 2013 (Schrallhammer et al., 2013).

Most phenotypic information for Rickettsiales as symbionts comes from *Wolbachia* (Anaplasmaceae). It is thought to infect around 50% of all terrestrial insects (Weinert *et*

al., 2015) and it is capable of reproductive manipulation (Duplouy and O'Neill, 2010; Altinli *et al.*, 2018), nutritional symbiosis (Hosokawa *et al.*, 2010), and protective symbiosis (Braquart-Varnier et al., 2015). Outside of *Wolbachia, Rickettsia* has several similar phenotypes to *Wolbachia* and stands as a viable alternative to *Wolbachia* as a study system (see Table 1.2), that may infect around 15-20% of arthropods (Weinert *et al.*, 2015).

Rickettsiaceae provide a useful snapshot into the evolution and function of endosymbionts in this group. Members infect a broad range of hosts from ciliates and algae to beetles and molluscs (Weinert *et al.*, 2009; Pilgrim *et al.*, 2021), providing a huge evolutionary span of both bacteria and hosts. Evidence suggests that they are important components of invertebrate biology as well as algae and ciliates where, in theory, they originated (Driscoll *et al.*, 2013; Schrallhammer *et al.*, 2013). Insects infected with *Rickettsia* commonly lose their infection when taken into the laboratory (Zélé *et al.*, 2020), and have varying frequencies within species (Cass et al., 2016), suggesting that these bacteria have costs to the host or experience segregational loss that is only countered by unknown environmental influences. The impact of Rickettsiaceae in most instances is unknown, but vertical transmission means is very unlikely that they have no impact (Jaenike, 2007).

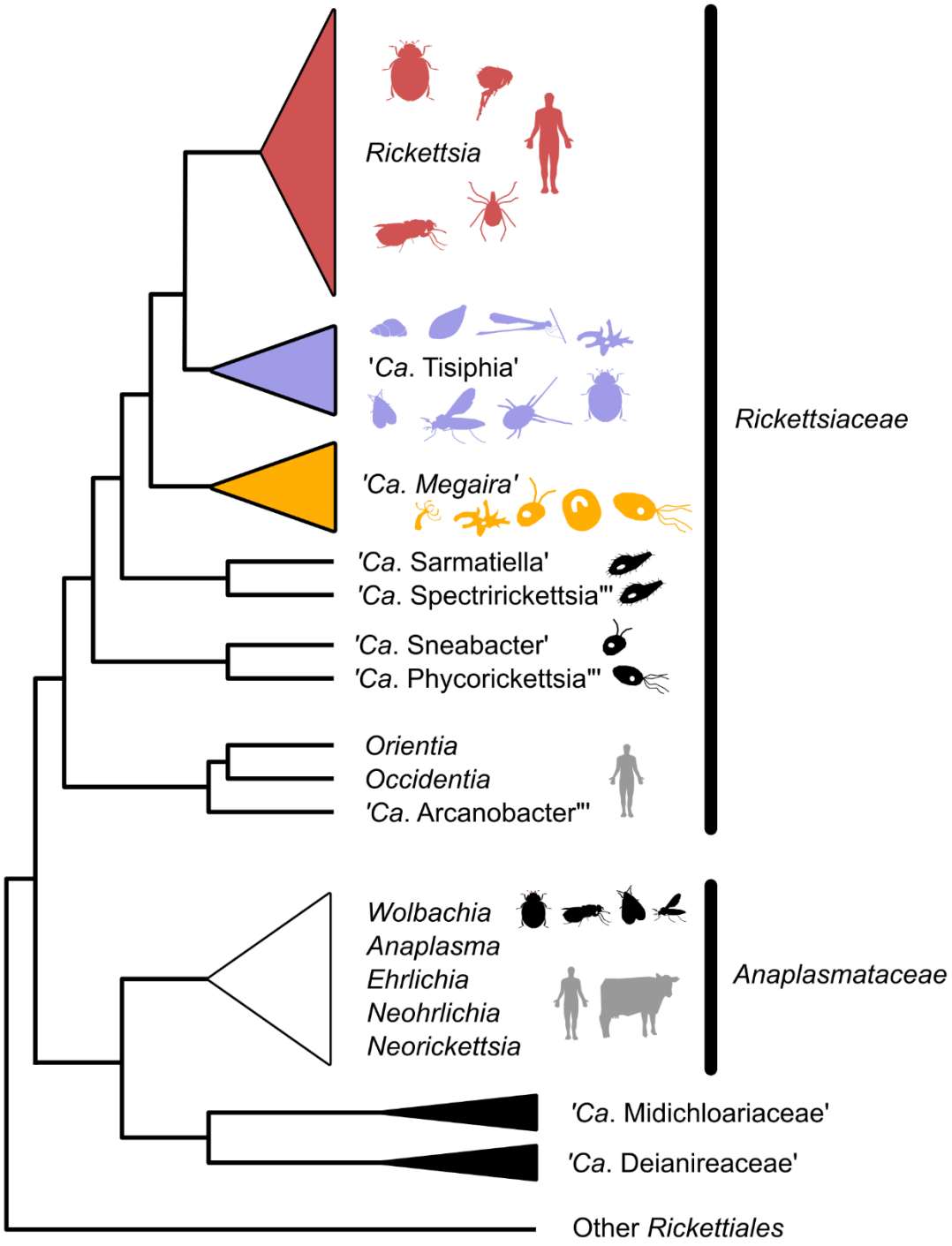


Figure 1.2. Cladogram of the main Rickettsiales clades mentioned in this thesis as currently recognised (note, '*Ca*. Tisiphia' is newly erected in Chapter 2). Any mammals shown indicate the presence of pathogenic strains.

Table 1.2. Known phenotypes of Rickettsia bacteria in various hosts

Reference	Host Order: Family	Host Species	*Rickettsia* Group	Phenotype	
(Himler *et al.*, 2011)	Hemiptera: Aleyrodidae	*Bemisia tabaci* (White fly, US)	Belli	Increased fecundity and offspring survival	Reproductive manipulation
(Hurst *et al.*, 1994; von der Schulenburg *et al.*, 2001)	Coleoptera: Coccinellidae	*Adalia decumpunctata* *Adalia bipunctata* (Ladybirds)	Adalia	Male killing	
(Lawson *et al.*, 2001)	Coleoptera: Buprestidae	*Brachys tessellatus* (Buprestid beetle)	Belli		
(Sakurai *et al.*, 2005)	Hemiptera: Aphididae	*Acyrthosiphon pisum* (Pea aphid)	Belli	Reduced fecundity	
(Hurst *et al.*, 1994)	Coleoptera: Coccinellidae	*Adalia bipunctata* (Ladybird)	Adalia		
(Giorgini *et al.*, 2010)	Hymenoptera: Eulophidae	*Pnigalio soemius* (Hymenoptera: Eulophidae, Parasitoid wasp)	Belli	Induces parthenogenesis	
(Himler *et al.*, 2011)	Hemiptera: Aleyrodidae	*Bemisia tabaci* (White fly,US)	Belli	Female bias	
(Perotti *et al.*, 2006)	Psocoptera: Liposcelididae	*Liposcelis bostrychophila*	Transitional	Egg viability	
(Biggs *et al.*, 2016; Nicholson and Paddock, 2017)	Broad reference for Ixodida ticks, fleas, and lice species unspecified	Multiple Ticks, Mites, Louse and Flea spp. between mammals, including humans	Typhus Spotted fever Transitional Orienta	Animal-animal vector transmission, pathogenic	Virulence efficiency and Vector Transmission
(Li *et al.*, 2017)	Hemiptera: Aleyrodidae Malvales: Malvaceae	*Bemisia tabaci* (White fly, MEAM1 and MED) *Gossypium hirsutum* L. var. Lumianyan no. 32 (Mexican cotton)	Belli	Plant mediated horizontal transmission	
(Kliot *et al.*, 2014)	Hemiptera: Aleyrodidae	*Bemisia tabaci* (White fly, B biotype)	Belli	Increases virulence of tomato yellow leaf curl virus	
(Sakurai *et al.*, 2005)	Hemiptera: Aphididae	*Acyrthosiphon pisum* (Pea aphid)	Belli	Suppression of coinfection	
(Kikuchi and Fukatsu, 2005)	Rhynchobdellida: Glossiphoniidae	*Torix* tagoi *Torix* tukubana (Leeches)	*Torix*	Effect on host body size/mass	Growth, nutrients, and life cycle
(Sakurai *et al.*, 2005)	Hemiptera: Aphididae	*Acyrthosiphon pisum* (Pea aphid)	Belli		
(Chiel, Inbar, *et al.*, 2009; Himler *et al.*, 2011)	Hemiptera: Aleyrodidae	*Bemisia tabaci* (White fly, B biotype and US)	Belli	Faster egg-adult development	
(Hurst *et al.*, 1994)	Coleoptera: Coccinellidae	*Adalia bipunctata* (Ladybird)	Adalia	Reduced longevity	
(Bodnar *et al.*, 2018)	Ixodida: Ixodidae	*Ixodes pacificus* (Tick)	Spotted fever	B vitamin production	
(Brumin, Kontsedalov and Ghanim, 2011)	Hemiptera: Aleyrodidae	*Bemisia tabaci* (White fly, B biotype)	Belli	Thermal tolerance	Defence and resistance
(Kontsedalov *et al.*, 2008)	Hemiptera: Aleyrodidae	*Bemisia tabaci* (White fly, B biotype)	Belli	Increased susceptibility to insecticides	
(Łukasik *et al.*, 2013)	Hemiptera: Aphididae	*Acyrthosiphon pisum* (Pea aphid)	Unspecified, likely belli	Fungal resistance	

1.3.2 Where do we find Rickettsia infection?

One of the best studied symbiont groups outside of *Wolbachia* are non-pathogenic *Rickettsia*. We know infections are variable in frequency within species (Perlman, Hunter and Zchori-Fein, 2006; Bockoven *et al.*, 2020) and are often lost when they are removed from their environment (Zélé *et al.*, 2020). This observation would suggest that there is some pressure in the environment that maintains the infection. The effects of symbionts are usually not tested in the real-world environment (Kontsedalov *et al.*, 2008; Brumin, Kontsedalov and Ghanim, 2011; Hendry, Hunter and Baltrus, 2014) and fieldwork is commonly limited to screening efforts that rarely record ecological information such as precipitation, temperature, or habitat type (Tsuchida *et al.*, 2002; Guo *et al.*, 2016; Thongprem *et al.*, 2020).

However, they do tend to record geographical coordinates which can be used to map incidence. Mapping the historical recorded incidence of non-pathogenic *Rickettsia* bacteria (including '*Ca*. Tisiphia') illustrates that it is a widespread, global infection (Figure 1.3). Several studies have suggested that *Rickettsia* are associated with wet or aquatic environments (Driscoll *et al.*, 2013; Schrallhammer *et al.*, 2013; Pilgrim *et al.*, 2021), but this has not been empirically studied. Another theory could be that *Rickettsia* are mostly

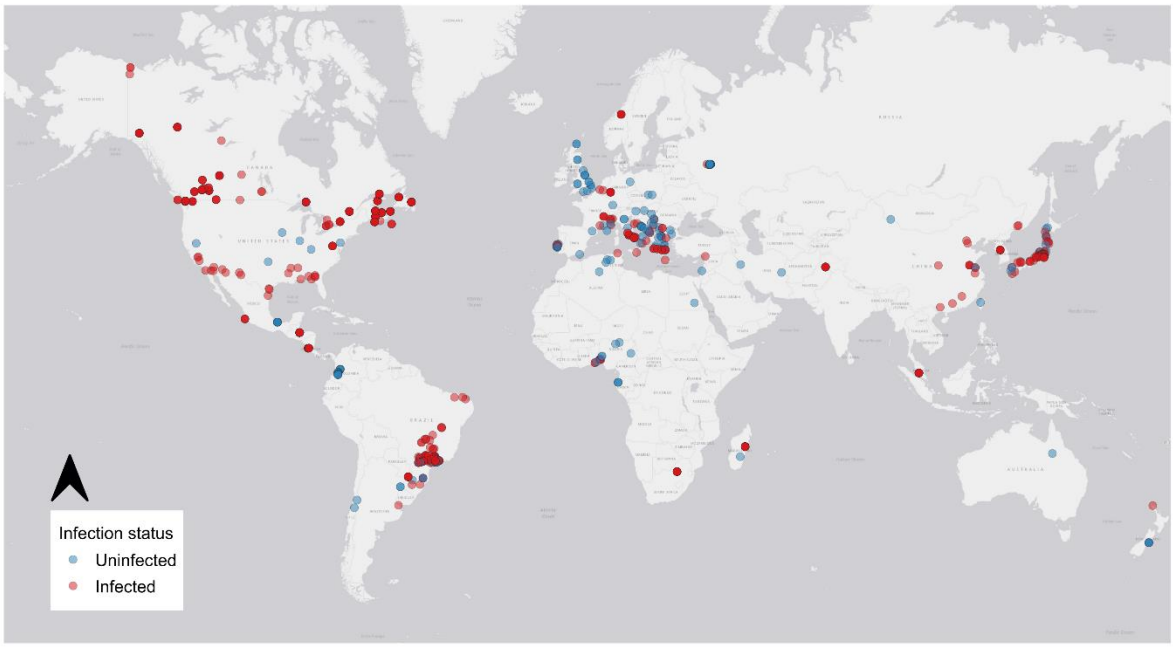


Figure 1.3. Map displaying the distribution of non-pathogenic *Rickettsia* incidences in Arthropoda. Based on 16S rRNA PCR screens in current literature. Includes species that have been tested and found negative.

protective symbionts, for instance providing defence against fungal infection. The broad scale distribution patterns of fungi can be modelled with precipitation and temperature data (Větrovský *et al.*, 2019), so if the theory is true, we may be able to model *Rickettsia* infection in the same way.

1.3.3 Why Chlamydiota

Like Rickettsiales, the class Chlamydiota comprises obligate intracellular bacteria, but with arguably poorer descriptions of evolution and phenotypes. The family Rhabdochlamydiaceae (Parachlamydiales) alone is thought to have over 388 undescribed genera based on environmental metagenome sampling (Halter *et al.*, 2022). Most are known to have pathogenic life histories in fish and mammals (Draghi *et al.*, 2004; Bayramova, Jacquier and Greub, 2018), and all have a unique biphasic lifestyle in which they live part of their lifecycle in the environment (non-reproductive), and part in the cell of a host (reproductive) (König *et al.*, 2017). However, little is known about their interaction with their native hosts, thought to be largely amoeba and other microeukaryotes (Horn, 2008; Halter *et al.*, 2022). In this thesis I will largely focus on the families Simkaniaceae and Rhabdochlamydiaceae displayed in Figure 1.4.

Though rarely observed, Chlamydiota are capable of beneficial symbiosis. Recently, *Parachlamydia acanthamoebae* has been shown to protect its *Acanthamoeba* host against viral infection by giant viruses (Arthofer *et al.*, 2022). The defence system seems to target viral factory formation, in a yet undescribed fashion, against several giant virus types including Viennavirus, Mimivirus and Tupanvirus.

Chlamydiota and Rickettsiales share several parallels in their evolution and life history. Chlamydiota share many plant-like protein pathways with plant chloroplasts and is speculated to share an ancestor with a cyanobacterium or facilitated the formation of the chloroplast in some way (Brinkman *et al.*, 2002; Horn, 2008). Similarly, Rickettsiales share characteristics of mitochondria and are thought to share ancestry with the bacteria that mitochondria arose from (Andersson *et al.*, 1998; Giannotti *et al.*, 2022). Both groups are hugely diverse and capable of infecting multiple orders of organisms across kingdoms (Horn, 2008; Weinert *et al.*, 2009; Pilgrim *et al.*, 2021; Halter *et al.*, 2022). Both are obligate intracellular bacteria, and both are capable of being pathogens, or symbionts. Both can also be passed to new hosts through the environment. Some Rickettsiales are

capable of persisting in plant phloem or transferring via contaminated piecing parts of parasites through the dirty needle effect (Chiel, Zchori-Fein, *et al.*, 2009; Li *et al.*, 2017), and all known Chlamydiota have a biphasic lifestyle (König *et al.*, 2017).

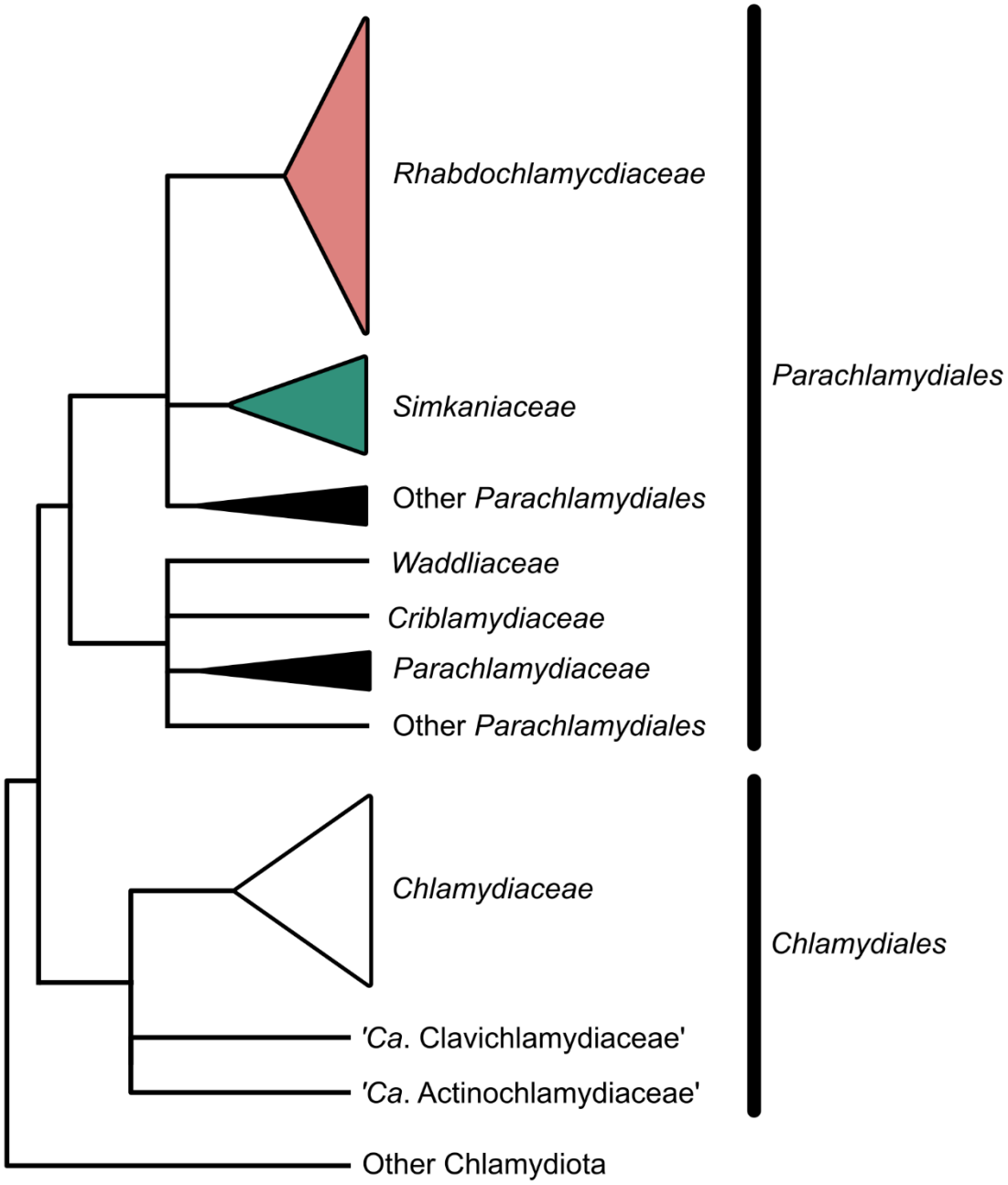


Figure 1.4. A simple cladogram showing the placement of Chlamydiota highlighting the clades examined in this thesis.

1.3.4 Applications and why we need a study system for obscure symbioses

Symbionts are now recognised as vital components of whole organism biology that can alter evolution (Charlat, Hurst and Merçot, 2003; Moran, 2007; Hurst, 2017). Beyond basic biology, symbionts may alter how an organism interacts with its environment in ways that could be applied to medicine or agriculture. Symbionts can alter how sensitive the host is to insecticides (Kontsedalov *et al.*, 2008) and thermal changes (Corbin *et al.*, 2017). They can also interfere with insect borne diseases to make vectors more (Kliot *et al.*, 2014) or less competent (Pereira *et al.*, 2018).

1.4 Outline of thesis

In this thesis I will clarify the evolution, diversity and potential function of understudied symbiotic bacteria that infect invertebrates, ciliates and algae. I will achieve this with a mixture of broad scale bioinformatics techniques to probe novel infections and new hosts, as well as laboratory work to develop new model systems.

Chapter 2. Here, I examine invertebrates for novel Rickettsiaceae. Large online databases and increasingly powerful bioinformatics tools allow for bacteria sequenced alongside their hosts to be extracted from raw genome sequence files that might otherwise not be screened. I searched the NCBI genome database for new Rickettsiaceae infection across invertebrates. In particular, I was interested in expanding the genome resources available for less well described symbiont groups like Adalia or Torix group *Rickettsia*. I apply phylogenomics and pangenomics to the new genomes to clarify *Rickettsia* evolution and examine their metabolic potential, looking for clues to their functional diversity.

Chapter 3. I use the same techniques from Chapter 2 to scour microeukaryotes for Rickettsiales symbionts. In addition, I cross reference environmental metagenomes assembles genomes to find additional bacteria within the '*Ca*. Megaira' clade that have not been classified as such. I further clarify the evolution and functions found in the more microeukaryote specific clade '*Ca.* Megaira' and highlight the huge potential for diversity in this order.

Chapter 4. While looking for Rickettsiales, it came to my attention that several ciliates and algae were infected with Chlamydiota, in particular the order Parachlamydiales. I use newly assembled genomes and uncharacterised environmental genomes to

add to the current known phylogeny of this enormously diverse group. Again, I examine metabolic potential for clues about their function in their host.

Chapter 5. Moving away from broad scale approaches, Chapter 5 focuses on one specific infection of '*Ca.* Tisiphia' (Torix group *Rickettsia*) in *Anopheles plumbeus* mosquitos. I visualise the infection to supplement phylogenomic information and theory about the infection being heritable. I also attempt to examine environmental factors that might influence rated of infection across the country of Germany.

Chapter 6. I finish by discussing the wider implications of my work, bringing together the conclusions drawn from both broad and local scale projects. I conclude that symbiont groups are more diverse than previously thought and that symbionts have a long and varied evolutionary history with their many hosts. I believe that studies to date, including my own, barely scratch the surface of host-symbiont interactions, and I hope that they prompt more specific studies of function, evolution and ecology.

Chapter 2. Rickettsiaceae in invertebrates

All analyses in this chapter are my own. The *Bryobia* mite and *Tsetse* fly *Rickettsia* genomes were initially identified and sequenced by Dr Nicky Wybouw and Dr Frances Blow respectively before incorporation into this project. The *'Ca.* Megaira' found in *Carteria* and sequenced and assembled by Stefanos Siozios and Simon Hunter-Barnett. The *'Ca.* Megaira' from a published *Mesostigma* SRA was extracted by Stefanos Siozios. '*Ca*. Tisiphia' from *Cimex lectularius* and *Culicoides newsteadi* were identified and assembled by Dr Jack Pilgrim.

This chapter has been published in two parts: Pilgrim et al 2021 and Davison et al 2022.

2.1 Abstract

Members of the bacterial genus *Rickettsia* were originally identified as causative agents of vector-borne diseases in mammals. However, many *Rickettsia* species are arthropod symbionts and close relatives of '*Ca.* Megaira', which are symbiotic associates of microeukaryotes. Here, we clarify the evolutionary relationships between these organisms by assembling 26 genomes of *Rickettsia* species from understudied groups, including the Torix group, and two genomes of '*Ca*. Megaira' from various insects and microeukaryotes. Our analyses of the new genomes, in comparison with previously described ones, indicate that the accessory genome diversity and broad host range of Torix group *Rickettsia* are comparable to those of all other *Rickettsia* combined. Therefore, the Torix clade may play unrecognized roles in invertebrate biology and physiology. We argue this clade should be given its own genus status, for which we propose the name '*Candidatus* Tisiphia'.

2.2 Introduction

Symbiotic bacteria are vital to the function of most living eukaryotes, including microeukaryotes, fungi, plants, and animals (Boettcher, Ruby and McFall-Ngai, 1996; Clay, Holah and Rudgers, 2005; Douglas, 2011; Fujishima and Kodama, 2012). The symbioses formed are often functionally important to the host with effects ranging from mutualistic to detrimental. Mutualistic symbionts may provide benefits through the biosynthesis of metabolites, or by protecting their hosts against pathogens and parasitoids (Oliver *et al.*, 2010; Hendry, Hunter and Baltrus, 2014). Parasitic symbionts can be detrimental to the

host due to resource exploitation or through reproductive manipulation that favour its own transmission over the host's (Engelstädter and Hurst, 2009; Leclair *et al.*, 2017). Across these different symbiotic relationships, symbionts are often important determinants of host ecology and evolution.

The Rickettsiales (Alphaproteobacteria) represent an order of largely obligate intracellular bacteria that form symbioses with a variety of eukaryotes (Weinert *et al.*, 2015). *Deianiraea*, an extracellular parasite of *Paramecium*, is the one known exception (Castelli *et al.*, 2019). Within Rickettsiales, the family Rickettsiaceae represent a diverse collection of bacteria that infect a wide range of eukaryotic hosts and can act as symbionts, parasites, and pathogens. Perhaps the best-known clade of Rickettsiaceae is the genus *Rickettsia*, which was initially described as the cause of spotted fever and other rickettsial diseases in vertebrates that are transmitted by ticks, lice, fleas and mites (Angelakis and Raoult, 2017).

Rickettsia have been increasingly recognised as heritable arthropod symbionts. Since the description of a maternally inherited male-killer in ladybirds (Werren *et al.*, 1994), we now know that heritable *Rickettsia* are common in arthropods (Weinert *et al.*, 2015; Pilgrim *et al.*, 2021). Further, *Rickettsia*-host symbioses are diverse, with different symbionts being capable of reproductive manipulation, nutritional and protective symbiosis, as well as influencing thermotolerance and pesticide susceptibility (Hurst *et al.*, 1994; Kontsedalov *et al.*, 2008; Chiel, Inbar, *et al.*, 2009; Giorgini *et al.*, 2010; Brumin, Kontsedalov and Ghanim, 2011; Łukasik *et al.*, 2013; Bodnar *et al.*, 2018).

Our understanding of the evolution and diversity of the genus *Rickettsia* and its allies has increased in recent years, with the taxonomy of Rickettsiaceae developing as more data becomes available (Gillespie *et al.*, 2007; Weinert *et al.*, 2009). Weinert et al. (2009) loosely defined 13 different groups of *Rickettsia* based on 16S rRNA phylogeny, which showed two early branching clades that appeared genetically distant from other members of the genus. One of these was a symbiont of *Hydra* and designated as Hydra group *Rickettsia*, which has since been assigned its own genus status, '*Candidatus* Megaira' (Schrallhammer *et al.*, 2013). '*Ca.* Megaira' forms a related clade to *Rickettsia* and is found in in ciliates, amoebae, chlorophyte and streptophyte algae, and cnidarians (Lanzoni *et al.*, 2019). Members of this clade are found in hosts from aquatic, marine and soil habitats

which include model organisms (e.g., *Paramecium*, *Volvox*) and economically important vertebrate parasites (e.g., *Ichthyophthirius multifiliis*, the ciliate that causes white spot disease in fish) (Lanzoni *et al.*, 2019). Whilst symbioses between '*Ca.* Megaira' and microeukaryotes are pervasive, there is no publicly available complete genome and the impact of these symbioses on the host are poorly understood.

A second early branching clade was described from *Torix tagoi* leeches and is commonly coined Torix group *Rickettsia* (Kikuchi and Fukatsu, 2005). Symbionts in the Torix clade have since been found in a wide range of invertebrate hosts from midges to freshwater snails to fish-parasitic amoeba (Pilgrim *et al.*, 2021). The documented diversity of hosts is wider than other *Rickettsia* groups, which are to date only found in arthropods and their associated vertebrate or plant hosts (Weinert *et al.*, 2009). Torix clade *Rickettsia* are known to be heritable symbionts, but their impact on host biology is poorly understood, despite the economic and medical importance of several hosts (inc. bed bugs, black flies, and biting midges). A few studies have described the potential effects on the host, which include: larger body size in leeches (Kikuchi and Fukatsu, 2005); a small negative effect on growth rate and reproduction in bed bugs (Thongprem *et al.*, 2020); and an association with parthenogenesis in *Empoasca* Leafhoppers (Aguin-Pombo *et al.*, 2021).

Current data suggest an emerging macroevolutionary scenario where the members of the *Rickettsia* clade originated as symbionts of microeukaryotes, before diversifying to infect invertebrates (Driscoll *et al.*, 2013; Schrallhammer *et al.*, 2013; Kang *et al.*, 2014). Many symbionts belonging to the Rickettsiaceae (e.g., '*Ca.* Megaira', '*Candidatus* Trichorickettsia', '*Candidatus* Phycorickettsia', '*Candidatus* Sarmatiella' and '*Candidatus* Gigarickettsia') circulate in a variety of microeukaryotes (Schrallhammer *et al.*, 2013; Vannini *et al.*, 2014; Sabaneyeva *et al.*, 2018; Yurchenko *et al.*, 2018; Castelli *et al.*, 2019). The Torix group *Rickettsia* retained a broad range of hosts from microeukaryotes to arthropods (Pilgrim *et al.*, 2021). The remaining members of the genus *Rickettsia* evolved to be arthropod heritable symbionts and vector-borne pathogens (Perlman, Hunter and Zchori-Fein, 2006; Weinert *et al.*, 2009). However, a lack of genomic and functional information for symbiotic clades limits our understanding of evolutionary transitions within *Rickettsia* and its related groups. No '*Ca.* Megaira' genome sequences are currently publicly available and of the 165 *Rickettsia* genome assemblies available on the NCBI (as

of 29/04/21), only two derive from the Torix clade and these are both draft genomes. In addition, dedicated heritable symbiont clades of *Rickettsia*, such as the Rhyzobius group, have no available genomic data, and there is a single representative for the Adalia clade. Despite the likelihood that heritable symbiosis with microeukaryotes and invertebrates was the ancestral state for this group of intracellular bacteria, available genomic resources are heavily skewed towards pathogens of vertebrates.

In this chapter I establish a richer base of genomic information for heritable symbiont *Rickettsia* and '*Ca.* Megaira', then use these resources to clarify the evolution of these groups. Genomic data is broadened through a combination of targeted sequencing of strains without complete genomes, and metagenomic assembly of *Rickettsia* strains from arthropod genome projects. These establish the first closed circular genome of a '*Ca.* Megaira' symbiont from a streptophyte alga (*Mesostigma viride*) and provide a draft genome for a second '*Ca.* Megaira' from a chlorophyte (*Carteria cerasiformis*). In addition, the complete genomes of two Torix *Rickettsia* from a midge (*Culicoides impunctatus*) and a bed bug (*Cimex lectularius*) are presented, as well as a draft genome for *Rickettsia* from a tsetse fly (*Glossina morsitans submorsitans*, an important vector species), and a new strain from a spider mite (*Bryobia graminum*). A metagenomic approach established a further 22 draft genomes for insect symbiotic strains, including previously unsequenced Rhyzobius and Meloidae group draft genomes. I utilize these to carry out pangenomic, phylogenomic and metabolic analyses of our extracted genome assemblies, with comparisons to existing *Rickettsia*.

2.3 Methods

2.3.1 *Genomic data collection and construction*

Two different workflows were employed to assemble genomes for '*Ca.* Megaira' and *Rickettsia* symbionts (Figure 2.1). a) Targeted sequencing and assembly of focal '*Ca.* Megaira' and Torix *Rickettsia*. b) Assembly from SRA deposits of '*Ca.* Megaira' from *Mesostigma viride* NIES296 and as well as 29 arthropod SRA deposits that potentially harbour *Rickettsia* identified in Pilgrim et al (2021). These 29 were identified as follows: 60,409 Arthropod SRA dataset from NCBI (as of the 20th May 2019) were filtered. To reduce the bias from over-represented laboratory model species (e.g. *Drosophila* spp., *Anopheles* spp.), a single dataset from 1,341 arthropod species was examined. Where

multiple data sets existed for a species, that with the largest read count was retained. Rickettsiaceae presence was identified with default phyloFlash parameters, which finds, extracts, and identifies SSU rRNA sequences (Gruber-Vodicka, Seah and Pruesse, 2020). The microbial composition of all SRA datasets that did not result in a reconstructed Rickettsiaceae 16S rRNA with phyloFlash were re-evaluated using Kraken2 (Wood, Lu and Langmead, 2019), a k-mer–based taxonomic classifier for short DNA sequences. A cut-off of ≥40,000 reads assigned to Rickettsia taxa was applied for reporting potential infections (theoretical genome coverage of ∼1–4× assuming an average genome size of ∼1.5 Mb). As *Rickettsia*-infected protists and parasitoids have previously been reported (Hagimori *et al.*, 2006; Dyková *et al.*, 2013; Galindo *et al.*, 2019), phyloFlash was also used to identify reads aligned to these taxa to account for potential positive results attributed to ingested protists or parasitisms. All new genome assemblies were analysed alongside previously assembled genomes from the genus *Rickettsia*, and the outgroup taxon *Orientia tsutsugamushi*, a distant relative of *Rickettsia* species (Tamura *et al.*, 1995). DNA preparation, sequencing strategies and symbiont assembly methodologies varied between species and are listed in the following sections and in Figure 2.1. The pipeline used to assemble genomes from Short Read Archive (SRA) data is deposited on Zenodo (Davison, 2022).

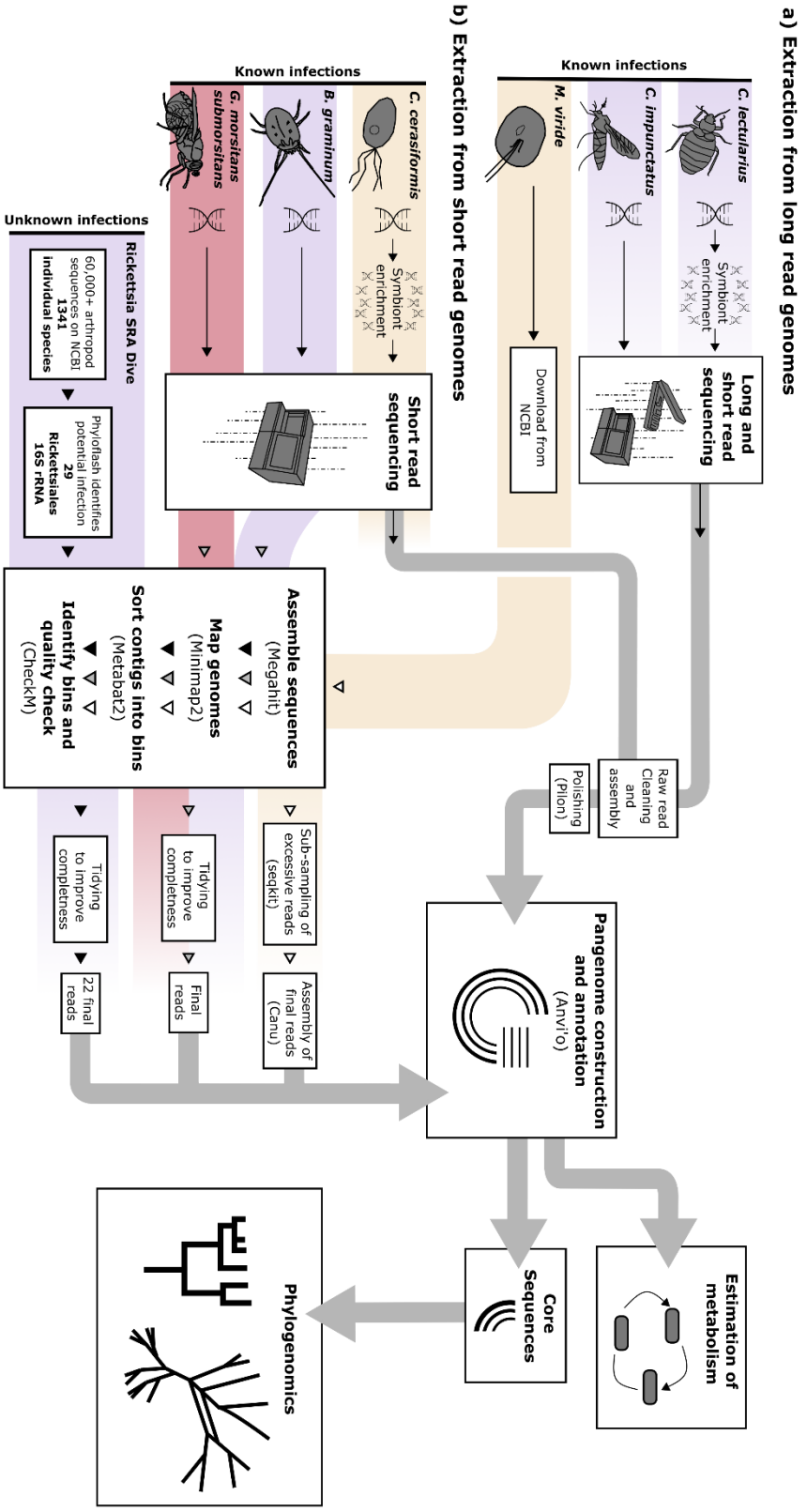


Figure 2.1. Workflow diagram for extraction, assembly and analyses performed in this chapter. Workflows for genome assemble are illustrated for a) long read host insect sequences and b) short read host insect sequences. Purple highlights Torix *Rickettsia* and orange highlights '*Ca*. Megaira' and red highlights Transitional *Rickettsia*. Sequencing technologies used vary with source and include Illumina short read sequencing, BGI DNBseq, Oxford Nanopore and PacBio.

2.3.2 *Sample collection for targeted genome assembly*

Cimex lectularius were acquired from the 'S1' isofemale colony maintained at the University of Bayreuth described in Thongprem et al (2020). *Culicoides impunctatus* females were collected from a wild population in Kinlochleven, Scotland (56° 42' 50.7''N 4° 57' 34.9''W) on the evenings of the 2nd and 3rd September 2020 by aspiration. *Carteria cerasiformis* strain NIES 425 was obtained from the Microbial Culture Collection at the National Institute for Environmental Studies, Japan. The *Glossina morsitans submorsitans* specimen Gms8 was collected in Burkina Faso in 2010 and *Rickettsia* infection was present alongside other symbionts as described in Doudoumis et al., (2017). The assembly itself is a result of later thesis work (Blow, 2017).

A *Bryobia* mite community was sampled from herbaceous vegetation in Turku, Finland. The Moomin isofemale line was established by isolating a single adult female and was maintained on detached leaves of *Phaseolus vulgaris* L. cv Speedy at 25 °C, 60 % RH, and a 16:8 light:dark photoperiod. The Moomin spider mite line was morphologically identified as *Bryobia graminum* by Prof Eddie A. Ueckermann (North-West University).

2.3.3 *Previously published Rickettsia genomes*

A total of 86 published *Rickettsia* genomes, and one genome from *Orientia tsutsugamushi,* were retrieved from the European Nucleotide Archive and assessed with CheckM v1.0.13 (Parks *et al.*, 2015). Inclusion criteria for genomes were high completeness (CheckM > 90%), low contamination (CheckM < 2%) and low strain heterogeneity (Check M < 50%) except in the case of Adalia for which there is only one genome (87.6% completeness). Filtering identified 76 high quality *Rickettsia* genomes that were used in all subsequent analyses (Appendix B.1).

2.3.4 *High molecular weight DNA extraction, assembly, and annotation of complete genomes for two 'Ca. Tisiphia' (==Torix group Rickettsia) from Culicoides impunctatus and Cimex lectularius*

High molecular weight (HMW) genomic DNA was prepared using four-hundred and eighty whole *C. impunctatus* and 45 *C. lectularius* heads, the latter of which had been symbiont-enriched using a protocol designed to eliminate host nuclei through filtration (Stouthamer, Kelly and Hunter, 2018). *Culicoides impunctatus* individuals were pooled and homogenised in two 1.5 ml Eppendorf tubes containing 0.9 ml of buffer G2 (Qiagen) using

a pestle while the filtrate from the enriched *C. lectularius* heads was also split and diluted to the same volumes. Twenty-five µL of proteinase K (50 mg/ml) was added to each Eppendorf before incubation at 56°C for 90 minutes with gentle inversion every 30 minutes. The respective lysates were centrifuged at 12,000 xg for 20 minutes before the supernatants were pooled and diluted to 3 ml with buffer G2. After equilibrating a Genomic-tip 20/G (Qiagen) with 1 ml QBT buffer, the lysate was gently inverted before being poured onto the tip membrane. The tip was washed four times with 1 ml of QC buffer (Qiagen) before elution of the DNA using buffer QF (Qiagen). Using wide-bore pipette tips, 667 µL of the eluate was pipetted into three 1.5 mL Eppendorf tubes before the addition of 467 µL isopropanol to each tube and mixing by gentle inversion 10 times. Genomic DNA was pelleted by centrifuging for 20 min at 15,000 xg at 4°C and washing twice with 70% ethanol before resuspending in buffer EB (Qiagen). Quality control of HMW DNA was then confirmed by running on a gel and assessment by Qubit fluorometric quantitation.

Long-read libraries for Oxford Nanopore sequencing were generated using the SQK-LSK109 Ligation Sequencing Kit and sequenced on Minion R9.4.1 flow cells at the Centre for Genomic Research, University of Liverpool, United Kingdom. Raw Nanopore reads were base called using Guppy version 4.0.15 (Oxford Nanopore Technologies Ltd., 2020) using the high accuracy model option (-c dna_r9.4.1_450bps_hac.cfg). All reads which were over 500bp in length and had an average phred (Q) score of > 10 were filtered using NanoFilt version 2.7.1 (De Coster *et al.*, 2018). These reads were assembled with Flye version 2.8.1 (Kolmogorov *et al.*, 2019) using default options.

Assembled circular contigs of ~1.5Mb in length were confirmed for *Rickettsia* identity by BLASTing against a *Rickettsia* genomic database. High quality short-read libraries were also generated from the same DNA samples and used to correct the nanopore assemblies. *C. impunctatus* paired-end library (2 x 150bp) was prepared using a Kapa HyperPrep kit (Roche) and sequenced by BGI Genomics (Hong Kong) on a DNBseq platform, whereas *C. lectularius* sequencing was carried out by BGI Genomics (Hong Kong) on a Hiseq Xten PE150 platform. Data cleaning and filtering was performed by BGI Genomics' using SOAPnuke version 2.1.4 (Chen *et al.*, 2018) removing adapters and any reads with 50% of bases having phred scores lower than 20.

Remaining reads were assembled with MEGAHIT version 1.2.9 (Li *et al.*, 2015) using default parameters and contigs were binned using MetaBAT2 version 2.12.1 (Kang *et al.*, 2019). The identities of bins were checked with CheckM version 1.1.3 (Parks *et al.*, 2015) and DNBseq reads were mapped to contigs from the *Rickettsia* allocated bin using 'perfect mode' in BBMap version 38.87 (Bushnell, 2015) and filtered using SAMtools version 1.11 (Li *et al.*, 2009). Filtered *Rickettsia* reads were then used to polish the Flye assembled *Rickettsia* genomes using two rounds of polishing with Pilon version 1.23 (Walker *et al.*, 2014) and the '--bases' option for correcting SNPs and small indels. Annotation of the polished genomes was accomplished using PROKKA version 1.13 (Seemann, 2014) and identification of polyketide and non-ribosomal peptide synthases was conducted by antiSMASH version 6.0 (Blin *et al.*, 2021).

2.3.5 *Extraction and assembly of a complete 'Ca. Megaira' from Mesostigma viride*

'Ca. Megaira' genome was extracted from recently published reads of *Mesostigma viride* NIES296 (from accession PRJNA509752). Illumina reads were de novo assembled using MEGAHIT version v1.2.9 (Li *et al.*, 2015), reads were mapped back to the assembled contigs. Contigs were clustered and binned based on nucleotide composition and coverage using MetaBAT2 v2:2.15 (Kang *et al.*, 2019) and a minimum contig length of 1.5kb. The quality of *'Ca.* Megaira' genome bin was inspected using CheckM (Parks *et al.*, 2015). The PacBio reads were mapped on the Illumina draft assembly and reads of '*Ca*. Megaira' origin were extracted. Due to the excessive number of obtained reads a sub-sample (reads > 10k and 1/3 of the total) was taken using seqkit (Shen *et al.*, 2016) and used for subsequent analysis. This sub-sample of PacBio reads was assembled using Canu version 1.8 (Koren *et al.*, 2017) under default parameters. The final assembly, consisting of two contigs, was manually inspected for circularization and trimmed accordingly. The final and circular assembly was further polished by a combination of PacBio and Illumina reads using Pilon v1.22 (Walker *et al.*, 2014).

2.3.6 *Extraction of Transitional Rickettsia, RiTSETSE, from Glossina morsitans submorsitans*

All methods described here for the extraction of *G. morsitans submorsitans* originates from a thesis by Frances Blow (Blow, 2017).

DNA was extracted immediately using the CTAB (Cetyl trimethylammonium bromide) method and was stored at -20°C. Whole Genome Shotgun (WGS) libraries were prepared with the Illumina TruSeq Nano DNA kit following the manufacturers' instructions. Samples were sequenced on two lanes of Illumina HiSeq with 250 bp paired-end reads. Raw sequencing reads were de-multiplexed and converted to FASTQ format with CASAVA version 1.8 (Illumina, 2011). Cutadapt version 1.2.1 (Martin, 2011) was used to trim Illumina adapter sequences from FASTQ files. Reads were trimmed if 3 bp or more of the 3' end of a read matched the adapter sequence. Sickle version 1.200 (Joshi and Fass, 2011) was used to trim reads based on quality: any reads with a window quality score of less than 20, or which were less than 10 bp long after trimming, were discarded.

Metagenomic reads were assembled with DISCOVAR (Broad Institute, 2013) and contigs shorter than 500 bp were removed and mapping with Bowtie2 (Langmead and Salzberg, 2012) was used to assess coverage. Taxonomy was assigned to contigs with BLAST and the GC content of contigs assessed with the Blobology package (Kumar *et al.*, 2013). Contigs were filtered based on GC content, coverage and taxonomy, and reads were extracted using scripts implemented in Blobology. Extracted reads were re-assembled with SPAdes version 3.7.1 (Nurk *et al.*, 2013) and mapped to contigs with Bowtie2. Assembly statistics were calculated with custom perl scripts and Qualimap version 2.2 (Okonechnikov, Conesa and García-Alcalde, 2016).

2.3.7 *DNA extraction of Moomin 'Ca. Tisiphia' (== torix Rickettsia) from Bryobia graminum str. moomin*

Genomic DNA was extracted from ~1000 adult females using the Quick-DNA Universal kit (BaseClear, the Netherlands) and was sequenced by GENEWIZ on an Illumina NovaSeq instrument. *Rickettsia* sequence was extracted from illumina reads as described for other MAGs.

2.3.8 *DNA extraction of 'Ca. Megaira' from Carteria cerasiformis*

Symbiont enriched DNA was extracted from culture using a modified version of the protocol of Stouthamer et al. (2018). Specifically, prior to homogenization the *Carteria cerasiformis* culture was filtered through a 100um filter/mesh to reduce bacterial contamination. DNA extraction was performed using the QIAGEN DNAeasy™ Blood & Tissue Kit. Short read sequencing was carried out by BGI Genomics (Hong Kong) on a Hiseq

Xten PE150 platform. *Rickettsia* sequences were assembled from Illumina reads as described for other MAGs.

2.3.9 *Assembly, and annotation of Rickettsia genomes from publicly available SRA data*

Within the study of Pilgrim et al. (2021), I identified 29 SRA deposits containing *Rickettsia* DNA. I used these datasets to extract and assemble 22 new high quality draft *Rickettsia* genomes. Briefly, short reads from each SRA library were assembled using MEGAHIT v1.2.9 (Li *et al.*, 2015), mapped with Minimap 2 v2.17-r941 (Li, 2018) and contigs were binned based on tetranucleotide frequencies using MetaBAT2 v2:2.15 (Kang *et al.*, 2019). *Rickettsia* like bins were quality inspected with CheckM v1.0.13 (Parks *et al.*, 2015). Bins with a completeness score of over 50% and contamination below 2% marked as Rickettsiales or *Rickettsia* were then retained onward for further refinement, annotation, and scrutiny.

To refine MAGs, insect SRA contigs were compared against a local *Rickettsia* genome database using Blastn (Camacho *et al.*, 2009). Contigs with significant matches to the database were extracted, non-*Rickettsia* contigs were identified with blastx against the nr database and contigs with atypical coverage were discarded. MetaBAT2 filtered out reads less than 1.5kb long for accuracy, but these reads are potentially informative in small symbiont genomes, so contigs with a length of 1-2.5kb were manually examined and added to MetaBAT2 assembled genomes. Those with improved CheckM score and no *Wolbachia* in the original host are used as the final draft genome for the *Rickettsia*. The additional genome for the leech *Rickettsia*, RiTBt, was found to contain *Cardinium* contamination during separate examination. RiTBt contigs identified as *Cardinium* using blastx were removed from the genome, reducing contamination from 9.48% to 0.95%. The final pipeline resulted in 22 MAGs each with completeness >90% and contamination <2%.

2.3.10 *Genome content comparison and pangenome construction*

Anvi'o 7 (Eren *et al.*, 2021) was used to construct a pangenome. Included in this were the 22 MAGs retrieved from SRA data, 2 '*Ca.* Megaira' genomes and 4 targeted Torix *Rickettsia* genomes, and one Transitional group *Rickettsia* genome acquired in this chapter. To these were added the 76 published and 1 *Orientia* described above, giving a total of 104 genomes. Individual Anvi'o genome databases were additionally annotated with HMMER,

KofamKOALA, and NCBI COG profiles (Eddy, 2011; Aramaki *et al.*, 2020; Galperin *et al.*, 2021). For the pangenome itself, orthologs were identified with NCBI blast, mcl inflation was set to 2, and minbit to 0.5. Average nucleotide sequence identity was calculated using pyANI (Pritchard *et al.*, 2016) within Anvi'o 7 and Average Amino Acid identity was calculated though the Kosta Lab online calculator (Rodriguez-R and Konstantinidis, 2016). Networks of ANI and AAI results were produced in Gephi 0.9.2 (Bastian, Heymann and Jacomy, 2009) with Frutcherman Reingold layout and annotated in Inkscape 0.92 (Inkscape Project, 2020). Exact code and a list of packages used is available on Zenodo (Davison, 2022).

KofamKOALA annotation (Aramaki *et al.*, 2020) in Anvi-o 7 was used to estimate completeness of metabolic pathways and Pheatmap (Kolde, 2019) in R 3.4.4 (R Core Team, 2020) was then used to produce heatmaps of metabolic potential. Annotations for function and *Rickettsia* group were added post hoc in Inkscape.

The biotin operon found in the genome Rhyzobius *Rickettsia*, Oopac6, was identified from metabolic prediction. To confirm Oopac6 carries a complete biotin pathway that shares ancestry with the existing *Rickettsia* biotin operon, Oopac6 biotin was compared to biotin pathways from five other related symbionts: *Cardinium*, *Lawsonia*, *Buchnera aphidicola*, *Rickettsia buchneri*, and *Wolbachia* (Seemann, 2014). Clinker (Gilchrist and Chooi, 2021) with default options was used to compare and visualise the similarity of genes within the biotin operon region of all 6 bacteria. Clinker by default displays the highest similarity comparisons based on an all-vs-all similarity matrix.

All generated draft and complete reference genomes were annotated using the NCBI's Prokaryotic Genome Annotation Pipeline (PGAP) (Tatusova *et al.*, 2016). Secondary metabolite biosynthetic gene clusters were identified using AntiSMASH version 6.0 (Blin *et al.*, 2021) along with Norine (Flissi *et al.*, 2019) which searched for similarities to predicted non-ribosomal peptides. BLASTp analysis was additionally used to identify the closest homologues of these biosynthetic gene clusters.

Functional enrichment analyses between the main *Rickettsia* clade and the Torix – '*Ca.* Megaira' clades were performed using the Anvi'o program anvi-get-enriched-functions-per-pan-group and the "COG_FUNCTION" as annotation source. A gene cluster presence

– absence table was exported using the command “anvi-export-tables”. This was used to create an UpSet plot using the R package ComplexUpset (Krassowski, Arts, and CyrilLagger, 2020) to visualize unique and shared gene clusters between different *Rickettsia* groups. A gene cluster was considered unique to a specified *Rickettsia* group when it was present in at least one genome belonging to that group. Gene cluster accumulation curves were performed for the pan-, core- and unique-genomes based on the same presence-absence matrix using a custom-made R script (Siozios, 2022). In each case the cumulative number of gene clusters were computed based on randomly sampled genomes using 100 permutations. The analysis was performed separately for Torix group and the combined remaining *Rickettsia*. Curves were plotted using the ggplot2 R package (Wickham, 2016).

All information on extra genomes can be found in Appendix B.1, and the code pipeline employed can be found on Zenodo (Davison, 2022).

2.3.11 *Phylogeny, Network, and recombination*

The single-copy core of all 104 genomes was identified in Anvi’o 7 and is made up of 74 single-copy gene (SCG) clusters. Protein alignments from SCG were extracted and concatenated using the command “anvi-get-sequences-for-gene-clusters”. Maximum likelihood phylogeny was constructed in IQ-TREE v2.1.2 (Nguyen *et al.*, 2015). Additionally, 43 ribosomal proteins were identified through Anvi’o 7 to test phylogenomic relationships. These gene clusters were extracted from the pangenome and used for an independent phylogenetic analysis. The best model according to the Bayesian Information Criterion (BIC) was selected with Model Finder Plus (MFP) (Kalyaanamoorthy *et al.*, 2017) as implemented in IQ-TREE; this was JTTDCMut+F+R6 for core gene clusters and JTTDCMut+F+R3 for ribosomal proteins. Both models were run with Ultrafast Bootstrapping (1000 UF bootstraps) (Hoang *et al.*, 2018) with *Orientia tsutsugamushi* as the outgroup.

The taxonomic placement of Oopac6, Ppec13 and Dallo3 genomes within the Rhyzobius, Meloidae and Belli groups respectively were confirmed in a smaller phylogenetic analysis, performed as detailed in (Pilgrim *et al.*, 2021) using reference MLST sequences (gltA, 16S rRNA, 17kDa OMP, COI) from other previously identified *Rickettsia* profiles (Appendix B.2).

The selected models used in the concatenated partition scheme were as follows: 16S rRNA: TIM3e+I+G4; 17Kda OMP: GTR+F+I+G4; COI: TPM3u+F+I+G4; gltA: K3Pu+F+I+G4a.

A nearest neighbour network was produced for core gene sets with default settings in Splitstree4 to further assess distances and relationships between *Rickettsia*, '*Ca.* Megaira' and Torix clades. All annotation was added *post hoc* in Inkscape. Recombination signals were examined by applying the Pairwise Homoplasy Index (PHI) test to the DNA sequence of each core gene cluster extracted with Anvio-7. DNA sequences were aligned with MUSCLE (Edgar, 2004) and PHI scores calculated for each of the 74 core gene cluster with PhiPack (Bruen, Philippe and Bryant, 2006).

The taxonomic identity for genomes was established with GTDB-Tk (Chaumeil *et al.*, 2020) to support the designation of taxa through phylogenetic comparison of marker genes against an online reference database.

2.3.12 Code accessibility

All code and bioinformatics pipelines used to extract and construct bacterial genomes from SRA data can be found here https://doi.org/10.5281/zenodo.6396821, and the R script for generating pangenome accumulation curves can be found on GitHub https://github.com/SioStef/panplots and here https://doi.org/10.5281/zenodo.6408803.

2.3.13 Data accessibility

The genomes and raw read sets generated in this chapter have been deposited in the GenBank database under accession code PRJNA763820. The assemblies produced from previously published third party data have been deposited in the GenBank database under accession code PRJNA767332. The genome content data and data for figures generated in this chapter are provided in the Appendix B.1 and B.2. Accessions and metadata for pre-existing genomic data are listed in Appendix B.1.

2.4 Results and Discussion

We have expanded the available genomic data for several *Rickettsia* groups through a combination of draft and complete genome assembly. This includes an eight-fold increase in available Torix-group genomes, and genomes for previously unsequenced Meloidae and Rhyzobius groups. We further report initial reference genomes for '*Ca.* Megaira'.

2.4.1 *Complete and closed reference genomes for Torix Rickettsia and 'Ca. Megaira'*

The use of long-read sequencing technologies produced complete genomes for two subclades of the Torix group limoniae (RiCimp) and leech (RiClec). Sequencing depth of the *Rickettsia* genomes from *C. impunctatus* (RiCimp) and *C. lectularius* (RiClec) were 18X and 52X respectively. The RiCimp genome provides evidence of plasmids in the Torix group (pRiCimp001 and pRiCimp002) (Table 2.1). Notably, the two plasmids share more similarities between them than to other *Rickettsia* plasmids. However, both plasmids contain distant homologs of the DnaA_N domain-containing proteins previously found in other *Rickettsia* plasmids (Gillespie et al., 2015). In addition, only two components of the type IV conjugative transfer system known as RAGEs (Rickettsiales Amplified Genetic Elements) (Gillespie *et al.*, 2012) were present on the plasmids including homologs of the proteins TrwB/TraD and TraA/MobA. The majority of the RAGE elements including both

the F-like (*tra*) and P-like type IV components have been incorporated in the main chromosome. The presence of RAGE elements, alongside the fact conjugation apparatuses have narrow host-ranges (Pukall, TschÃ¤pe and Smalla, 1996), suggest horizontal transfer of these plasmids is likely within the Rickettsiaceae and could occur between Torix and the main *Rickettsia* clade, considering co-infections of these genera have been noted previously (Yan *et al.*, 2019; Dally *et al.*, 2020). We additionally assembled a complete closed reference genome of '*Ca.* Megaira' from *Mesostigma viride* (MegNEIS296) from previously published genome sequencing efforts. Likewise, MegNEIS296 genome contains a plasmid which bears features of other *Rickettsia* plasmids including the presence of a tra conjugative element and the presence of two DnaA_N-like protein paralogs.

General features of both genomes are consistent with previous genomic studies of the Torix group (Table 2.1). A single full set of rRNAs (16S, 5S and 23S) and a GC content of ~33% was observed. Notably, the two complete Torix group genomes show a distinct lack

Table 2.1. Summary of the closed 'Ca. Megaira' and Torix Rickettsia genomes completed in this project.

Group	'*Ca.* Megaira'	Torix *Rickettsia*	
Strain Name	MegNIES296	RiCimp	RiClec
Symbiont genome accession	GCA_020410825.1	GCA_020410785.1	GCA_020410805.1
Host	*Mesostigma viride* NIES-296	*Culicoides impunctatus*	*Cimex lectularius*
Raw reads accession	SRR8439255, SRX5120346	SRR16018514, SRR16018513	SRR16018512, SRR16018511
Total nucleotides	1,532,409	1,566,468	1,611,726
Chromosome size (bp)	1,448,425	1,469,631	1, 611,726
Plasmids	1 (83,984 bp)	2 (77550bp + 19287bp)	None
GC content (%)	33.9	32.9	32.8
Number of CDS	1,359	1,397	1,544
Avg. CDS length (bp)	998	900	874
Coding density (%)	88.5	86	84
rRNAs	3	3	3
tRNAs	34	34	35

of synteny (Appendix figure A.1), a genomic feature that is compatible with our phylogenetic analyses that placed these two lineages in different subclades (leech/limoniae) (Figure 2.2 and Appendix figure A.2). Gene order breakdown due to intragenomic recombination has been previously associated with the expansion of mobile genetic elements in both *Rickettsia* (Fuxelius *et al.*, 2007) and *Wolbachia* (Comandatore *et al.*, 2015), another member of the Rickettsiales. Both RiCimp and RiClec genomes predicted to encode for a high number of transposable elements with circa 96 and 119 annotated putative transposases respectively. This expansion of transposable elements along with their phylogenetic distance is likely responsible for the extreme synteny breakdown between RiCimp and RiClec.

Of note within the closed reference genomes MegNEIS296 and RiCimp is the presence of a putative non-ribosomal peptide synthetase (NRPS) and a hybrid non-ribosomal peptide/polyketide synthetase (NRPS/PKS) respectively (Appendix figure A.3). Although, the exact products of these putative pathways are uncertain, *in silico* prediction by Norine suggests some similarity with both cytotoxic and antimicrobial peptides, hinting at a potential defensive role (Appendix figure A.3). Further homology comparison with other taxa did not provide links with any specific functions or phenotypes. Previously, an unrelated hybrid NRPS/PKS cluster has been reported in *Rickettsia buchneri* on a mobile genetic element, providing potential routes for horizontal transmission (Hagen et al., 2018). The strongest blastp hits of MegNEIS296 NRPS proteins occur in *Cyanobacteria* (Appendix figure A.3) (Hagen *et al.*, 2018). In addition, putative toxin-antitoxin systems similar to one associated with cytoplasmic incompatibility in *Wolbachia* have recently been observed on the plasmid of *Rickettsia felis* in a parthenogenetic booklouse (Gillespie *et al.*, 2015). Toxin-antitoxin systems are thought to be part of an extensive bacterial mobilome network associated with reproductive parasitism (Gillespie *et al.*, 2018). A BLAST search found a very similar protein in Oopac6 to the putative large pLbAR toxin found in *R. felis* (88% aa identity), and a more distantly related protein in the *C. impunctatus* plasmid (25% aa identity).

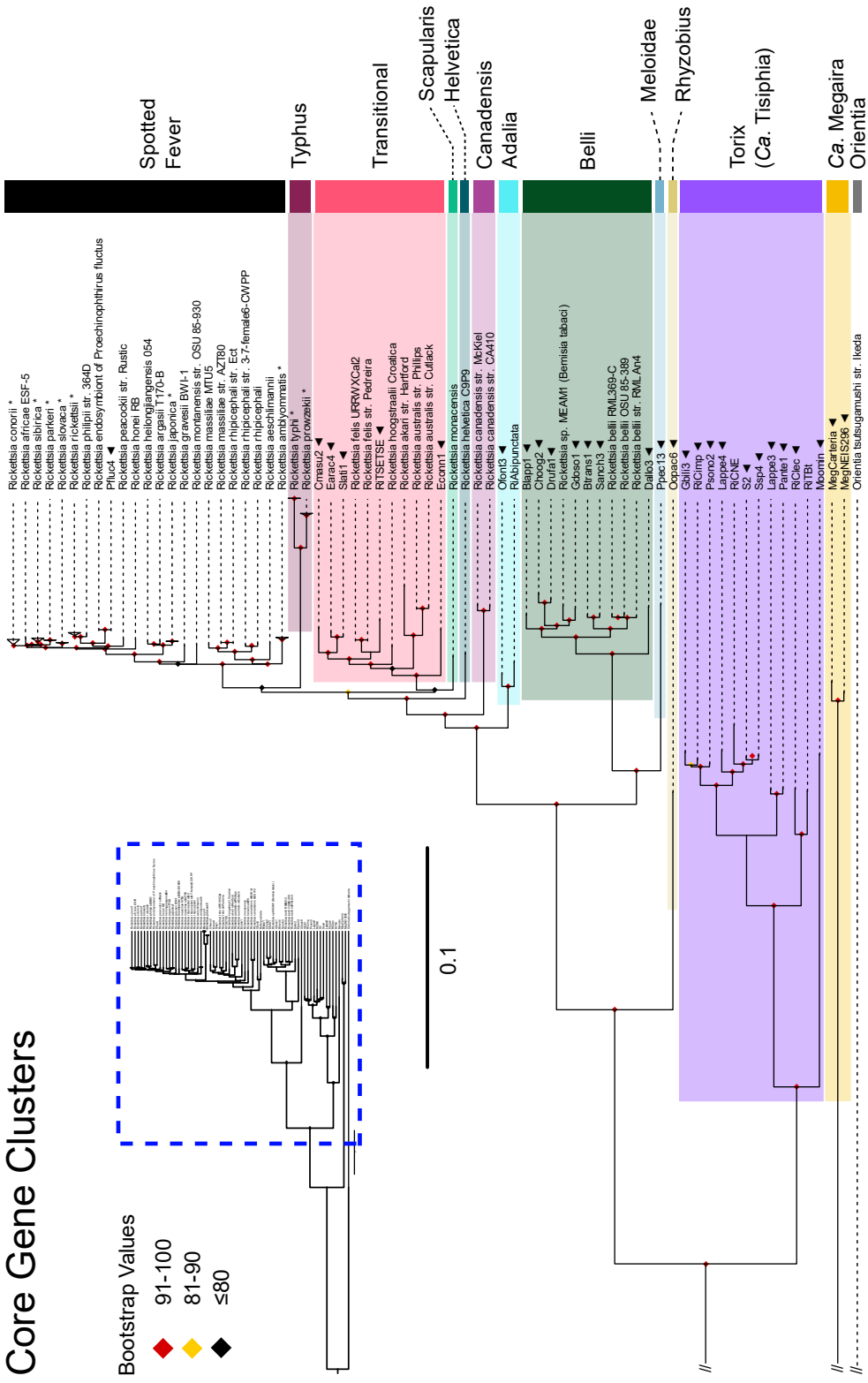


Figure 2.2. Genome wide phylogeny of *Rickettsia* and '*Ca*. Megaira'. Maximum likelihood (ML) phylogeny of *Rickettsia* and '*Ca*. Megaira' constructed from 74 core gene clusters extracted from the pangenome. New genomes are indicated by ◄ and bootstrap values based on 1000 replicates are indicated with coloured diamonds (red = 91-100, yellow = 81-90, black <= 80). New complete genomes are: RiCimp, RiClec and MegNEIS296. Asterisks indicate collapsed monophyletic branches and "//" represent breaks in the branch. Source data are provided in Appendix B.2.

2.4.2 *Sequencing and de novo assembly of other Rickettsia and 'Ca. Megaira' genomes.*

Our direct sequencing efforts enabled assembly of draft genomes for a second '*Ca.* Megaira' strain from the alga *Carteria cerasiformis*, and for *Rickettsia* associated with tsetse flies and *Bryobia* spider mites. The *Rickettsia* genome retrieved from a wild caught tsetse fly, RiTSETSE, is a potentially chimeric assembly of closely related Transitional group *Rickettsia*. We identified an excess of 3584 biallelic sites (including 3369 snps and 215 indels) when the raw Illumina reads were mapped back to the assembly. High read depth indicates that this could be a symbiotic association, reflecting previous observations in tsetse fly cells (Mediannikov *et al.*, 2012) However, there is a possibility that RiTSETSE is not a heritable symbiont, but comes from transient infection from a recent blood meal.

From the SRA accessions, the metagenomic pipeline extracted 29 full symbiont genomes for Rickettsiales across 24 host species. Five of 29 were identified as *Wolbachia* and discarded from further analysis, one was a *Rickettsia* discarded for low quality, and another was a previously assembled Torix *Rickettsia*, RiCNE (Pilgrim *et al.*, 2017). Thus, 22 high quality *Rickettsia* metagenomes were obtained from 21 host species. One beetle (SRR6004191) carried coinfecting *Rickettsia* Lappe3 and Lappe4 (Table 2.2). The high-quality *Rickettsia* covered the Belli, Torix, Transitional, Rhyzobius, Meloidae and Spotted Fever Groups (Table 2.2 and Appendix B.1).

Beetles, particularly rove beetle (*Staphylinidae*) species, appear as a possible hotspot of *Rickettsia* infection. *Rickettsia* has historically been commonly associated with beetles, including ladybird beetles (*Adalia bipunctata*), diving beetles (*Deronectes sp*.) and bark beetles (*Scolytinae*) (Hurst *et al.*, 1994; Perlman, Hunter and Zchori-Fein, 2006; Zchori-Fein, Borad and Harari, 2006; Küchler, Kehl and Dettner, 2009; Weinert *et al.*, 2009). Though a plausible and likely hotspot, this observation needs be approached with caution as this could be an artefact of skewed sampling efforts.

Table 2.2. Summary of draft genomes generated during the current project and their associated hosts. Full metadata including checkM completeness scores and levels of contamination can be found in Appendix B.1.

Strain	Symbiotic bacteria assembly accession	Group	Number of contigs	Total length (bp)	Host name	Host Order
Blapp1	GCA_020404495.1	Belli	171	1266633	*Bembidion lapponicum*	Coleoptera
Btrans1	GCA_020404375.1	Belli	241	1417452	*Bembidion nr. transversale* OSAC:DRMaddison DNA3205	Coleoptera
Choog2	GCA_020404365.1	Belli	16	1357829	*Columbicola hoogstraali*	Phthiraptera
Cmasu2	GCA_020404525.1	Transitional	196	1295004	*Ceroptres masudai*	Hymenoptera
Dallo3	GCA_020404485.1	Belli	196	990679	*Diachasma alloeum*	Hymenoptera
Drufa1	GCA_020404445.1	Belli	14	1364611	*Degeeriella rufa*	Phthiraptera
Earac4	GCA_020881375.1	Transitional	96	1350066	*Ecitomorpha arachnoides*	Coleoptera
Econn1	GCA_020881315.1	Transitional	238	1070326	*Eriopis connexa*	Coleoptera
Gbili3	GCA_020881275.1	Torix limoniae (*'Ca.* Tisiphia*')*	171	1188102	*Gnoriste bilineata*	Diptera
Gdoso1	GCA_020881245.1	Belli	34	1420758	*Graphium doson*	Lepidoptera
Lappe3	GCA_020881125.1	Torix limoniae (*'Ca.* Tisiphia*')*	122	1368980	*Labidopullus appendiculatus*	Coleoptera
Lappe4	GCA_020881075.1	Torix limoniae (*'Ca.* Tisiphia*')*	154	1332357	*Labidopullus appendiculatus*	Coleoptera
MegCarteria	GCA_020881215.1	*'Ca.* Megaira'	72	1298707	*Carteria cerasiformis*	Chlamydom-onadales
Ofont3	GCA_020404465.1	Adalia	91	1529137	*Omalisus fontisbellaquei*	Coleoptera
Oopac6	GCA_020881235.1	Rhyzobius	181	1497231	*Oxypoda opaca*	Coleoptera
Pante1	GCA_020881195.1	Torix limoniae (*'Ca.* Tisiphia*')*	70	1472610	*Pseudomimeciton antennatum*	Coleoptera
Pfluc4	GCA_020404545.1	Spotted Fever	7	1251895	*Proechinophthirus fluctus*	Phthiraptera
Ppec13	GCA_020404425.1	Belli	90	1426047	*Pyrocoelia pectoralis*	Coleoptera
Psono2	GCA_020881175.1	Torix limoniae (*'Ca.* Tisiphia*')*	163	1492063	*Platyusa sonomae*	Coleoptera
RiTSETSE	GCA_020881295.1	Transitional	172	1451997	*Glossina morsitans submorsitans*	Diptera
Sanch3	GCA_020881115.1	Belli	181	1487154	*Stiretrus anchorago*	Hemiptera
Slati1	GCA_020881155.1	Transitional	109	1301763	*Sceptobius lativentris*	Coleoptera
S2	GCA_020404555.1	Torix limoniae (*'Ca.* Tisiphia*')*	103	1251484	*Sericostoma*	Trichoptera
Ssp4	GCA_020404565.1	Torix limoniae (*'Ca.* Tisiphia*')*	87	1231013	*Sericostoma sp.* HW-2014	Trichoptera
Moomin	GCA_020881085.1	Torix moomin (*'Ca.* Tisiphia*')*	204	1137559	*Bryobia graminum*	Trombidi-formes

2.4.3 *Phylogenomic analyses and taxonomic placement of assembled genomes*

The network and phylogeny illustrate the distance of Torix from '*Ca.* Megaira' and other *Rickettsia*, along with an extremely high level of within-group diversity in Torix compared to any other group (Figure 2.2, Figure 2.3 and Appendix figure A.2 and A.4). No significant discordance was detected between the core and ribosomal phylogenies. The phylogenies generated using core genomes are consistent with previously identified *Rickettsia* and host associations using more limited genetic markers (Weinert *et al.*, 2009; Boyd *et al.*, 2016; Guillotte *et al.*, 2021; Pilgrim *et al.*, 2021). For instance, Pfluc4 from *Proechinophthirus fluctus* lice is grouped on the same branch as a previously sequenced *Rickettsia* from a different individual of *P. fluctus* (Boyd *et al.*, 2016). The following groups were identified in the 22 genomes assembled from the SRA screening: 4 Transitional, 1 Spotted Fever, 1 Adalia, 8 Belli and 7 Torix limoniae. Targeted sequences were confirmed as: Torix limoniae (RiCimp), Torix leech (RiClec), Transitional (RiTSETSE), '*Ca.* Megaira' (MegCarteria and MegNEIS296), and a deeply diverging Torix clade provisionally named Moomin (Moomin) (Table 2.2, Figure. 2.2, Appendix figure A.2 and A.4). The extracted Torix genomes include one double infection giving a total of 10 new genomes across 9 potential host species. The double infection is found within the rove beetle *Labidopullus appendiculatus*, forming two distinct lineages, Lappe3 and Lappe4 (Figure 2.2 and Appendix figure A.2).

I also report a putative Rhyzobius group *Rickettsia* genomes extracted from the staphylinid beetle *Oxypoda opaca* (Oopac6) and Meloidae group *Rickettsia* from the firefly *Pyrocoelia pectoralis* (Ppec13). They have high completeness, low contamination, and consistently group away from the other draft and completed genomes (Figure 2.2, Figure 2.3, and Appendix figure A.2). MLST analyses demonstrate that these bacteria are most like the Rhyzobius and Meloidae groups described by Weinert et al. (2009) (Appendix figure A.5). Phylogenies of Oopac6 and Ppec13 suggest that Rhyzobius potentially sits as sister group to all other *Rickettsia* groups, and Meloidae is more closely associated with Belli (Figure 2.2, Appendix figure A.2, A.4 and A.5). Further genome construction will help clarify this taxon and its relationship to the rest of the Rickettsiaceae. The sequencing data for the wasp, *Diachasma alloeum*, used here has previously been described to contain a pseudogenised nuclear insert of *Rickettsia* material, but not a complete *Rickettsia* genome

(Tvedte *et al.*, 2019). The construction of a full, non-pseudogenised genome with higher read depth than the insect contigs, low contamination (0.95%) and high completion (93.13%) suggests that these reads likely represent a viable *Rickettsia* infection in *D. alloeum*. However, these data do not exclude the presence of an additional nuclear insert. It is possible for a whole symbiont genome to be incorporated into the host's DNA like in the case of *Wolbachia* (Hotopp *et al.*, 2007), or the partial inserts of '*Ca.* Megaira' genomes in the *Volvox carteri* genome (Kawafune *et al.*, 2015). The presence of both the insert and symbiont need confirmation through appropriate microscopy methods.

Recombination is low within the core genomes of *Rickettsia* and '*Ca.* Megaira' but may occur between closely related clades that are not investigated here. Across all genomes, the PHI score is significant in 6 of the 74 core gene clusters, suggesting putative recombination events. However, it is reasonable to assume that most of these may be a result of systematic error due to the divergent evolutionary processes at work across *Rickettsia* genomes. Patterns of recombination can occur by chance rather than driven by evolution which cannot be differentiated by current phylogenetic methods (Murray *et al.*, 2016). The function of each respective cluster can be found in Appendix B.2.

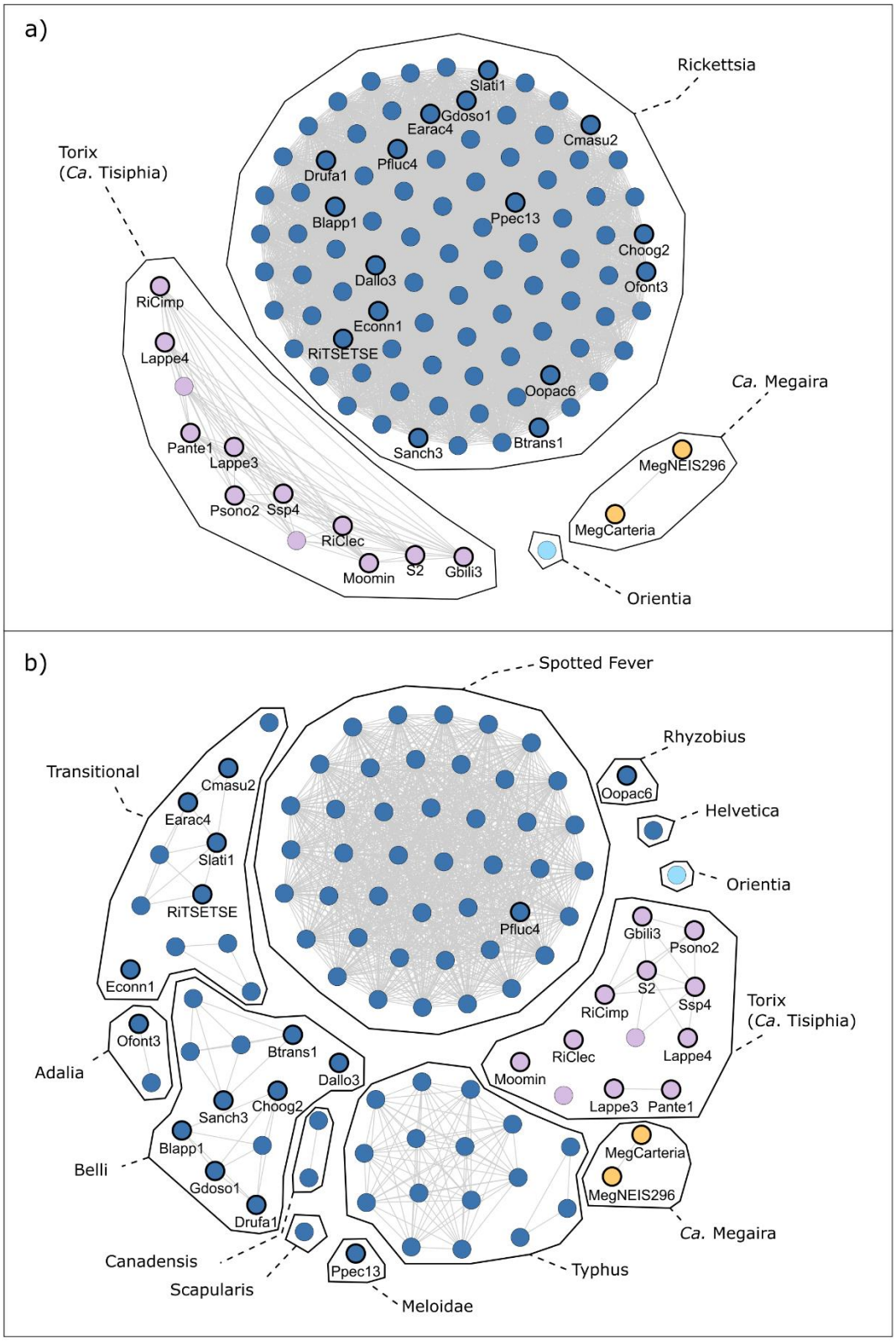


Figure 2.3. Genus and species level clustering across *Rickettsia* and '*Ca*. Megaira'. Frutcherman Reingold networks of pairwise a) Average Amino Acid Identity (AAI) with edge weights >65% similarity and b) Average Nucleotide Identity (ANI) with edge weights >95% similarity across all genomes. AAI and ANI illustrate genus and species boundaries, respectively. The 13 current cluster names are annotated over the 23 species clusters found in the ANI network. New genomes are named and have a thick black outline. Node fill colours indicate *Rickettsia* (Dark blue), '*Ca*. Megaira' (orange), Torix ('*Ca*. Tisiphia', purple), *Orientia* outgroup (light blue). Source data are provided in Appendix B.2.

2.4.4 *Gene content, pangenome and metabolic analysis*

Across all genomes used in the gene content comparison analysis (Appendix B.2 and Appendix figure A.6), Anvi'o identified only 208 core gene clusters of which 74 are represented by single-copy genes. It is particularly evident the large size of the accessory genome across the main *Rickettsia* and the Torix clades. Out of the 2470 predicted ortholog clusters for the Torix clade 1296 (52.5%) are uniquely found among the Torix genomes, while for *Rickettsia* 2460 unique ortholog clusters were predicted from a total of 3811 (64.5%) (Figure 2.4). However, if we account for the number of genomes available in each clade then Torix shows higher rates of gene cluster and unique gene clusters accumulation with each additional genome (Figure 2.5). Our results indicate that the main *Rickettsia* clade and especially the Torix clade, seem to have a high degree of genome diversity, suggesting a wider repertoire of genes and potentially greater rates of gene turnover. As expected, the more genomes that are included in analyses, the smaller the core genome extracted. However, gene content analysis results of increasingly diverged genomes should be always interpreted with caution as true homology relationship between genes/proteins might get obscured by their sequence divergence.

Torix is a distinctly separate clade sharing less than 65% AAI similarity to any *Rickettsia* or '*Ca.* Megaira' genomes (Figure 2.3). It contains at least five species-level clusters with >95% ANI similarity that reflect its highly diverse niche in the environment (Figure 2.3) (Jain *et al.*, 2018; Pilgrim *et al.*, 2021; Rodriguez-R *et al.*, 2021). With only two examples, the true diversity of '*Ca.* Megaira' is underestimated here. Overall, our results indicate higher genomic plasticity within Torix clade in terms of gene content compared to *Rickettsia*.

I also investigated whether Torix and *Rickettsia* clades are enriched for particular COGs (Appendix B.1 and B.2). Among the most highly enriched genes in Torix clade were genes encoding for invasion associated proteins like the exopolysaccharide synthesis protein ExoD (COG3932) and the invasion associated protein IalB (COG5342), a carbonic anhydrase (COG0288) and a Chloramphenicol resistance associated protein (COG3896). Both carbonic anhydrase and ExoD homologs has been already reported in Torix clade (Pilgrim et al. 2017) and our results here further support their important role in Torix biology. ExoD has been previously reported as essential for successful nodule invasion of

the nitrogen-fixing endosymbiont *Rhizobium* (Reed and Walker, 1991). When we consider both Torix and '*Ca.* Megaira' clades the genes involved in the non-oxidative phase of the PPP pathway were the most highly enriched genes (Appendix B.1). It is noteworthy that a large fraction of the enriched genes in both *Rickettsia* and Torix clades are related to cell wall and membrane biogenesis. These are likely associated with differences in the biology of the two clades at the host-microbe interface.

Rickettsia lineages group together based on gene presence/absence and produce repeated patterns of accessory genes that reliably occur within each clade (Appendix figure A.6). AAI scores separate Torix group, *Rickettsia* and '*Ca.* Megaira' into genus groups with no score above 65% similarity outside of each respective clade (Figure 2.3a) (Konstantinidis, Rosselló-Móra and Amann, 2017). ANI scores suggest that Torix and the remaining *Rickettsia* clades are multispecies clusters with less than 95% similarity between genomes in the same groups except for the Spotted Fever Group (Figure 2.3b) (Konstantinidis, Rosselló-Móra and Amann, 2017).

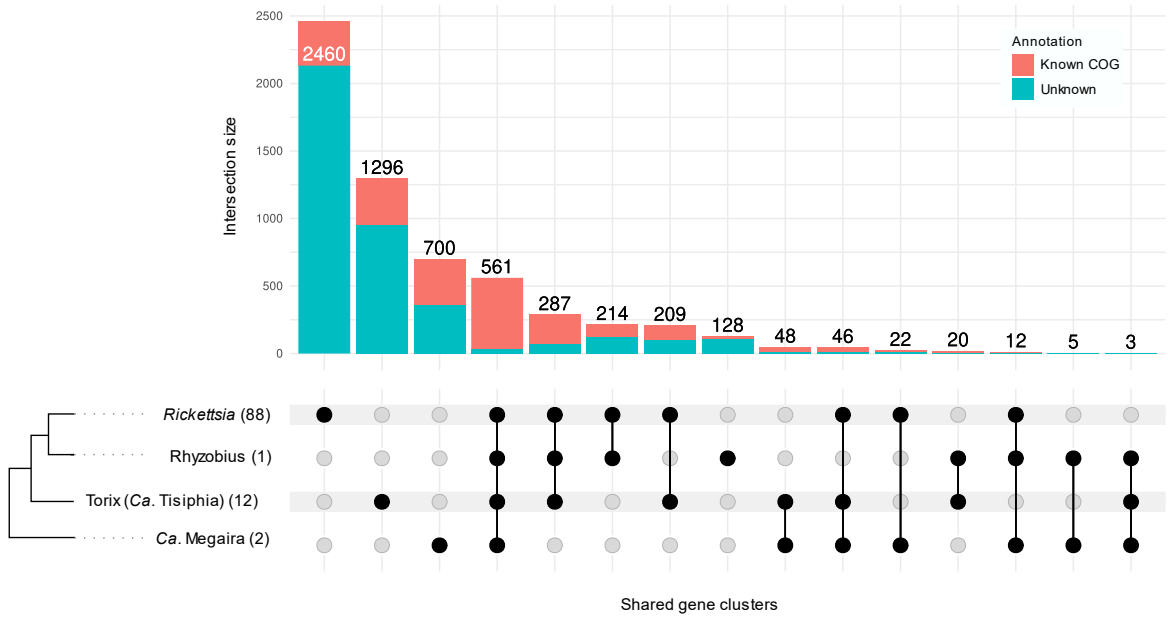


Figure 2.4. Gene content comparison. Shared and unique gene clusters across genus putative genus clusters *Rickettsia*, Rhyzobius, Torix and '*Ca*. Megaira' as suggested by GTDB-tk. Vertical coloured bars represent the size of intersections (the number of shared gene clusters) between genomes in descending order with known COG functions displayed in coral and unknown in blue. Black dots mean the cluster is present and connected dots represent gene clusters that are present across groups. Source data are provided in Appendix B.2.

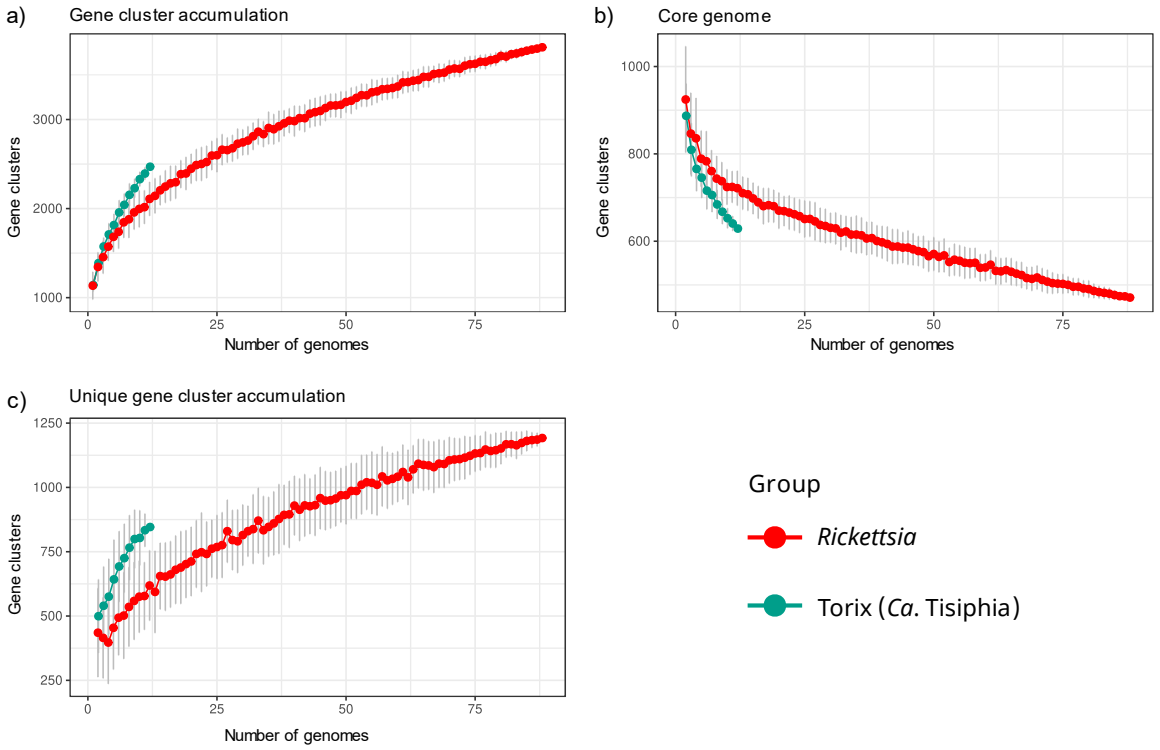


Figure 2.5. Gene cluster accumulation analysis. a) Pangenome accumulation curves, b) core genome accumulation curves and c) the unique genome of *Rickettsia* (red) and Torix (turquoise) clades as a function of the number of genomes sequenced. Each point represents the mean value while error bars represent ± standard deviation based on 100 permutations. Source data are provided in in Appendix B.2.

Rickettsial genomes extracted from SRA samples are generally congruent with the metabolic potential of their respective groups (Figure 2.6). Torix and '*Ca.* Megaira' all have complete pentose phosphate pathways (PPP), a unique marker for these groups which seems to have been lost in the other *Rickettsia* clades (Pilgrim *et al.*, 2017). The PPP generates NADPH, precursors to amino acids, and is known to protect against oxidative injury in some bacteria (Christodoulou *et al.*, 2018), as well as enabling conversion of hexose monosaccharides into pentose used in nucleic acid and exopolysaccharide synthesis. The PPP has also been associated with establishing symbiosis between the *Alphaproteobacteria Sinorhizobium meliloti* and its plant host *Medicago sativa* (Hawkins, Ordonez and Oresnik, 2018). This pathway has previously been highlighted in the Torix group (Pilgrim *et al.*, 2017) and its presence in all newly assembled Torix and '*Ca.* Megaira' draft genomes consolidates its importance as an identifying feature for these groups (Figure 2.6 and Appendix B.2). Considering the trend towards gene loss, the PPP is likely an ancestral feature that was lost in the main *Rickettsia* clade (Driscoll *et al.*, 2017; Pilgrim *et al.*, 2017).

Metabolic pathways for Glycolysis, gluconeogenesis, and cofactor/vitamin synthesis are absent or incomplete across all *Rickettsia* included in these analyses, except in the Rhyzobius group member, Oopac6. Oopac6 has a putatively complete biotin synthesis pathway (Figure 2.6, Appendix figure A.7) and is likely a separate genus according to GTDBtk analysis (Appendix B.1). The Oopac6 biotin synthesis pathway is related to, but distinct from, the *Rickettsia* biotin pathway from *Rickettsia buchneri* (Gillespie *et al.*, 2012) with which it shares between 85% to 92% amino acid sequence similarity across genes (Appendix figure A.7) (Gillespie *et al.*, 2012).

Moreover, there is no sequence similarity outside of the biotin operon. This, along with the presence on a plasmid in *Rickettsia* buchneri makes it likely that Oopac6 operon is a result of horizontal gene transfer. Animals cannot synthesize B-vitamins, so they either acquire them from diet or from microorganisms that can synthesize them. Oopac6 has retained or acquired a complete biotin operon where this operon is absent in other members of the genus. Biotin pathways in insect symbionts are commonly considered to be an indicator of nutritional symbioses (Douglas, 2017), so Rhyzobius *Rickettsia* could contribute to the feeding ecology of the beetle *O. opaca*. However, like other aleocharine rove beetles, *O. opaca* is likely predaceous, omnivorous, or fungivorous (analysis of gut contents from a related species, *O. grandipennis*, revealed a high prevalence of yeasts: (Klimaszewski *et al.*, 2013)).

I posit no obvious reason for how these beetles benefit from harbouring a biotin-producing symbiont. One theory is that this operon could be a hangover from a relatively recent host shift event and may have been functionally important in the original host. Similarly, if the symbiont is undergoing genome degradation, a once useful biotin pathway may be present but not functional (Van Ham *et al.*, 2003; Blow, 2017). BioH, a vital pathway which produces pimeloyl-ACP, is partially present (Figure 2.6) but is not found within the biotin operon (Appendix figure A.7) suggesting that this pathway may not be functional (as observed in some *Buchnera aphidicola* (Van Ham *et al.*, 2003; Manzano-Marín *et al.*, 2020)) or that it may be used in a different way. As this is the only member of this group with a whole genome so far, further research is required to firmly establish the presence and function of this pathway.

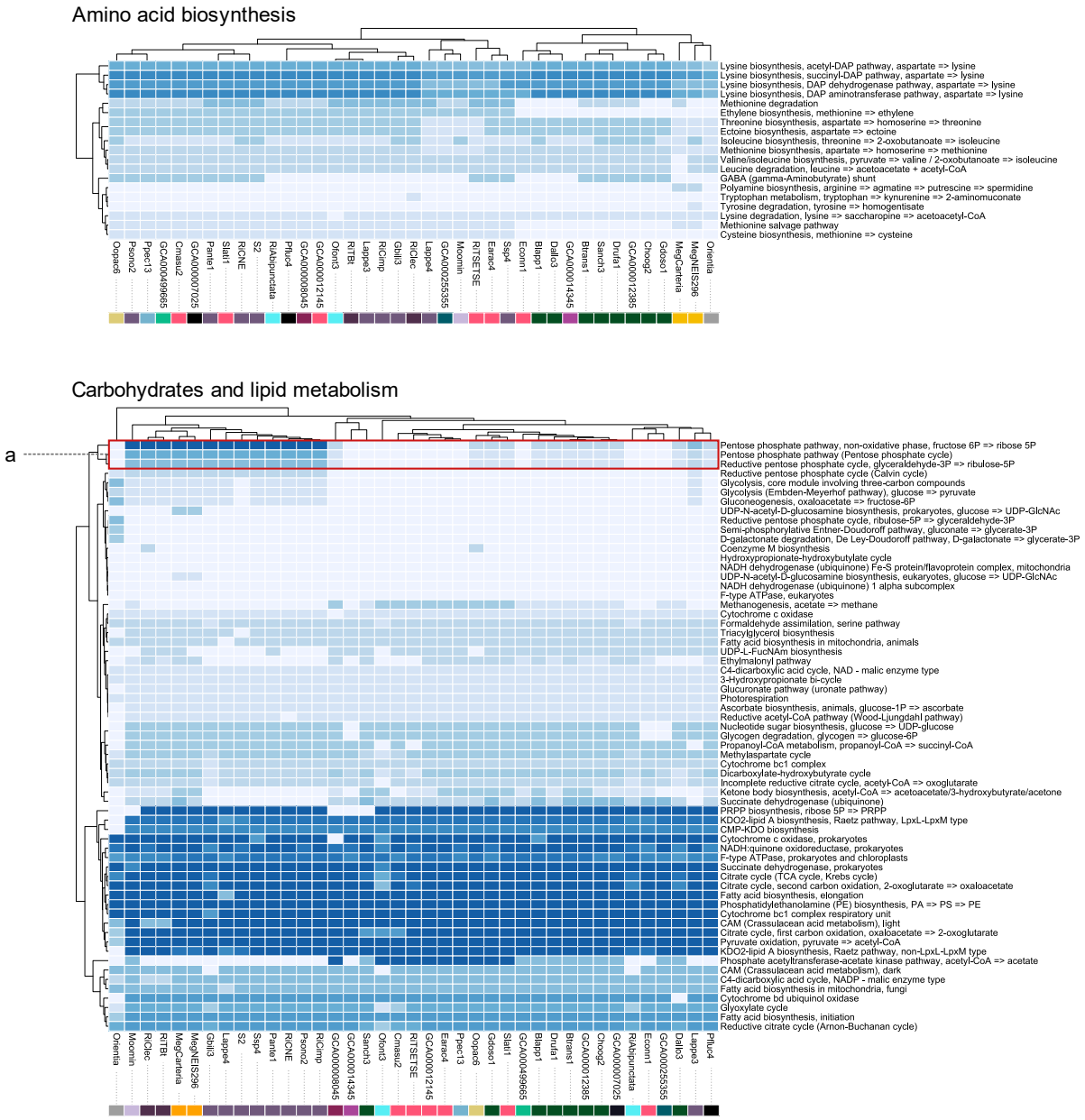


Figure 2.6. Comparison of metabolic potential across selected *Rickettsia* and '*Ca*. Megaira'. Heatmaps of predicted KEGG pathway completion estimated in Anvi'o 7, separated by function and produced with Pheatmap. High to low completeness is coloured dark to light blue. Species groups are indicated with a unique colour as shown in the legend. Pathways of interest are highlighted in red: a) The pentose phosphate pathway only present in Torix and '*Ca*. Megaira', b) the biotin pathway present only in the Rhyzobius *Rickettsia* Oopac6, c) NAD biosynthesis only present in Moomin *Rickettsia*, d) dTDP-L-rhamnose biosynthesis pathway in Gdoso1, Choog2, Drufa1, and Blapp1. SFG is Spotted Fever Group. Source data are provided in Appendix B.2.

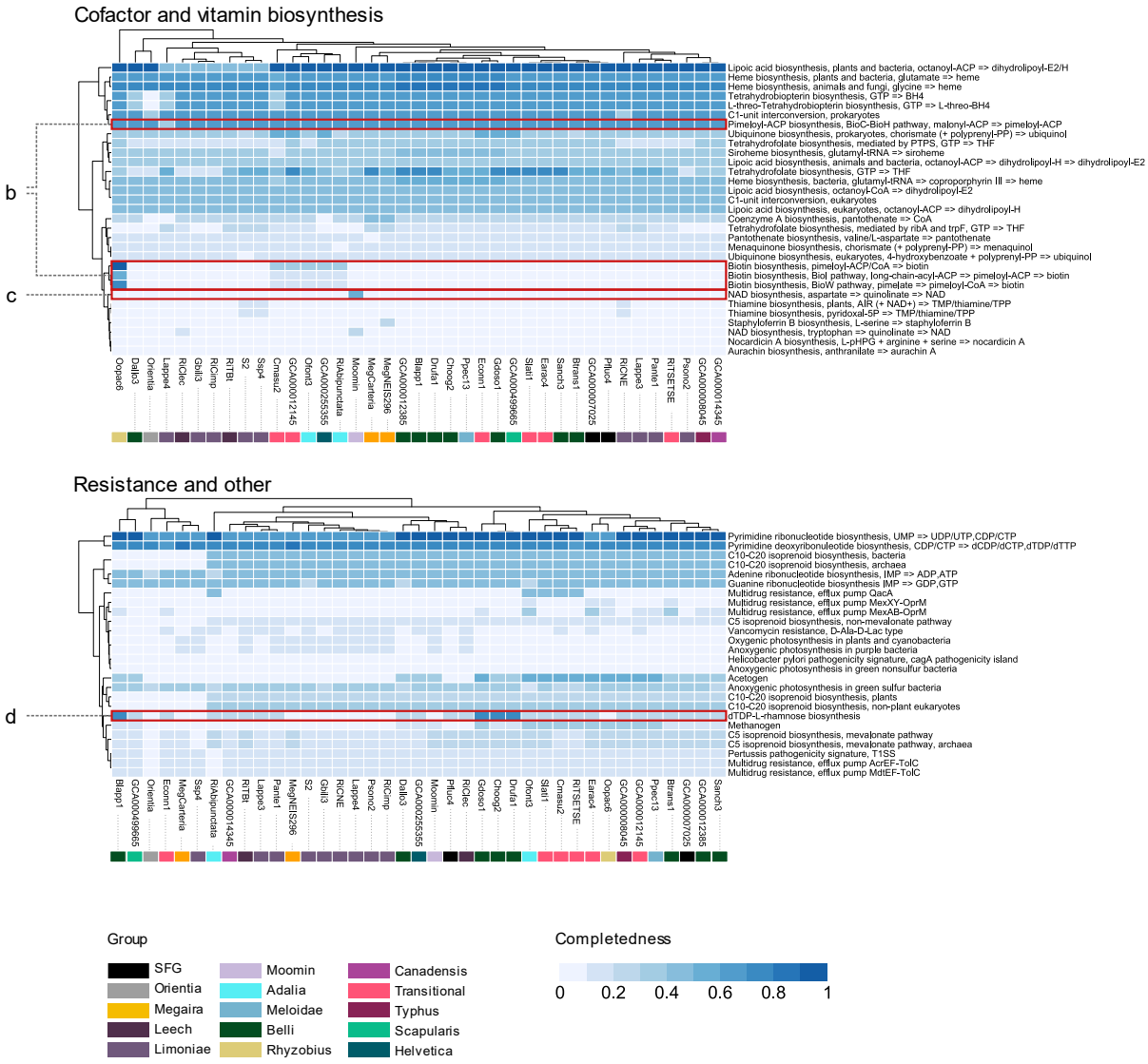


Figure 2.6. (cont.) Comparison of metabolic potential across selected *Rickettsia* and '*Ca*. Megaira'. Heatmaps of predicted KEGG pathway completion estimated in Anvi'o 7, separated by function and produced with Pheatmap. High to low completeness is coloured dark to light blue. Species groups are indicated with a unique colour as shown in the legend. Pathways of interest are highlighted in red: a) The pentose phosphate pathway only present in Torix and '*Ca*. Megaira', b) the biotin pathway present only in the Rhyzobius *Rickettsia* Oopac6, c) NAD biosynthesis only present in Moomin *Rickettsia*, d) dTDP-L-rhamnose biosynthesis pathway in Gdoso1, Choog2, Drufa1, and Blapp1. SFG is Spotted Fever Group. Source data are provided in Appendix B.2.

A 75% complete dTDP-L-rhamnose biosynthesis pathway was observed in four of the draft belli assemblies (Gdoso1, Choog2, Drufa1, Blapp1) (Figure 2.6). Two host species are bird lice (*Columbicola hoogstraali, Degeeriella rufa*), one is a butterfly (*Graphium doson*), and one is a ground beetle (*Bembidion lapponicum*). dTDP-L-rhamnose is an essential component of human pathogenic bacteria like *Pseudomonas, Streptococcus* and *Enterococcus*, where it is used in cell wall construction (van der Beek *et al.*, 2019). This pathway (Jiang *et al.*, 2021) may be involved in the moulting process of *Caenorhabditis elegans* (Feng, Shou and Butcher, 2016), and it is a precursor to rhamnolipids that are used in quorum sensing (Daniels, Vanderleyden and Michiels, 2004). In the root symbiont *Azospirillium*, disruption of this pathway alters root colonisation, lipopolysaccharide structure and exopolysaccharide production (Jofré, Lagares and Mori, 2004). No *Rickettsia* from typically pathogenic groups assessed in Figure 2.6 has this pathway, and the hosts of these four bacteria are not involved with human or mammalian disease. Presence in feather lice provides little opportunity for this *Rickettsia* to be pathogenic to their vertebrate hosts because feather lice are not blood feeders, and Belli group *Rickettsia* are rarely pathogenic. Further, this association does not explain its presence in a butterfly and ground beetle; it is most likely that this pathway, if functional, would be involved in establishing infection in the insect host or host-symbiont recognition.

A partial NAD biosynthesis pathway is present only in the Moomin genome. NAD is used as a coenzyme in numerous reactions as well as a substrate in some synthesis pathways, such as ADP-Ribosyltransferases which are used in bacterial toxin-antitoxin systems (Aravind *et al.*, 2015; Poltronieri and Čerekovic, 2018). NAD pathways have previously found in two other members of Rickettsiaceae, '*Ca.* Sarmatiella mevalonica' and *Occidentia massiliensis* (Mediannikov *et al.*, 2012; Castelli *et al.*, 2021). The most likely explanation for rare occurrence in Rickettsiaceae is either a lateral transfer event or remnants from ancestral occurrence.

2.4.5 *Designation of 'Candidatus Tisiphia'*

In all analyses, Torix group consistently clusters away from the rest of *Rickettsia* as a sister taxon. Despite the relatively small number of Torix genomes, within group diversity is greater than any divergence between previously described *Rickettsia* in any other group (Figure 2.2, Appendix figure. 2 and 4). Additionally, Torix shares characteristics with both '*Ca.* Megaira' and *Rickettsia*, but with many of its own unique features (Figure 2.4 and Figure 2.5). The distance of Torix from other *Rickettsia* and '*Ca.* Megaira' is confirmed in both the phylogenomic and metabolic function analyses to the extent that Torix should be separated from *Rickettsia* and assigned its own genus. This is supported by GTDB-Tk analysis which places all Torix genomes separate from *Rickettsia* (Appendix B.1) alongside AAI percentage similarity scores less than 65% in all cases (Figure 2.3a). To this end, I propose the name '*Candidatus* Tisiphia'. This name follows the fury Tisiphone, reflecting the genus '*Ca.* Megaira' being named after her sister Megaera.

2.4.6 *Conclusions*

The bioinformatics approach has successfully extracted a substantial number of *Rickettsia* and '*Ca.* Megaira' genomes from existing SRA data, including genomes for putative Rhyzobius *Rickettsia* and several '*Ca.* Tisiphia' (formerly Torix group *Rickettsia*). Successful completion of two '*Ca.* Megaira' and two '*Ca.* Tisiphia' genomes provide solid reference points for the evolution of *Rickettsia* and its related groups. From this, I can confirm the presence of a complete Pentose Phosphate Pathway in '*Ca.* Tisiphia' and '*Ca.* Megaira', suggesting that this pathway was lost during *Rickettsia* evolution. I also describe previously unsequenced Meloidae and Rhyzobius *Rickettsia* and show that Rhyzobius group *Rickettsia* has the potential to be a nutritional symbiont due to the presence of a complete biotin pathway. These genomes provide a much-needed expansion of available data for symbiotic *Rickettsia* clades and clarification on the evolution of *Rickettsia* from '*Ca.* Megaira' and '*Ca.* Tisiphia'.

Chapter 3. Rickettsiales in Ciliates and Algae

This chapter is currently in press:

Davison, H.R., Hurst, G.D.D., Siozios, S. (2023). 'Candidatus Megaira' are diverse symbionts of algae and ciliates with the potential for defensive symbiosis, *Microbial Genomics*, DOI: 10.1099/mgen.0.000950

All data collection and analyses in this chapter were performed by me.

3.1 Abstract

Symbiotic microbes from the genus *'Candidatus* Megaira' (Rickettsiales) are known to be common associates of algae and ciliates. However genomic resources for these bacteria are scarce, limiting our understanding of their diversity and biology. I therefore utilized SRA and metagenomic assemblies to explore the diversity of this genus. I successfully extracted four draft *'Ca.* Megaira' genomes including one complete scaffold for a *'Ca.* Megaira' and identified an additional 14 draft genomes from uncategorised environmental Metagenome-Assembled Genomes. I use this information to resolve the phylogeny for the hyper-diverse *'Ca.* Megaira', with hosts broadly spanning ciliates, micro- and macro-algae, and find that the current single genus designation *'Ca.* Megaira' significantly underestimates their diversity. I also evaluate the metabolic potential and diversity of *'Ca.* Megaira' from this new genomic data and find no clear evidence of nutritional symbiosis. In contrast, I hypothesize a potential for defensive symbiosis in *'Ca.* Megaira'. Intriguingly, one symbiont genome revealed a proliferation of ORFs with ankyrin, tetratricopeptide and Leucine rich repeats redolent to that observed in the genus *Wolbachia* where they are considered important for host-symbiont protein-protein interactions. Onward research should investigate the phenotypic interactions between *'Ca.* Megaira' and their various potential hosts, including the economically important *Nemacystus decipiens*, and target acquisition of genomic information to reflect the diversity of this massively variable group.

3.2 Introduction

A wide range of bacteria species reside as endosymbionts in both microeukaryotes, and algae (Hackstein and Vogels, 1997; Vannini *et al.*, 2013; Kawafune *et al.*, 2015; Watanabe *et al.*, 2016; Castelli *et al.*, 2019, 2021; Lanzoni *et al.*, 2019). Symbiont presence can affect the biology of their host in significant ways, from reproductive manipulation (Sonneborn, 1943; Schrallhammer, Castelli and Petroni, 2018) to stress tolerance (Duncan *et al.*, 2010), nutrient production (van Bruggen, Stumm and Vogels, 1983; Du, Maslov and Chang, 1994) and methanogenesis (van Hoek *et al.*, 2000). Symbionts in microeukaryotes were recognised as early as 1902 in the amoeba *Pelomyxa* (Penard, 1902). Whilst some systems are well understood, such as *Caedibacter* and *Paracaedibacter* in *Paramecium* (Schrallhammer, Castelli and Petroni, 2018), our knowledge of symbiont evolution and function in microeukaryotes is fragmentary in comparison to symbioses in animals and terrestrial plants. For instance, the effects of endosymbiotic bacteria in algae are currently unknown, with studies rarely extending beyond the presence of the symbioses and the phylogenetic affiliation of the symbiont (Kochert and Olson, 1970; Nozaki *et al.*, 1989; Kawafune *et al.*, 2015).

In the last decade Rickettsiales have been identified as a group that commonly form symbioses with microeukaryotes as well as invertebrates and algae (Weinert *et al.*, 2009, 2015; Schrallhammer *et al.*, 2013; Schulz *et al.*, 2016; Sabaneyeva *et al.*, 2018; Lanzoni *et al.*, 2019; Castelli *et al.*, 2021; Pilgrim *et al.*, 2021). The origins of some families within the Rickettsiales, like the Rickettsiaceae, may derive from symbioses with microeukaryotes (Schrallhammer *et al.*, 2013). *'Ca.* Megaira' is a member of Rickettsiales and a relative of '*Ca.* Tisiphia', *Rickettsia* and *Wolbachia* which are prolific endosymbionts with wide ranging effects on their hosts (Werren *et al.*, 1994; Stouthamer, Breeuwer and Hurst, 1999; Charlat, Hurst and Merçot, 2003; Duron *et al.*, 2008; Brumin, Kontsedalov and Ghanim, 2011; Hendry, Hunter and Baltrus, 2014). As such, *'Ca.* Megaira' has the potential to impact its hosts in many ways. However, only a single functional study has been completed, which found that *'Ca.* Megaira' presence improved growth in some *Paramecium* (Lanzoni *et al.*, 2019; Pasqualetti *et al.*, 2020). In contrast to *Rickettsia* and

even *'Ca.* Tisiphia', there is currently very limited genomic data for *'Ca.* Megaira*'*, with a single closed and a single draft genome, both from algae[1] (Davison *et al.*, 2022).

The increasing power and reliability of bioinformatic tools now enable us to extract high quality microbial symbiont genomes from the Sequence Read Archive (SRA) deposits (Sangwan, Xia and Gilbert, 2016; Davison *et al.*, 2022). We can search for symbiotic bacteria in hosts without *a priori* hypothesis to establish novel symbiotic interactions with target microbes, and then assemble draft genome sequences for the symbionts. Declining costs have driven a surge in sequencing non-model taxa like microeukaryotes and environmental DNA, providing ample data from which to extract symbiont genomic data. For taxa like *'Ca.* Megaira*'* where there is little genomic information available, this data then provides us with the opportunity to explore their evolution and diversity in more detail and generate hypotheses as to the function of the symbioses found.

In this chapter, I search and extract potential *'Ca.* Megaira*'* symbionts in GenBank SRA data for ciliates and all current classifications of micro- and macro-algae. In addition, I identified *'Ca.* Megaira*'* genomes amongst publicly available Metagenome-Assembled Genomes (MAGs) in GenBank. These data collectively expand the known whole genomes of *'Ca.* Megaira*'* from two to twenty genomes and enable phylogenomic and metabolic analyses.

3.3 Methods

3.3.1 Collection of external genomes for metagenomics and phylogenomics

Illumina SRA data for all ciliates and current classifications of Algae as of 05 May 2021 were downloaded from NCBI to screen for symbiont genomes. These were*: Bacillariophyceae, Charophyceae, Chlorarachniophytes, Chlorophyceae, Chlorophyta, Chrysophyceae, Cillophora, Cryptophyceae, Dictyochophyceae, Dinophyceae, Euglenophyceae, Eustigmatophyceae, Haptophyta, Mesostigmatophyceae, Phaeophyceae, Rhodophyta, Synurophyceae, Ulvophyceae, Ulvophyceae*. Libraries were excluded if they were extremely shallow sequencing efforts below 500 megabases, macronucleus-only sequencing, mutant resequencing, listed as antibiotic treated, or dd-

[1] Both genomes were produced in chapter 2 of this thesis and are published in Davison et al. (2022).

RAD sequence. In total 1113 of 3445 algae and 464 of 547 ciliate libraries were identified for onward analysis.

3.3.2 Metagenomic identification, assembly of genomes and phylogenomic analysis

SRA deposits were screened for the presence of Rickettsiales using Phyloflash (Gruber-Vodicka, Seah and Pruesse, 2020). Rickettsiales-positive libraries were taken forward for metagenomic assembly and binning to extract full genome sequences as described in Davison *et al*. (2022). Briefly, metagenomic assembly, binning and quality check was performed with MEGAHIT, MetaBAT2 and CheckM (Li *et al.*, 2015; Parks *et al.*, 2015; Li, 2018; Kang *et al.*, 2019; Davison *et al.*, 2022). Samples that contained >50% complete symbiont genomes with <5% contamination were taken forward for further examination and manual refinement. GTDBtk (Chaumeil *et al.*, 2020) was used for taxonomic classification of each extracted genome and identify their nearest relatives. Genome bins identified as Rickettsiales were named as follows: first three letters of their closest relative + first letter of host genus + first four letters of host species + bin number. For example, a *'Ca.* Megaira' from a *Mesostigma viride* SRA in bin 4 would be labelled MegMviri4.

Nemacystus decipiens (bioproject PRJDB7493) had multiple SRA libraries from the same biosample which I co-assembled with MEGAHIT. Then, each library was individually mapped back to the assembly with bowtie2 (Langmead and Salzberg, 2012) and symbiont bins were identified with MetaBAT2. Nine of the libraries were mate-pair reads with insert sizes ranges from 2-13 kbp and these were used to scaffold the draft assembly and close the genome using BESST algorithm (Sahlin *et al.*, 2014).

Additional putative *'Ca.* Megaira' genomes were identified on GenBank as following. I performed blastp searches of core '*Ca*. Megaira' proteins from our new draft genomes to identify homologs in the non-redundant protein sequence database using default settings (Camacho *et al.*, 2009). Amongst the top hits were protein sequences from 14 existing but unclassified environmental MAGs. These MAGs were retrieved and their affiliation to '*Ca*. Megaira' was confirmed using the GTDBtk database.

In order to anchor our genomes against previous knowledge of *'Ca.* Megaira' diversity, 16S rRNA sequences were assembled for *'Ca.* Megaira' symbionts where possible. Although, due to the limitations of metagenomic binning and assembly, 16S rRNA retrieval

was not possible for several environmental metagenomes. MegHsini1 is a partial genome and two 16S rRNA sequences can be extracted with Anvi'o 7 (Eren *et al.*, 2021). The most complete of these was used for 16S rRNA sequence placement. The less complete one seems to be related to *Deineraceae* and was deemed a likely contaminant. Additional sequences can be found in Appendix C.1.

The draft genome data were used to enable a phylogenomic approach to *'Ca.* Megaira' diversity alongside existing known *'Ca.* Megaira' genomes (Appendix C.1 and Appendix C.1). Orthologous genes across the 20 *'Ca.* Megaira' genomes were identified using Anvi'o 7 (Eren *et al.*, 2021) for the purpose of extracting the core gene clusters (50 gene clusters). Average Nucleotide Identity (ANI) was calculated through pyANI within Anvi'o 7 (Appendix C.1). Average Amino-acid Identity (AAI) was calculated pairwise for each genome pair through the AAI-Matrix calculator from the enveomics toolbox (Appendix C.1) (Rodriguez-R and Konstantinidis, 2016). Synteny between JAFLDA01 and MegNEIS296 was established with PROmer in MUMmer3 package with default setting (Kurtz *et al.*, 2004). Maximum likelihood trees were produced with IQ-Tree and automatic best model selection using ModelFinder (Kalyaanamoorthy *et al.*, 2017; Minh *et al.*, 2020) with 1000 replicates of UltraFast Boostrap (Hoang *et al.*, 2018) and SH-like Approximate Likelihood Ratio Test (Guindon *et al.*, 2010). Models selected for each tree are as follows: *'Ca.* Megaira' core amino acids = LG+F+I+G4, *'Ca.* Megaira' 16S rRNA = GTR+F+R3. Bayesian phylogenetic inference was performed in Phylobayes-mpi (Lartillot *et al.*, 2013) and the CAT-GTR model. Two independent chains were run in parallel for at least 40.000 cycles each until convergence was observed (maxdiff < 0.1).

3.3.3 Examining metabolic potential, annotation and identifying NRPS systems

High quality genomes from the above were defined as >90% complete and contamination <10%. This process defined 2 existing *'Ca.* Megaira' genomes (MegCarteria, and MegNEIS296), 3 novel genomes derived from the SRA (MegSroe9, MegMviri4, and MegNdeciBESST), and 5 novel genomes derived from MAGs (JAFLDA01, VGEX01, JAJTEJ01, NVVL01, JAFLCZ01) as high quality, and these were analysed alongside a *'Ca.* Tisiphia' genome and *Orientia*. Metabolic potential was predicted based on KEGG annotations by Anvi'o 7 (Eddy, 2018; Eren *et al.*, 2021). Heatmaps of pathway completeness were sorted by phylogeny and plotted in Python with Seaborn (Rossum and Drake, 2009; Waskom and

Seaborn development team, 2020). An upsetplot of shared gene clusters between genomes was constructed with ComplexUpset (Krassowski, Arts, and CyrilLagger, 2020) in R 4.1.0 (R Core Team, 2020).

AntiSMASH (Blin *et al.*, 2021) was then used on the eight high quality genomes to predict secondary metabolites such as those produced by the non-ribosomal peptide synthetase (NRPS) systems. These have been identified previously in the existing *'Ca.* Megaira' genome, MegNEIS296 (ASM2041082v1). Clinker was used to visualise the similarity between the resulting systems found (Gilchrist and Chooi, 2021). Further annotations were made with InterProScan 5 (Jones *et al.*, 2014) using Pfam, TIGRFAM, PANTHER and GOterms.

3.4 Results

3.4.1 Assembly of genomes

After metagenomic binning, four SRA deposits were identified as harbouring *'Ca.* Megaira' and taken forward for further analysis. All but one genome is >90% complete according to checkM results (Table 3.1). MegHsini1 is derived from a single cell genomics approach and was just 62.84% complete and thus not included in onward metabolic analyses; core genes clusters and marker genes were nevertheless retained for phylogenetic placement. No Rickettsiales other than *'Ca.* Megaira' were recovered. The *'Ca.* Megaira' from *Nemacystus decipiens* (PRJDB7493) was the only genome that could be assembled into one scaffold, albeit not closed, using the available mate-pair data. This genome, named here as MegNdeciBESST, has a total size of about 1.3Mb and contains 20 gaps, ranging from 346 to 5679 bp. 14 additional environmental MAGs, previously characterized as unclassified Rickettsiales, were identified in GenBank. These environmental MAGs are of similar quality as the MAGs constructed from SRA databases here (Table 3.1).

Table 3.1. '*Ca*. Megaira' genome statistics and sources. In depth metadata including SRA sample accessions can be found in Appendix C.1.

Name	Bacteria accession	'*Candidatus* Megaira' Clade	Source accession	Host Type	Source	CheckM completion score	CheckM contamination	Genome size	Number of contigs	GC content %	Completion status
Genomes assembled in this chapter											
MegNdeciBESST	SAMN30190846	n/a	PRJDB7493	Algae	Nemacystus decipiens	96.21	0.71	1,273,930	23	31.75	Single scaffold
MegMviri4	SAMN30190847	Clade A	PRJNA517804	Algae	Mesostigma viride	96.21	3.32	1,410,865	28	33.66	Contigs
MegSroe9	SAMN30190848	Clade A	PRJNA507905	Ciliate	Stentor roeselii strain:QDSR01	95.50	1.94	1,258,451	82	33.65	Contigs
MegHsini1	SAMN30190849	n/a	PRJNA546036	Ciliate	Hartmannula sinica	62.84	0.95	702,013	183	28.35	Contigs
Existing unclassified MAGs											
RFMR01	GCA_009927585.1	Clade A	PRJNA495371	Unknown	Freshwater	86.63	3.12	1,145,548	209	33.63	Contigs
RGPV01	GCA_010026065.1	Clade A	PRJNA495371	Unknown	Freshwater	62.76	14.26	2,044,025	736	33.78	Contigs
RGWT01	GCA_010029695.1	Clade A	PRJNA495371	Unknown	Freshwater	54.07	4.55	1,011,306	546	34.82	Contigs
JAFLCZ01	GCA_017302665.1	Clade A	PRJNA704939	Unknown	Activated sludge	98.58	7.11	1,646,433	29	33	Contigs
JAGOTB01	GCA_018062005.1	Clade A	PRJNA524094	Unknown	Wastewater	76.13	6.79	1,346,348	220	33.5	Contigs
JAGWWU01	GCA_018970295.1	Clade A	PRJNA675967	Unknown	Mine drainage	87.68	4.74	1,251,243	73	33.5	Contigs
JAJTEJ01	GCA_021300375.1	Clade A	PRJNA464361	Unknown	Lake water	95.34	2.84	1,300,143	76	33.5	Contigs
VGEX01	GCA_016869095.1	Clade A	PRJNA523022	Unknown	Freshwater	98.58	2.37	1,657,923	74	33	Contigs
JAFLDA01	GCA_017302595.1	Clade E	PRJNA704939	Unknown	Activated sludge	99.53	0.95	1,325,166	3	34.5	Contigs
RFTG01	GCA_009923565.1	Clade A	PRJNA495371	Unknown	Freshwater	68.27	3.28	918,542	365	34	Contigs
NVVL01	GCA_002402195.1	n/a	PRJNA391950	Unknown	Marine	91.07	6	1,905,515	111	40.5	Contigs
JAIELT01	GCA_019752735.1	Clade E	PRJNA745370	Unknown	Drinking water	56.92	1.34	897,267	126	35	Contigs
RXKF01	GCA_003963235.1	n/a	PRJNA490743	Unknown	Freshwater	86.97	4.66	1,361,144	88	30.5	Contigs
JACCWQ01	GCA_013697555.1	Clade E	PRJNA630822	Unknown	Soil	69.18	0.96	895,256	209	34.5	Contigs

3.4.2 Phylogeny and evolution

ANI and AAI scores, alongside phylogenetic analysis, suggest that the whole of *'Ca.* Megaira' genus is deeply divergent (Figures 3.1 to 3.4, and Appendix figure A.8). For instance, AAI scores between Clade A and Clade E are <65% (Figure 3.3) – where Clade refers to previously identified pseudo-species groups based on 16S rRNA phylogeny – and there is no synteny between MegNEIS296 and JAFLDA01, representatives of each group (Appendix figure A.9). The existing *'Ca.* Megaira' clades do not sufficiently describe the diversity seen within the group and our genomic data suggest that the *'Ca.* Megaira' clade groups may represent different genera.

Four of the *'Ca.* Megaira' draft genomes (MegHsini1, MegNdeciBESST, NVVL01, and RXKF01) represent new *'Ca.* Megaira' clades (Figure 3.1 and Figure 3.2.). AAI scores of <65% suggest that these four are sufficiently derived to be considered new genera (Figure 3.3). However, the placement of MegHsini1 within the Rickettsiales is currently uncertain (Figure 3.1 to Figure 3.2., and Appendix figure A.8). For instance, GTDBtk classification does not assign MegHsini1 a genus or species (Appendix C.1). Based on available 16S rRNA and supporting AAI scores, most of the MAGs clustered within Clade A; three MAGs fall into Clade E (and possibly Clade C); and two form a new group within Clade A which share an ANI similarity score of <95% (Figure 3.2. and Figure 3.3). Two MAGs lack 16S rRNA sequence and cannot currently be associated with any group as 16S rRNA is the only marker used to date to classify *'Ca.* Megaira'.

In several instances the genomes used in this chapter are the only ones available for their lineage (Figure 3.1). In addition, MegHsini1 is very incomplete (62%) in comparison to the majority others (11 of 18 are >85% complete, Table 3.1), despite having high depth of coverage (~245X, Appendix C.1). Although a 16S rRNA sequence was also recovered, MegHsini1 also is weakly placed in phylogenetic estimation (Figure 3.1, Figure 3.2, and Appendix figure A.8) and potentially suffers from long branch attraction. At this stage I do not know if MegHsini1's uniqueness is a genuine feature, or a symptom of fragmentation caused by amplification bias during the enrichment steps of single cell genomics. Further expansion of genomic data for *'Ca.* Megaira' is required to refine the phylogeny of the bacteria in this species, and I would recommend any future screening efforts use other

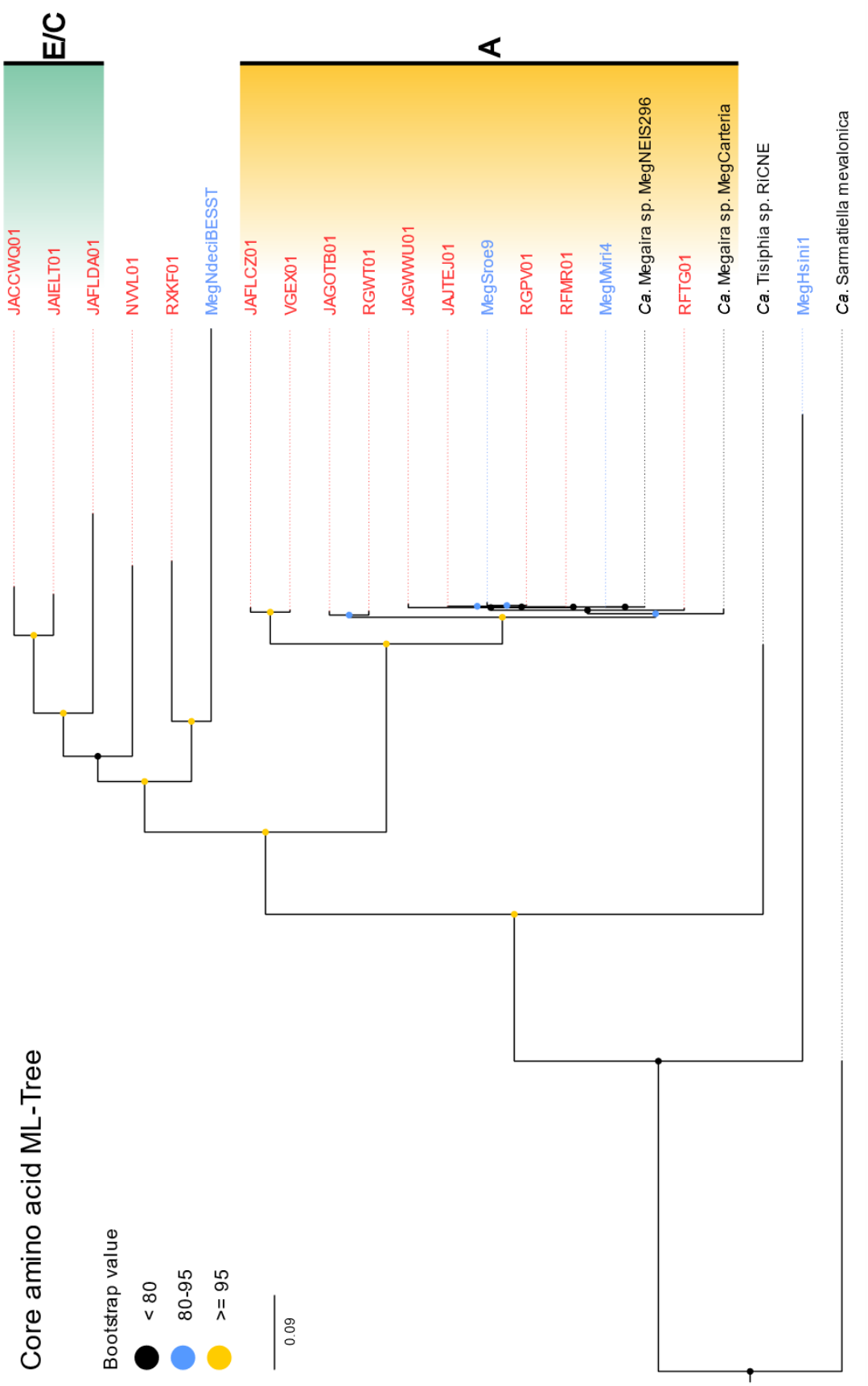


Figure 3.1. '*Ca*. Megaira' core genome maximum likelihood tree 1000 ultrafast bootstrap (UFB). Support for each split is shown as coloured circles, with strong support being >=95. Samples from this chapter are blue and existing environmental metagenomes are red.

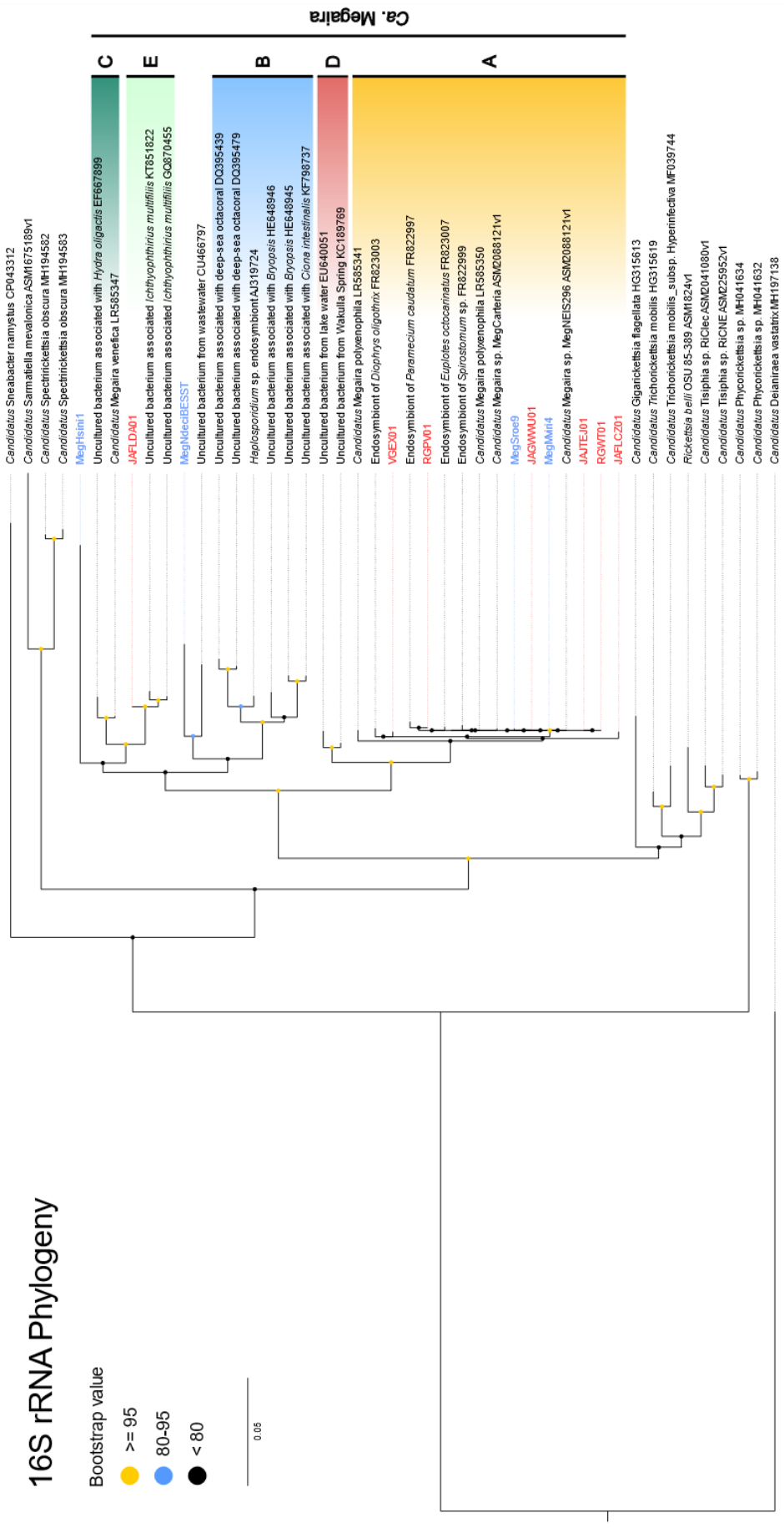


Figure 3.2. '*Ca*. Megaira' 16S rRNA maximum likelihood tree with 1000 ultrafast bootstrap (UFB). Support for each split is shown as coloured circles, with strong support being >=95. Samples from this chapter are blue and existing environmental metagenomes are red. Metadata is in Appendix C.1.

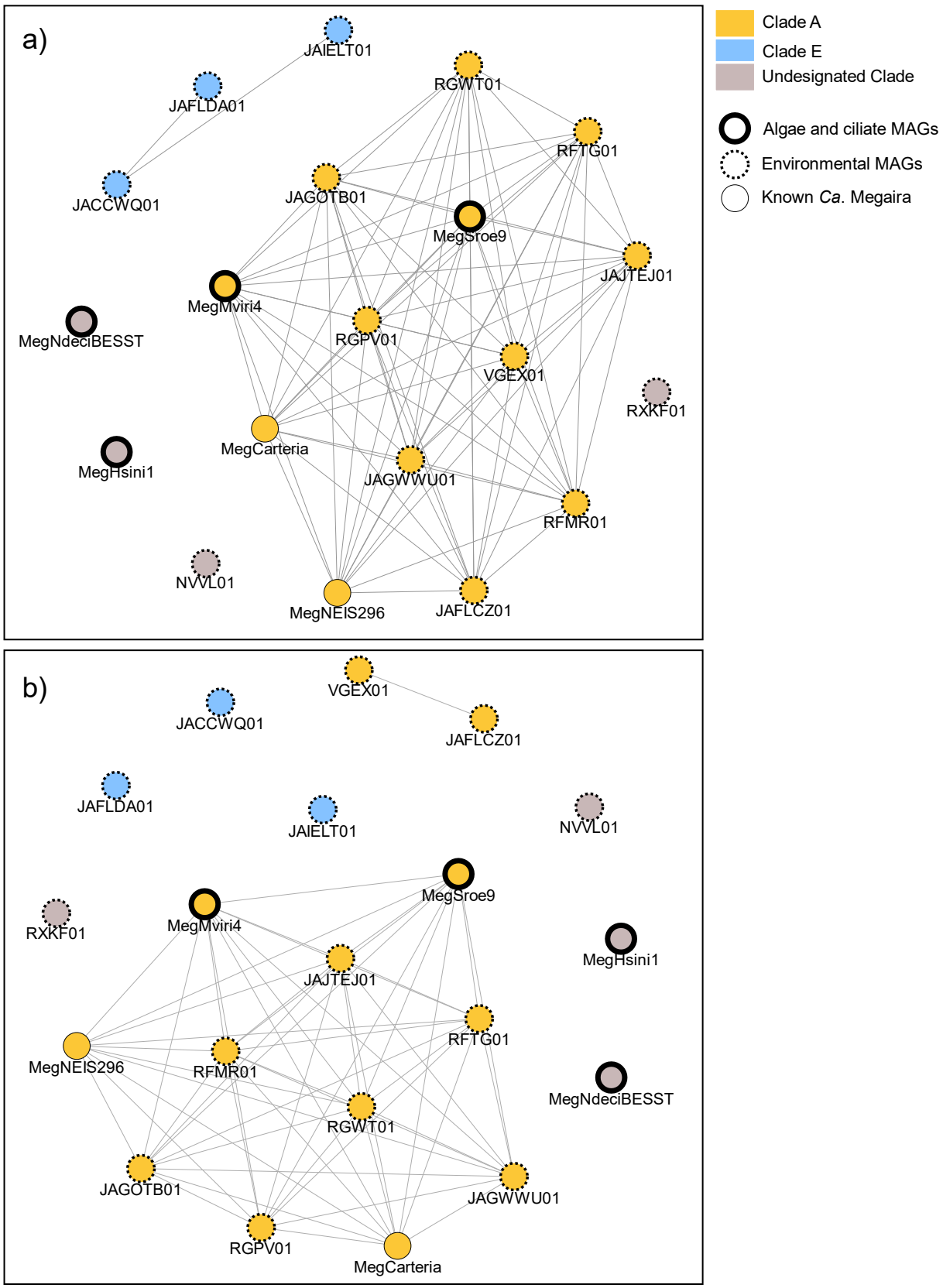


Figure 3.3. AAI and ANI map for '*Ca*. Megaira' showing a) genomes sharing >65% AAI similarity and b) genomes with >95% ANI similarity. Raw data can be found in Appendix C.1.

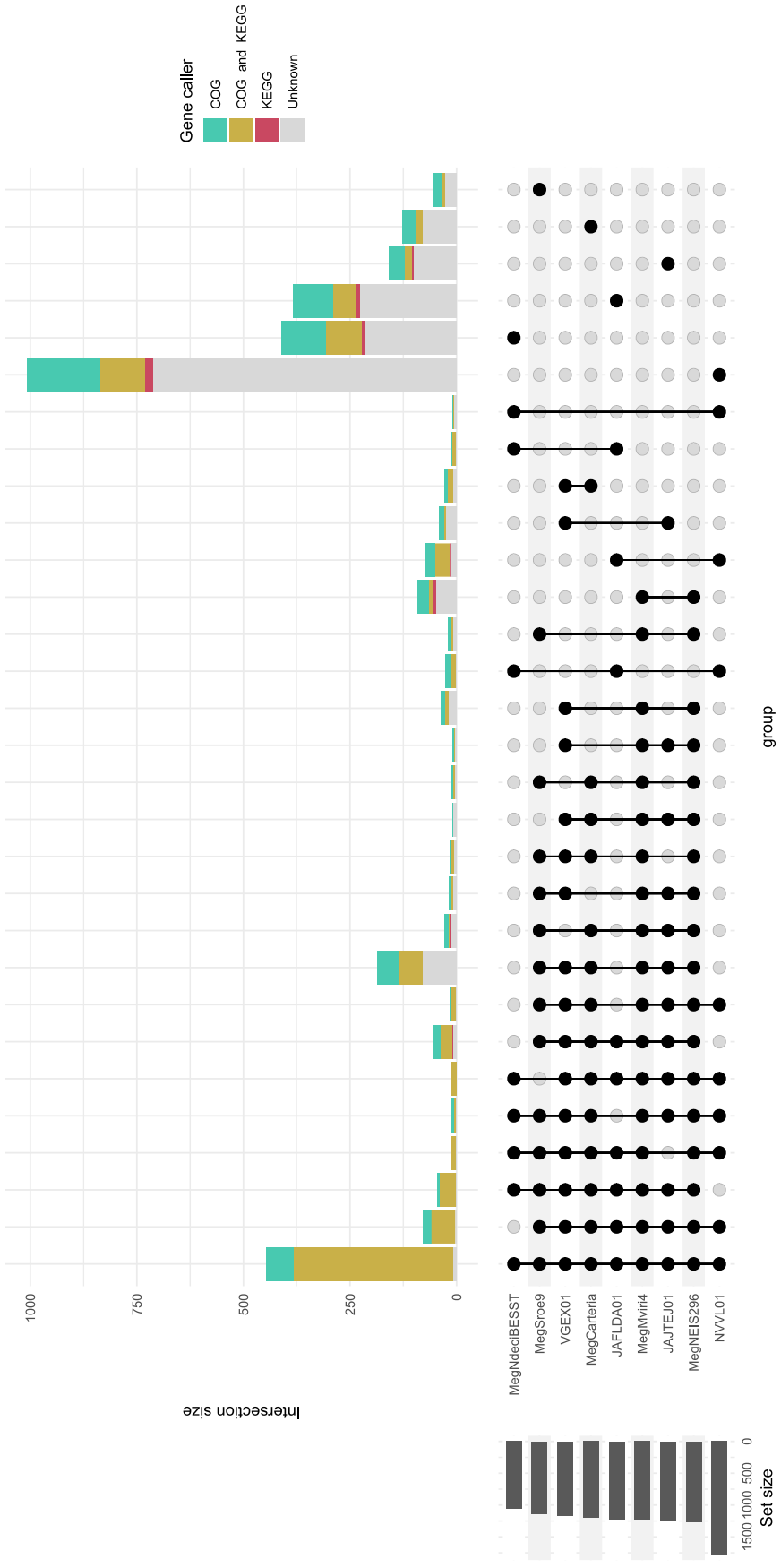


Figure 3.4. Gene content comparison for '*Ca*. Megaira. An upset plot showing the number of gene clusters (bars) shared between genomes ordered by intersection size and degree. Genomes being compared are indicated with black circles and lines. The number of known genes and the caller that identified them are indicated by bar size and colour. Presence-absence data can be found in Appendix C.1.

indicator genes alongside 16S rRNA. The genomic information obtained here will enable development of these markers and PCR protocols.

Gene content analysis across the '*Ca.* Megaira' clades mirror these findings. The A group '*Ca.* Megaira' have a common shared unique gene set and have similar patterns of gene presence absence (Figure 3.4 and Appendix figure A.8). Outside of clade A stains, NVVL01 is highly distinct, having over double the number of unique gene clusters compared to all other taxa; a large number of unique gene clusters were additionally observed in the other two non-A group strains, MegNdeciBESST, and JAFLDA01 (Figure 3.4).

3.4.3 Metabolism, secondary compound synthesis, secretion systems and potential symbiosis factors

'Ca. Megaira' are not predicted to encode complete cofactor or vitamin pathways as would be typically observed in nutritional symbioses (Figure 3.5, Appendix C.1). The genome JAFLDA01 is predicted to encode partial thiamine pathway and NVVL01 a partial biotin biosynthesis pathway, neither of which are not predicted to be functional without external inputs. All *'Ca.* Megaira' are predicted to have complete Non-oxidative Pentose Phosphate Pathways like their relatives, *'Ca.* Tisiphia' (== Torix Group *Rickettsia*) (Davison et al., 2022; Pilgrim et al., 2017). MegNdeciBESST and JAFLDA01 have complete dTDP-L-Rhamnose pathways (Figure 5). Clade A *'Ca.* Megaira', excluding MegCarteria, and clade E *'Ca.* Megaira' appear to be enriched for terpenoid and polyketide biosynthesis pathways compared to other taxa (Figure 3.5 and Figure 3.6).

AntiSMASH identified five putative Non-Ribosomal Peptide Synthetase (NRPS/PKS) systems in four of eight genomes examined (Figure 3.7). It also predicted the presence of three predicted cyclodipeptide synthases (CDPS), and two ribosomally synthesized and post-translationally modified peptides systems (RiPPs), including one synthesizing a lasso peptide (Figure 3.7, Appendix C.2). Blastp found that the MegMviri4 contig containing the putative NRPS has 100% similarity with the NRPS found previously in MegNEIS296, albeit it is only a partial fragment. Considering the highly repetitive structure of the NRPS modules, such systems are poorly assembled with only short reads. I also observed that MegMviri4 and VGEX01 share extremely similar CDPS systems (Figure 3.7). Overall,

Figure 3.5. Heatmap for metabolic pathways of interest in '*Ca*. Megaira'. '*Ca*. Tisiphia', RiCimp and *Orientia tsutsugamushi* are outgroups. Kofam module completeness from highest to lowest is shown with dark to light blue shading and pathways of interest are highlighted and circled with orange. Full metadata and additional pathways can be found in Appendix C.1. Samples from this chapter are blue and existing environmental metagenomes are red.

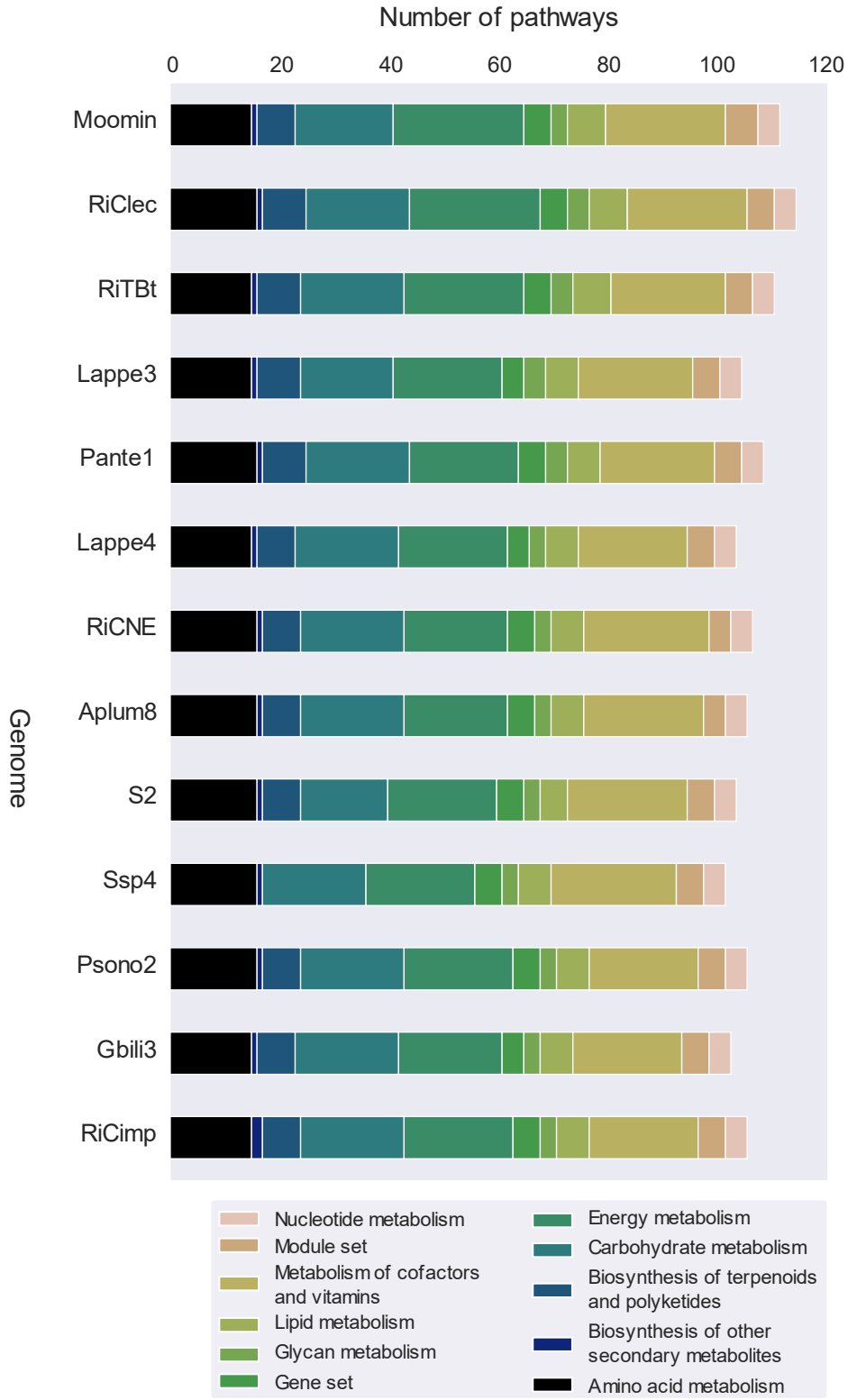


Figure 3.6. KEGG module distribution in '*Ca*. Megaira. The number of pathways found per genome annotated by KEGG module category for '*Ca*. Megaira', with '*Ca*. Tisiphia' RiCimp and *Orientia tsutsugamushi* as outgroups. Full metadata can be found in Appendix C.1. Samples from this chapter are blue and existing environmental metagenomes are red.

according to blastp, the CDPS, NRPS and RiPP systems were most like those found in the two existing *'Ca.* Megaira*'* genomes, MegCarteria and MegNEIS296 (Appendix C.1).

A mostly complete flagellar apparatus was also identified in JAFLDA01 (Appendix C.1 Appendix figure A.10). Partial flagella pathways are also annotated in the genomes NVVL01 and RXKF01 (Appendix C.1). Aside these, *'Ca.* Megaira' strains all carry Sec and Tat systems for translocation of proteins to the periplasmic space, alongside one or more Type IV secretion systems (Appendix C.1).

I examined the *'Ca.* Megaira*'* genomes for ORFs with three classes of motif associated with protein-protein interaction considered important in symbiont-host interactions: ankyrin repeat domains, Tetratricopeptide repeats, and Leucine Rich Repeats. These genes sets were not generally common across *'Ca.* Megaira*'* (Figure 3.7 and Table 3.2). However, the MegNDeciBESST genome was notably enriched, including 15 ORFs carrying ankyrin repeats, 20 with predicted tetratricopeptide repeat motifs and four with leucine rich repeat genes. Two other strains NVVL01 and JAFLCZ01 have modestly increased complement of ORFs in this class (Figure 3.7. and Table 3.2).

Table 3.2. Number of ORFs in '*Ca*. Megaira' genomes containing putative protein-protein interaction domains as recognised in pfam searches.

	ORF feature		
	Ankyrin domains	Tetratricopeptide repeats	Leucine Rich Repeats
MegNdeciBESST	15	20	4
MegSroe9	1	0	1
MegCarteria	1	1	1
MegMviri4	1	3	1
MegNEIS296	1	4	1
NVVL01	9	7	2
JAFLDA01	4	3	0
JAFLCZ01	5	4	7
JAJTEK01	2	3	1
VGEX01	4	4	2

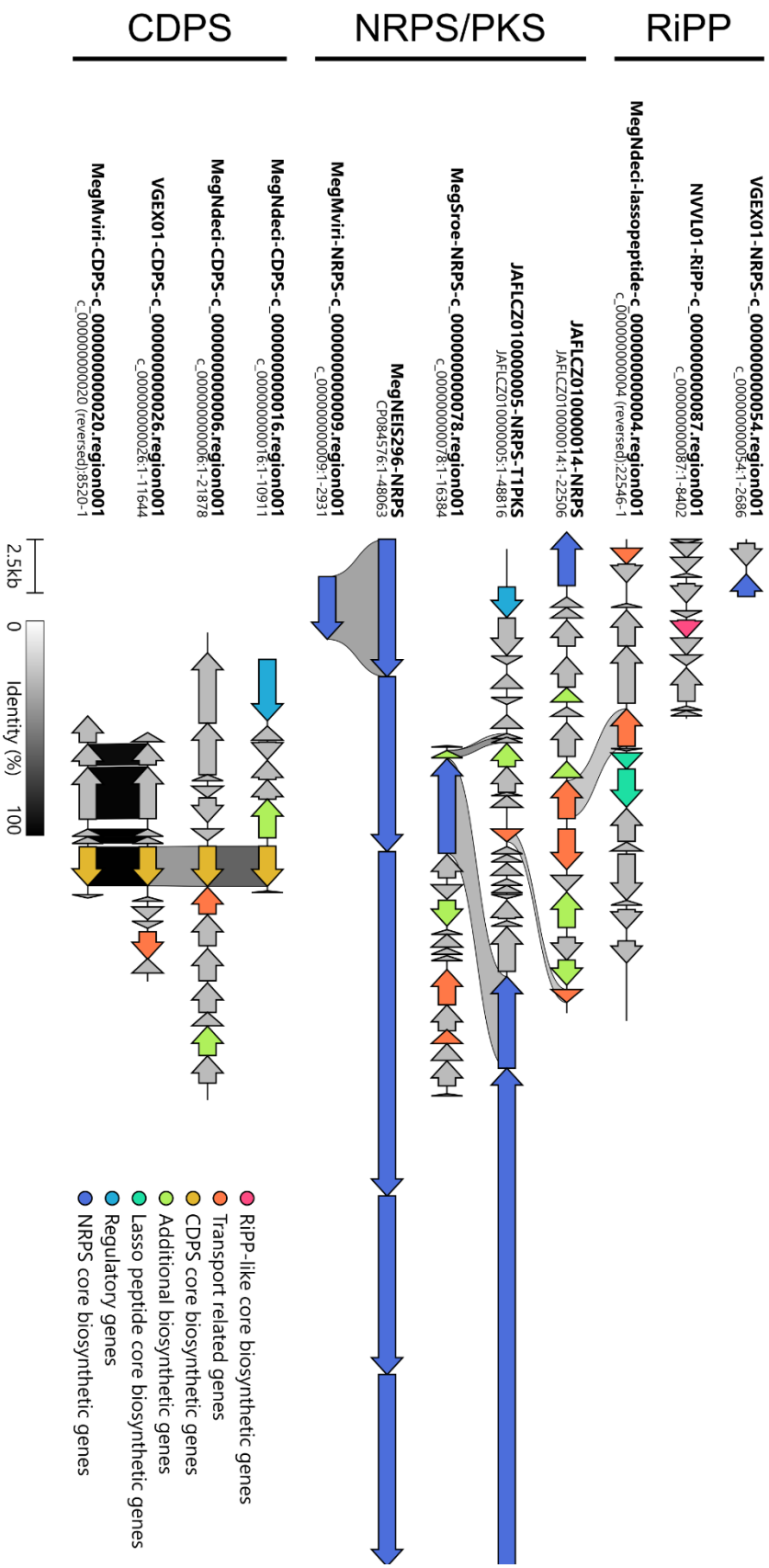


Figure 3.7. Clinker similarity diagram of RiPP, NRPS and CDPS gene regions found across '*Ca*. Megaira' by antiSMASH. Similarities between genes are indicated with grey shaded links between genes, and colours represent the types of genes present as found by antiSMASH. Rows are ordered by best overall similarity according to clinker defaults. A fully interactive clinker diagram with more details on each gene function can be found in Appendix C.2.

3.5 Discussion

Advances in metagenomic methods and data-mining techniques are enriching our understanding of microbial symbiont diversity. The genus *'Ca.* Megaira' represents a common and hyperdiverse clade of intracellular symbionts associated with microeukaryotes and algae. Using a metagenomic approach, I have assembled draft genomes for four *'Ca.* Megaira' species. One of these genomes was assembled into a single scaffold using mate pair reads. In addition, I identified 14 previously existing MAGs in GenBank derived from previous environmental metagenome projects (Tully *et al.*, 2018; Goh *et al.*, 2019; Kantor, Miller and Nelson, 2019; Rodriguez-R *et al.*, 2020; Schneider *et al.*, 2020; Eren *et al.*, 2021; McDaniel *et al.*, 2021; Ortiz *et al.*, 2021; Tran *et al.*, 2021; Vosloo *et al.*, 2021; Chen *et al.*, 2022; Yancey *et al.*, 2022). Of these, five can be considered high quality (>90% complete, <10% contamination).

Our data indicate *'Ca.* Megaira' is diverse enough to be considered its own family within Rickettsiales. The available genomes for previously recognised Clades of *'Ca.* Megaira' share AAI similarity below 65% as well as very low synteny between the two most complete genomes JAFLDA01, *'Ca.* Megaira' Clade E and *'Ca.* Megaira' Clade A from *Mesostigma viride*, (Figure 3.1 to 3.4, and Appendix figure A.9). In addition, NVVL01, while firmly positioned within *'Ca.* Megaira', has an enormous number of unique and unclassified gene clusters that exceed all other genomes described here; this novelty indicates a potentially enormous scope for further genomic diversity within the *'Ca.* Megaira' clades. Our data also indicate a new species group within the current Clade A (Figure 3.1 and Figure 3.3). Overall, the analysis of our current and limited genomic data suggest that *'Ca.* Megaira' lineage consist of at least 6 genus-level clades and 9 species.

Nevertheless, our understanding of *'Ca.* Megaira' genomic diversity remains limited, as I am unable to consolidate the taxonomy for single genomes that fall outside the main clades or that lack 16S rRNA resulting from metagenomic assembly (Yuan *et al.*, 2015). As such, whilst our data indicates taxonomic revision is necessary, I have chosen not to challenge current levels of taxonomic classification to avoid confusion while our knowledge of this family of bacteria is still relatively small. Instead, I encourage future studies to diversify the markers that they use for identifying *'Ca.* Megaira' beyond 16S

rRNA, and to obtain greater genomic information, particularly beyond clade A strains, to allow firm resolution of *'Ca.* Megaira*'* genomic diversity to allow this revision.

All *'Ca.* Megaira*'* genomes obtained have similar predicted metabolic potential which match the two currently available genomes for this group (Figure 3.5). Many algae depend on external sources of biotin, thiamine, and cobalamin, including from bacteria (Tandon, Jin and Huang, 2017). But apart from some partial B vitamin pathways in JAFLDA01 and NVVL01 there is little evidence of capacity for vitamin dependent nutritional symbioses in these bacteria (Figure 3.5 and Appendix C.1). Although the external provision of intermediate metabolites could in theory complement an incomplete pathway, I currently have no evidence that this is the case in *'Ca.* Megaira*'*. Indeed, NVVL01 lacks both bioA and bioD genes which makes the functionality of the whole pathway questionable.

Most Clade A strains encode a large number of proteins related to terpenoid and polyketide pathways (Figure 3.5 and Figure 3.6). These are known to be associated with plant-mycorrhizal and sponge-alphaproteobacteria defensive symbioses (Karimi *et al.*, 2019). Terpenes are also produced by algae for defence systems, and some red algae appear to be reliant on bacteria-like terpene pathways to do so (Wei *et al.*, 2019). Terpenoids and polyketides can also increase host tolerance to various environmental stresses including pathogenic bacteria and heavy metal pollution (French, 2017; Karimi *et al.*, 2019). In addition, MegNdeciBESST, which was recovered from a brown alga genome project, has a complete dTDP-L-Rhamnose biosynthesis pathway which can be associated with establishing symbiosis in plants (Ma, Pan and McNeil, 2002; Jofré, Lagares and Mori, 2004). Therefore, it is possible that *'Ca.* Megaira*'* form a type of defensive symbiosis with their hosts. However, these terpenoids could alternatively be part of establishing infection in the host algae, rather than a defensive symbiosis because bacteria use them to produce components of their cell walls (van der Beek *et al.*, 2019).

The presence of systems predicted to synthesize secondary metabolites (NRPS, CDPS and RiPPs, including a lasso peptide) provide additional evidence that *'Ca.* Megaira*'* could be involved in protective symbiosis, or a toxin-antitoxin system which can be associated with reproductive manipulation (Massey and Newton, 2022). These peptide groups cover a wide variety of bacterial secondary metabolites, many of which are associated with antimicrobial, antifungal, antiviral, and antibiotic properties (Hegemann *et al.*, 2015;

Wenski, Thiengmag and Helfrich, 2022); lasso peptides additionally show very high level of tolerance to environmental extremes of temperature and pH (Hegemann *et al.*, 2015). Alternatively, the products of these systems could be actively harmful to the host as some of these molecules, like the RiPP nostocyclamide, have been shown to have anti-algal properties (Todorova *et al.*, 1995). It is currently unknown if these systems are functional or how the products might affect their hosts. However, they do seem to be common in *'Ca.* Megaira*'* as they are present in six of the eight genomes examined here.

Some intracellular symbionts deploy an array of proteins which interact with host proteins to modify host cellular systems and establish symbiosis. The most widely recognised of these is the expansion of genes carrying ankyrin domains in *Wolbachia* (Siozios *et al.*, 2013; Rice, Sheehan and Newton, 2017). MegNDeciBESST is evolutionary distant from other *'Ca.* Megaira*'* and has a clearly expanded repertoire of genes encoding ankyrin domains, tetratricopeptide domains and leucine rich repeat domains which are associated with protein-protein interaction. This distinction likely makes the molecular basis of its symbioses distinct from that of the other strains. The MegNDeciBESST genome is particularly interesting, as it indicates the expansion of potential effectors functioning through protein-protein interaction that is observed in *Wolbachia* is not unique and has independently evolved in other intracellular symbionts. This aspect of the MegNDeciBESST genome also supports the biological diversity of symbiosis that exists within the current clade *'Ca.* Megaira*'*.

I also found evidence for a complete flagellar apparatus in a clade D *'Ca.* Megaira*'*, JAFLDA01. Although Rickettsiaceae do not typically have flagella, microscopy results suggested the presence of a putative flagellar structure in *'Ca.* Megaira venefica' (Lanzoni *et al.*, 2019), a member of *'Ca.* Megaira*'* clade C. The apparatus is also present in a few related genera such as '*Ca.* Trichorickettsia' and '*Ca.* Gigarickettsia' (Vannini *et al.*, 2014). The presence of flagellar genes in a deep member of the Rickettsiaceae (Martijn *et al.*, 2015) further suggest that a flagellar assembly apparatus might have been an ancestral feature of Rickettsiaceae which was subsequently lost from most of the lineages. I do not know if these pathways are functional, but it is notable that complete or near complete sets of these genes are found in several *'Ca.* Megaira*'* species, while the majority of Rickettsiaceae lack them entirely.

In conclusion, *'Ca.* Megaira' is emerging as a diverse, cosmopolitan clade of bacteria that often form symbioses with a variety of ciliate, micro- and macro-algae. It is commonly found in aquatic metagenomes (Lanzoni *et al.*, 2019) and is likely associated with many other microeukaryotes. I assembled 4 new draft genomes and identify 14 existing environmental MAGs. It is still unclear how these bacteria interact with their hosts, but the presence of partial terpene pathways, alongside the occurrence of various ORFs, NRPS, CDPS and RiPPs across *'Ca.* Megaira' could point towards defensive symbioses. I do not believe that the current taxonomy of *'Ca.* Megaira' sufficiently describes the diversity I observe here. However, further investigation is needed to fully consolidate the identity of genomes lacking 16S rRNA and increasing genome representation to avoid clades being represented by a single genome clade. Once this is complete, the diversity and biology of this hyper diverse group can be established with greater power.

Chapter 4. Novel Chlamydiota diversity emerging from metagenomic data, including algal and ciliate genomes.

4.1 Abstract

Chlamydiota are an ancient and hyperdiverse order of obligate intracellular bacteria. The most well-known representatives are pathogens or parasites of mammals, but it is thought that their primary hosts are microeukaryotes like Amoebozoa. Outside of pathogenic genera like *Chlamydia*, the taxonomy, evolution, and function of Chlamydiota are poorly described. Here I use bioinformatics techniques to extend our current knowledge of Chlamydiota diversity and its hosts, in particular the family Parachlamydiales. I extract one Rhabdochlamydiaceae and three Simkaniaceae genomes from NCBI Short Read Archive deposits of ciliate and algal genome sequencing projects. I then use these to identify a further 14 and 8 genomes respectively amongst existing environmental assemblies. From these data I identify two novel clades with host associated data, for which I propose names '*Candidatus* Sacchlamydia' (Family Rhabdochlamydiaceae*)* and '*Candidatus* Amphrikania' (Family Simkaniaceae), as well as a third novel clade of environmental MAGs '*Candidatus* Acheromydia' (Family Rhabdochlamydiaceae). The extent of uncharacterized diversity within the Rhabdochlamydiaceae and Simkaniaceae is indicated by 16 of the 22 MAGs being evolutionarily distant from currently characterised clades. Within the data, I find huge diversity in Parachlamydiales metabolism and evolution, including the potential for metabolic and defensive symbioses as well as pathogenicity. These data provide an imperative to link genomic diversity in metagenomics data to their associated eukaryotic host, and to develop onward understanding of the functional significance of symbiosis with this hyperdiverse clade.

4.2 Introduction

Chlamydiota is a phylum of bacteria that are highly diverged from any other microbial clade, such that they were for some time considered a separate kingdom of life (Prowazek, 1912; Horn, 2008). Species within the group uniquely share a biphasic lifestyle, with an inert environmental stage and a reproductive intracellular stage (Kahane *et al.*, 1993; Bastidas *et al.*, 2013; König *et al.*, 2017, p. 201). These bacteria also share proteins that

are associated with chloroplast function in plants, which has been speculated to suggest a long-standing relationship with cyanobacteria and algae (Brinkman *et al.*, 2002; Horn *et al.*, 2004). It has also been proposed that ancient members of Chlamydiota facilitated the establishment of the first photosynthetic proto-algal eukaryotes (Huang and Gogarten, 2007).

Chlamydiota are most well known as mammalian pathogens and were first described by the ancient Chinese and Egyptians (Gruber, Lipozenčić and Kehler, 2015). However, the status of the clade has been revised since the advent of DNA taxonomy and it is now recognised that their host range is much broader. They are commonly found in free living amoeba species and environmental samples, though the original host is often not known (Horn, 2008). Besides amoeba, they can also infect mammals, fish, worms, and a variety of arthropods (Draghi *et al.*, 2004; Horn, 2008; Kjeldsen *et al.*, 2010; Halter *et al.*, 2022). There are potentially many hundreds of unexplored families within Chlamydiota (Horn, 2008; Lagkouvardos *et al.*, 2014; König *et al.*, 2017). One order of Chlamydiota, the Parachlamydiales, are widely observed and sequenced in environmental samples, but have little to no functional information and recognised taxonomic affiliations are currently fluid (Horn, 2008; Nylund *et al.*, 2018; Halter *et al.*, 2022).

Very recently, *Parachlamydia acanthamoebae* has been shown to protect its amoeba host from giant viruses (Arthofer *et al.*, 2022). The mechanism for this is not yet known, but *P. acanthamoebae* seems to prevent the virus from co-opting the amoeba machinery, thus preventing the formation of viral factories. These data indicate the potential these symbioses have as hidden players in the biology, ecology, and evolution of microeukaryotes beyond a role as pathogens. Moreover, this creates an imperative to investigate symbioses in microeukaryotes outside of amoeba, which are currently rare.

Sequence data presents an opportunity to estimate the level of uncharacterised diversity in microbial symbionts. Two approaches are possible. First, genome sequencing projects for eukaryotes commonly contain the reads from their associated symbionts within them. Second, environmental sequencing projects allow the identification and assembly of draft genomes without host association data. In this Chapter, I first explore the NCBI Sequence Read Archive (SRA) databases for ciliates, macro- and micro-algae for the presence of Chlamydiota and assemble 1 and 3 new draft genomes for the Parachlamydiales families,

Rhabdochlamydiaceae and Simkaniaceae respectively. I then use these genomes to identify 22 related MAGs from environmental metagenomic studies on online databases. Finally, I apply phylogenetic and metabolic analyses to these, and existing genomes, to better define to the taxonomy these genomes, and to explore potential impacts on their hosts.

4.3 Methods

4.3.1 Data collection

The presence of Chlamydiota bacteria was investigated across Algae and Ciliate SRA databases on NCBI as described previously in Chapter 3 using PhyloFlash to screen for the presence of further Chlamydiota bacteria. Those SRA providing evidence of Chlamydiota presence were taken forward to metagenomic binning and assembly. I used GTDBtk to screen bins and found that all genomes recovered were related to Rhabdochlamydiaceae or Simkaniaceae, so analyses from this point on were focused to Parachlamydiales. Genomes assembled here were then used as baits to identify related metagenomically assembled genomes (MAGs) in the NCBI nr database that were previously not fully classified.

Existing Parachalmydiales genomes, as well as a *Clavichalmydia* and a *Chlamydia* genome were downloaded from NCBI (see Appendix D.1) and the GEM catalogue for genomes identified by Köstlbacher et al. (Köstlbacher *et al.*, 2021; Nayfach *et al.*, 2021).

4.3.2 Metagenomic assembly and annotation

Minimap2, MEGAHIT and MetaBAT2 (Li *et al.*, 2015; Li, 2018; Kang *et al.*, 2019) were used to assemble and bin metagenomic bacterial genomes from SRA data using the pipeline in Davison et al. (2022). Genome quality and identities were checked with CheckM and NCBI Blastn (Camacho *et al.*, 2009; Parks *et al.*, 2015).

Four newly assembled genomes, 18 environmental MAGs, 82 previously established Parachlamydiaceae, and 2 outgroup Chlamydiaceae were then passed to Anvio-7 to be annotated with COG20 and KEGG kofams through their pangenomics pipeline (Aramaki *et al.*, 2020; Eren *et al.*, 2021; Galperin *et al.*, 2021; Davison, 2022).

4.3.3 Phlyogenomics and metabolic predictions

Amino acid gene clusters found in 90% of the 106 genomes were identified and extracted through Anvio-7. I extracted 34 single copy gene clusters that contain a total of 3604 genes. A phylogeny partitioned by gene cluster was constructed with IQTREE, using model finder plus, 1000 ultrafast bootstraps, and 1000 SH-alrt replicates. Models per partition can be found in Appendix D.2.

GTDBtk was used to validate the need for new genera, as well as identify family level taxonomy for each genome. AAI was calculated for all genomes through kostalabs online enveomics suite (Rodriguez-R and Konstantinidis, 2016). ANIb was calculated with pyANI through anvio-7 (Appendix D.1) (Pritchard *et al.*, 2016). The network was visualised in Gephi 0.9 (Bastian, Heymann and Jacomy, 2009), with edges filtered for >0.6 AAI and >0.95 ANI scores. Annotations were added in Inkscape (Inkscape Project, 2020).

Metabolic completeness was assessed through Anvio-7 using KEGG kofams for high quality genomes exceeding a CheckM score of 90% completeness. Heatmaps were constructed with Seaborn 0.12.1 in Python 3.10 and annotated with Inkscape (Rossum and Drake, 2009; Waskom and Seaborn development team, 2020). Toxin-antitoxin systems were also identified with antiSMASH (Blin *et al.*, 2021). CrisprCAS finder (Couvin *et al.*, 2018) was used to establish the presence of cas systems in all extracted genomes.

4.4 Results and Discussion

4.4.1 Genomes

Four genomes were assembled from SRA data and additional 22 recovered from the NCBI GenBank non-redundant database (Table 4.1).

Table 4.1. Select Chlamydiota genome metadata for newly assembled Rhabdochlamydiaceae and Simkaniaceae genomes and previously assembled NCBI environmental MAGs. Full metadata can be found in Appendix D.1.

Name	Bacteria accession	Clade	Source accession	Host Type	Source	CheckM completion score	CheckM contamination	Genome size	Number of contigs	GC content %	Completion status
Genomes assembled in this chapter											
SimiAspAT52	SAMN32091777	Simkaniaceae	SRR3080743	Algae	*Amoebophrya* sp. AT5.2	91.55	0	1,800,129	71	36.81	contig
SimkCsp1	SAMN32091778	Simkaniaceae	SRR9841583	Ciliate	*Chilodochona* sp.	90.65	0.23	1,516,029	217	39.52	contig
RhabSjap6	SAMN32091779	Rhabdochlamydiaceae	SRR2043156	Algae	*Saccharina japonica*	91.05	0.9	1,734,283	4	45.53	contig
SimkSjap42	SAMN32091780	Simkaniaceae	SRR2043156	Algae	*Saccharina japonica*	65.79	1.73	972,015	316	43.90	contig
Existing MAGs											
JAAKFB01	GCA_011065235.1	Rhabdochlamydiaceae	PRJNA504765	Unknown	marine sediment	82.56	2.73	1,738,011	252	44.0	contig
JAAKFC01	GCA_011065225.1	Rhabdochlamydiaceae	PRJNA504765	Unknown	marine sediment	78.01	2.03	1,662,393	270	41.5	contig
JAAKFF01	GCA_011065165.1	Simkaniaceae	PRJNA504765	Unknown	marine sediment	90.2	0	1,625,346	143	46.5	contig
JAAKFO01	GCA_011064755.1	Rhabdochlamydiaceae	PRJNA504765	Unknown	marine sediment	82.02	3.6	1,505,120	432	42.5	contig
JAAKFS01	GCA_011064965.1	Rhabdochlamydiaceae	PRJNA504765	Unknown	marine sediment	92.91	1.35	1,614,069	25	46.5	contig
JAAKFU01	GCA_011064915.1	Rhabdochlamydiaceae	PRJNA504765	Unknown	marine sediment	84.12	2.25	1,677,468	185	42.5	contig
JACPOL01	GCA_016185065.1	Rhabdochlamydiaceae	PRJNA640378	Unknown	groundwater	95.61	0.68	1,865,878	10	48.5	contig
JACPWB01	GCA_016197265.1	Rhabdochlamydiaceae	PRJNA640378	Unknown	groundwater	76.96	5.69	1,071,557	362	46.5	scaffold
JACRBE01	GCA_016213235.1	Simkaniaceae	PRJNA640378	Unknown	groundwater	82.09	1.04	1,468,519	59	36.0	contig
JAGOLG01	GCA_017994195.1	Rhabdochlamydiaceae	PRJNA524094	Unknown	wastewater	89.02	1.49	1,679,597	130	42.5	contig
JAGROF01	GCA_020440405.1	Simkaniaceae	PRJNA432264	Unknown	Saline, activated sludge	80.23	0.34	1,731,737	299	44.5	scaffold
JAGROI01	GCA_020440305.1	Simkaniaceae	PRJNA432264	Unknown	saline, activated sludge	96.28	0.56	2,560,732	197	39.0	scaffold
JAGROK01	GCA_020440285.1	Simkaniaceae	PRJNA432264	Unknown	Saline, activated sludge	88.85	1.52	1,696,508	180	42.5	contig
JAGROM01	GCA_020440265.1	Simkaniaceae	PRJNA432264	Unknown	Saline, activated sludge	92.44	5.63	2,391,591	302	41.5	Scaffold
JAGXTH01	GCA_018333475.1	Rhabdochlamydiaceae	PRJNA672823	Unknown	groundwater	84.46	4.05	1,945,191	100	42.0	Contig
JAHCAX01	GCA_019634675.1	Rhabdochlamydiaceae	PRJNA725625	Unknown	activated sludge	96.96	0.68	2,665,416	21	46.5	Contig
JAIELY01	GCA_019752555.1	Rhabdochlamydiaceae	PRJNA745370	Unknown	drinking water	95.61	0.68	2,067,720	42	44.0	contig
JAIETG01	GCA_019745395.1	Rhabdochlamydiaceae	PRJNA745370	Unknown	drinking water	93.69	1.35	2,621,400	37	39.5	Contig
JAJACU01	GCA_022601575.1	Simkaniaceae	PRJNA742377	Unknown	sponge, *Haliclona oculata*	95.95	0	1,780,284	3	45.0	Scaffold
JAJACS01	GCA_022601395.1	Simkaniaceae	PRJNA742377	Unknown	sponge, *Haliclona oculata*	94.93	0.68	1,870,253	48	38.5	contig
MGLO01	GCA_001796065.1	Rhabdochlamydiaceae	PRJNA288027	Unknown	groundwater	84.46	0.17	1,177,368	32	49.0	Scaffold
RGUS01	GCA_010027655.1	Rhabdochlamydiaceae	PRJNA495371	Unknown	Freshwater	91.44	13.91	2,063,700	445	38.0	contig

4.4.2 Phylogeny

I observe high diversity in the currently named families, genera, and even species of Chlamydiota, in line with other studies (Horn, 2008; Lagkouvardos *et al.*, 2014; König *et al.*, 2017). The current taxonomy is not sufficient and has been left undefined to the detriment of these bacteria. This has resulted in some genomes such as *Rubidus massiliensis* being apparently miscategorised as a Chlamydiales instead of *Parachlamydiales* (Figure 4.1 and Figure 4.2). AAI and ANI results shown in Figure 4.2 support most existing genera and species clusters, but also indicates the potential need for some genera to be split (e.g. *Simkania*) and some species to be taxonomically combined (*Ca*. Rhabdochlamydia oedothora, '*Ca*. Rhabdochlamydia' sp. W815 RHOW815).

Using AAI values with a cut off of 65%, I find objective support for naming the clade containing MGLO01 as its own genus, '*Ca*. Arenachlamydia', as suggested in (Pillonel, Bertelli and Greub, 2018). However, contrary to Pillonel et al. (2018), I do not find that they are unique within Chlamydiota in not possessing a Menaquinone biosynthesis/futalosine pathway (Appendix D.1). At this stage I do not believe splitting the family Rhabdochlamydiaceae into multiple families is helpful. No other environmental MAGs used in this chapter has to our knowledge been assigned a taxonomic classification lower than family.

Several groups, like SM23-39 and most MCF types, are sufficiently different to warrant their own family name. SM23-39 has been referred to as '*Ca*. Limichlamydiaceae' or '*Ca*. Anoxychlamydiales' (Pillonel, Bertelli and Greub, 2018; Dharamshi *et al.*, 2020). Based on GTDBtk analysis I have opted for using '*Ca*. Limichlamydiaceae' as all genomes share family level classification by this method. Higher classifications can be added later as needed. There is no set threshold for classifying family or order level, and although GTDBtk does overclassify (Chaumeil *et al.*, 2020), it is currently the only objective way of designating family level classification.

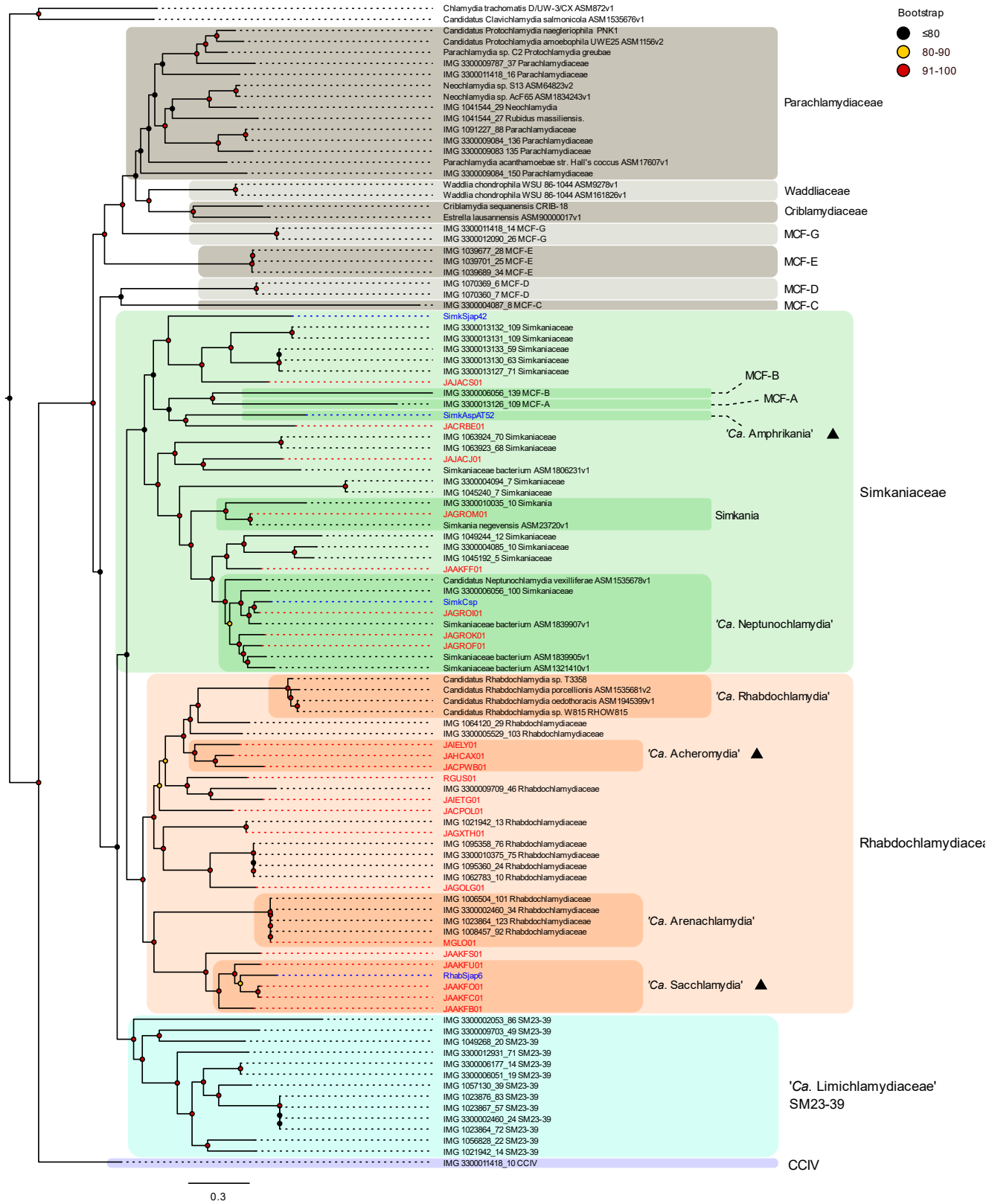


Figure 4.1 Genome wide phylogeny of Parachlamydiales. Maximum likelihood (ML) phylogeny of Parachlamydiales constructed from 34 single copy gene clusters that contain a total of 3604 genes. New genomes are indicated by ▲ and bootstrap values based on 1000 replicates are indicated with coloured circles (red = 91-100, yellow = 81-90, black <= 80).

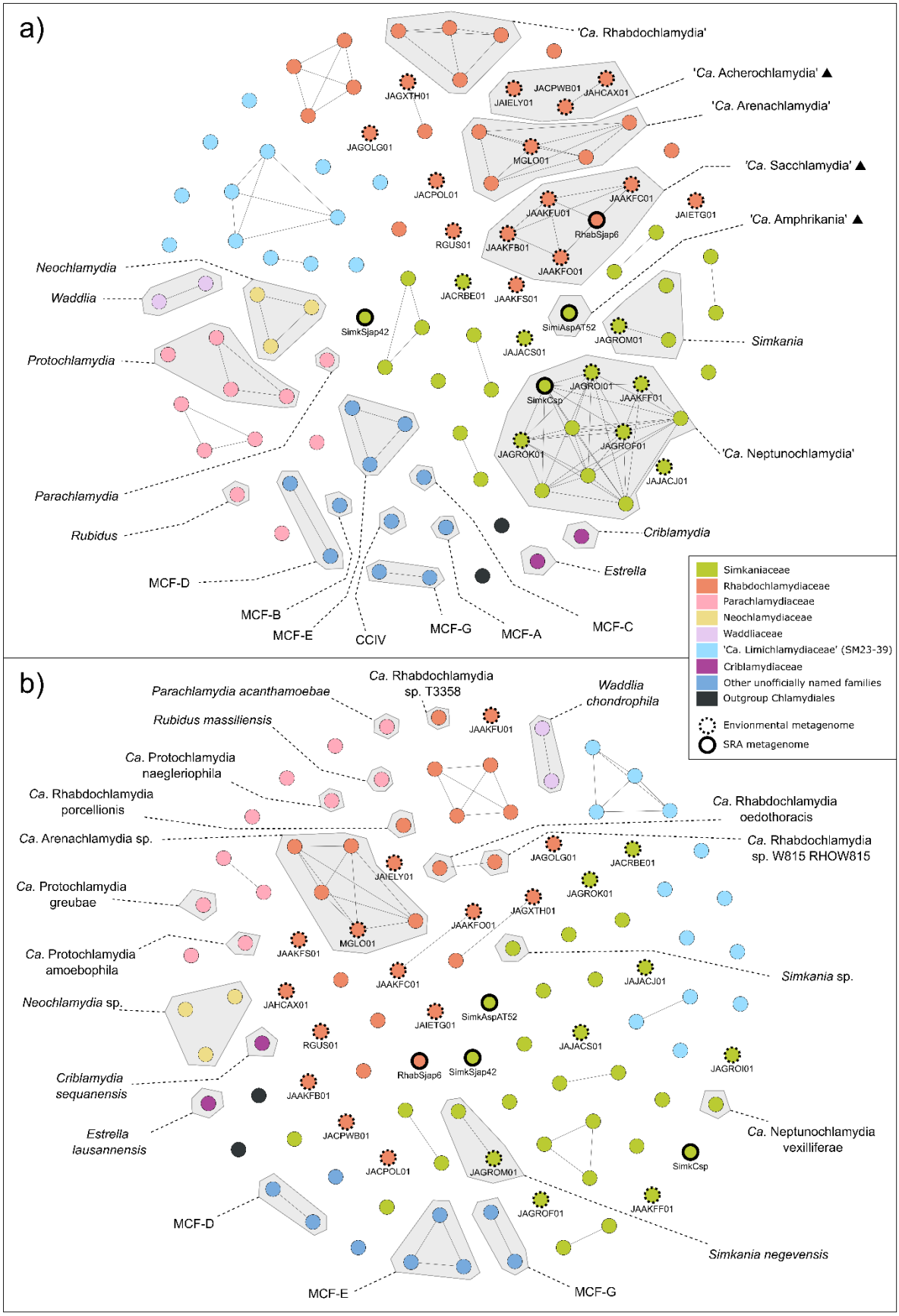


Figure 4.2. Genus and species level clustering across Parachlamydiales. Frutcherman Reingold networks of pairwise a) Average Amino Acid Identity (AAI) with edge weights >65% similarity and b) Average Nucleotide Identity (ANI) with edge weights >95% similarity across all genomes. AAI and ANI illustrate genus and species boundaries, respectively. Proposed genus names are indicated with a ▲.

Other strains form genera groups of four or more genomes that are simply unnamed. Simkaniaceae and Rhabdochlamydiaceae in particular lack solid categorisation, each having only two and one named genera respectively. Further, the genomic diversity displayed within both families is extensive. Some single genomes possess more than ten times the number of unique genes compared to those shared across all genomes (Figure 4.3). It is very likely that Parachlamydiales will need to be split into multiple orders as more data becomes available, and revision of family level affiliation, to better describe this ancient clade of bacteria.

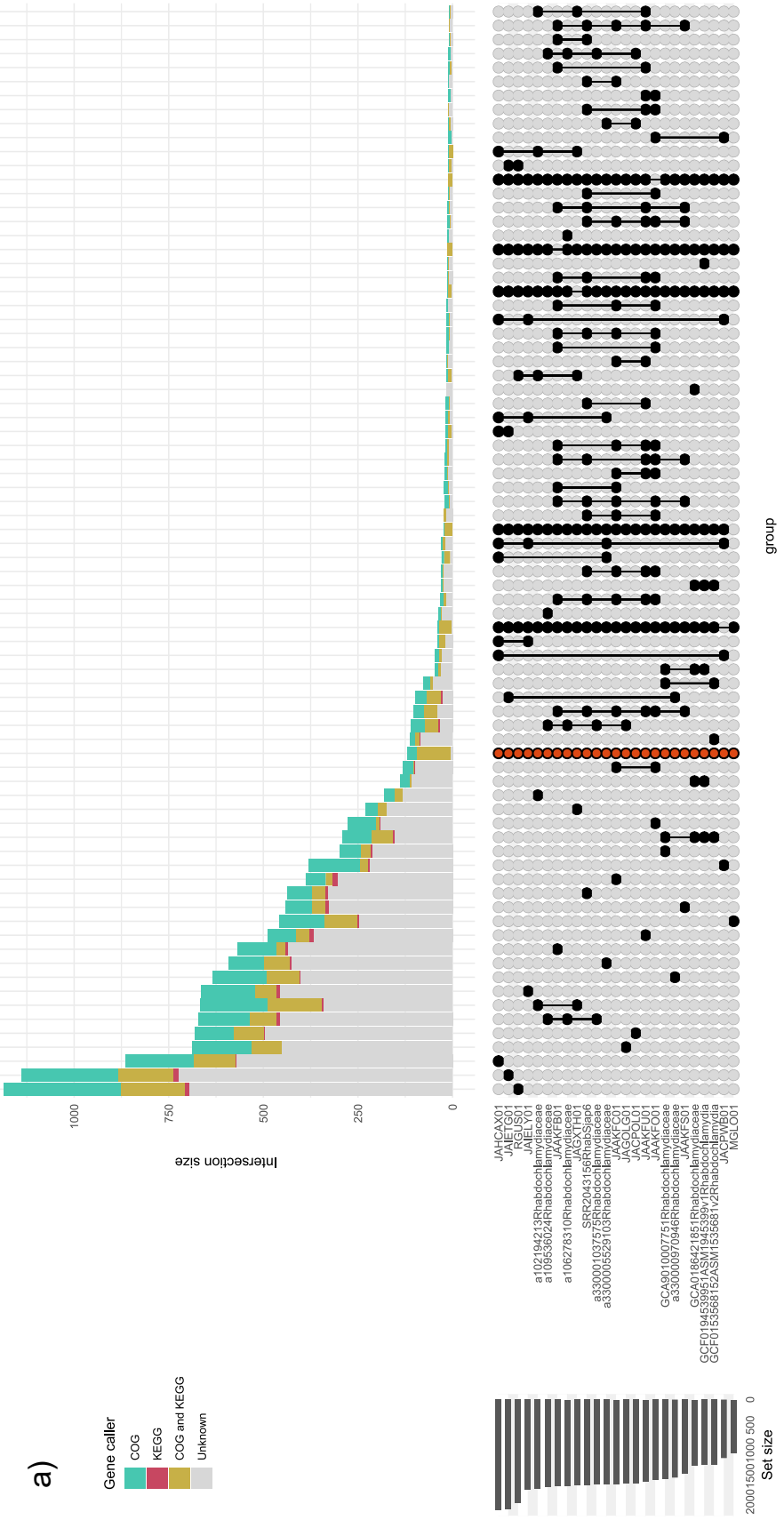


Figure 4.3a. Gene content comparison for Rhabdochlamydiaceae and Simkaniaceae. An upsetplot illustrating the disparity of gene presence absence across the families a) Rhabdochlamydiaceae and b) Simkaniaceae. Vertical bars illustrate the number of genes present shared across compared genomes, indicated by black circles beneath each bar. Orange circles indicate the core genome shared by all genomes in the comparison.

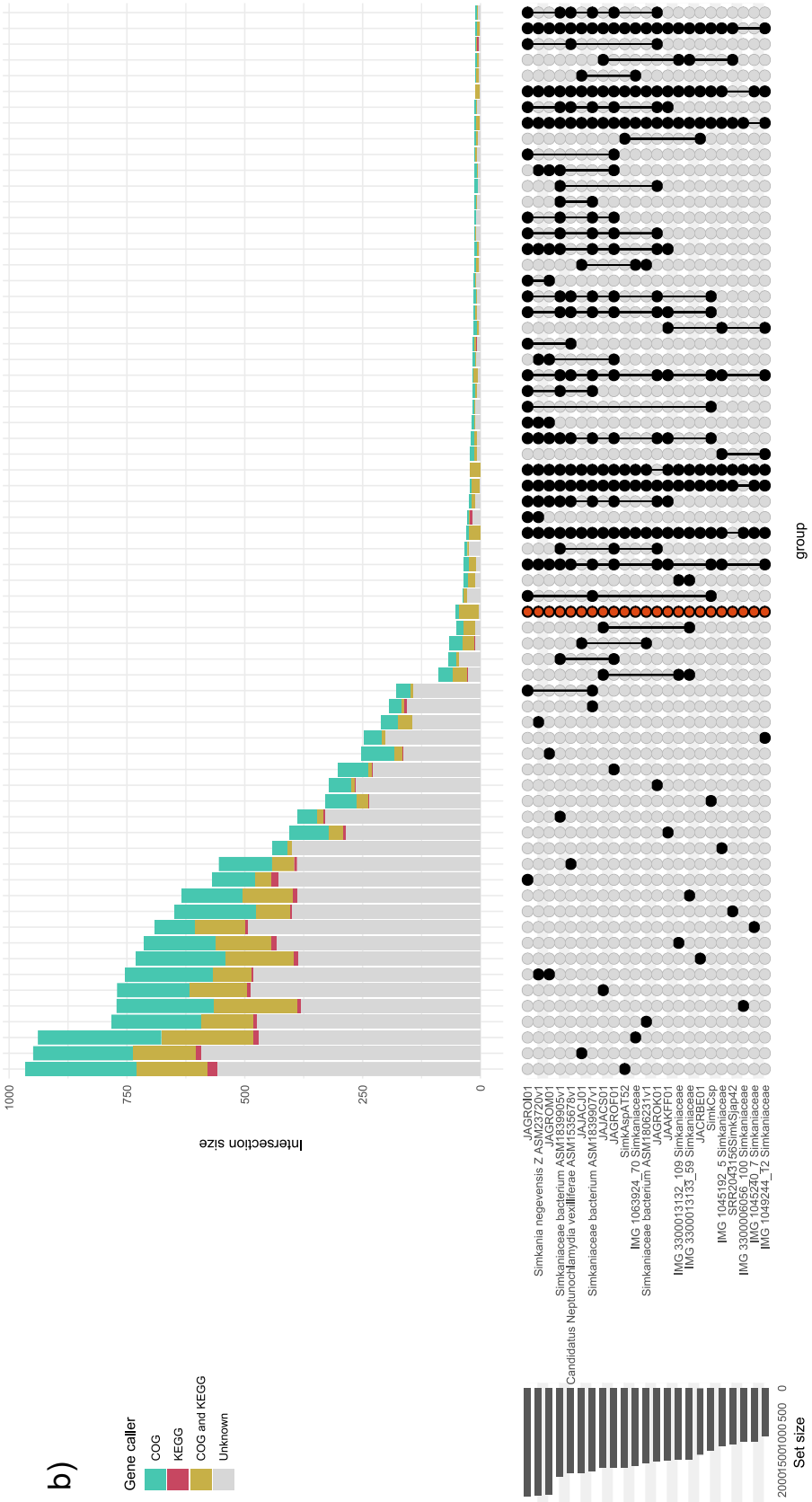


Figure 4.3b. Gene content comparison for Rhabdochlamydiaceae and Simkaniaceae. An upset plot illustrating the disparity of gene presence absence across the families a) Rhabdochlamydiaceae and b) Simkaniaceae. Vertical bars illustrate the number of genes present shared across compared genomes, indicated by black circles beneath each bar. Orange circles indicate the core genome shared by all genomes in the comparison.

4.4.3 Proposed Taxonomy

Based on AAI, ANI, GTDBtk analysis, metabolic results, and phylogeny, I find evidence for at least three new genus-level clusters within Rhabdochlamydiaceae and Simkaniaceae. The genera I propose consist either of multiple high-quality genomes or are high-quality and host associated assemblies.

'*Candidatus* Sacchlamydia' gen. nov. - For the cluster that contains RhabSjap6, a host associated genome, named after the putative host *Saccharina japonica* and the pattern of using "-chlamydia" as the suffix for Chlamydiota genera.

'*Candidatus* Amphrikania' gen. nov. - A second new clade with host association data lies within the Simkaniaceae (the genome named SimkAspAT52). Amphri- after Amphritrite Queen of the Sea and echoing its marine host Amoebophyra host, and -kania after the type genus for the family Simkaniaceae, *Simkania*.

Similarly to SimkAspAT52, clade SimkSjap42 is likely a new species group but will not be named because it has genome that is only 65% complete in CheckM analysis (Figure 4.1., Table 4.1).

'*Candidatus* Acheromydia' gen. nov. - For the clade including the MAGs JAHCAX01 and JACPWB01 which share >65% AAI similarity. For convenience, and so as not to over inflate this group with new names, I also include the relatively closely related genome JAIEL01 (>62% AAI similarity). This genus is named after Acheron (Acher-) one of five Greek rivers of the underworld that occasionally surfaces above ground. The MAGs come from drinking water, activated sludge and ground water. -mydia from *Chlamydia*.

4.4.4 Metabolism

Parachlamydiales have a broad array of metabolic pathways (Figure 4.4 and Appendix D.1). Of note, several have mostly complete B vitamin synthesis pathways (biotin, riboflavin and thiamine), which are known to be associated with beneficial contributions to the host in other symbioses with intracellular bacteria. However, all characterised Parachlamydiales have biphasic lifestyles, and many are known pathogens, so the presence of B vitamin modules does not necessarily indicate potential benefit to the host. For instance, pathogenic *Chlamydia* with bioY genes have previously been observed to uptake biotin from host cells (Fisher *et al.*, 2012). All Parachlamydiales genomes here also have complete Type III secretion systems and incomplete gene sets associated with flagellar assembly (Appendix D.1).

If Parachlamydiales do form beneficial symbioses with their host it is feasibly associated with photosynthesis through vitamin K1 synthesis, CAM, or chlorophyll synthesis. Many Parachlamydiales studied here appear to have CAM (light) pathways with 50% completeness, and 100% in JAGROM01. The reason for the presence of these systems is unclear, but it is persistent across genomes from different studies. Some Parachlamydiales have uroporphyrinogen decarboxylase and uroporphyrinogen-III synthase which form part of several biosynthesis pathways including heme, chlorophyll, cobalamin, and siroheme. In addition, some have incomplete phylloquinone (vitamin K1) biosynthesis pathways; phylloquinone is produced and used by plants, algae, and cyanobacteria in photosynthesis. Most genomes also have a complete chorismate synthase pathway (Figure 4.5, Appendix D.1), an important intermediate product for phylloquinone production, as well as intermediates for alkaloids and salicylic acid that are important in plant defence systems (Hamberger *et al.*, 2006; Shanmugabalaji *et al.*, 2022). It should be noted that chorismate can also be used for a variety of other vital aromatic compounds such as ubiquinone which is common in bacteria (Dosselaere and Vanderleyden, 2001).

There is also a possibility that these bacteria form defensive symbioses, as seen in the case of viral protection of amoebae by *P. acanthamoeba* (Arthofer *et al.*, 2022). Unfortunately, we do not know how *P. acanthamoeba* blocks viral factory formation, so I cannot identify any associated pathways through homology. CRISPR/Cas Phage defence

systems have also previously been found in *Protochlamydia* genomes and other *Chlamydiota* (Deveau, Garneau and Moineau, 2010; Bertelli *et al.*, 2016; Köstlbacher *et al.*, 2021), which could be associated with the anti-viral activity seen in *P. acanthamoeba* (Arthofer et al., 2022). I find that KEGG Kofam identifies cas2 in all *Parachlamydiales* studied in this chapter, including *P. acanthamoeba*. Several have additional types of CRISPR/Cas associated genes (Appendix D.1). CRISPRcas-finder only identifies cas and cas spacers in the four MAGs extracted from the SRA samples (Appendix D.1). Whether these genes are functional in defense against phages is unknown; the class of Cas is not predictive of their function and CRISPR/Cas are common across bacteria but notably rare in microbial symbionts (Deveau et al., 2010).

Several terpenoid synthesis pathways are present, including dTDP-L-rhamnose biosynthesis which can be associated with plant-bacteria symbioses as well as pathogenicity (Ma, Pan and McNeil, 2002; Jofré, Lagares and Mori, 2004; French, 2017; Jiang *et al.*, 2021). Terpenoids are also used by algae themselves in defence against threats such as bacteria and heavy metals (French, 2017; Karimi *et al.*, 2019). Some Red algae use bacteria like pathways to produce their terpenoids (Wei *et al.*, 2019).

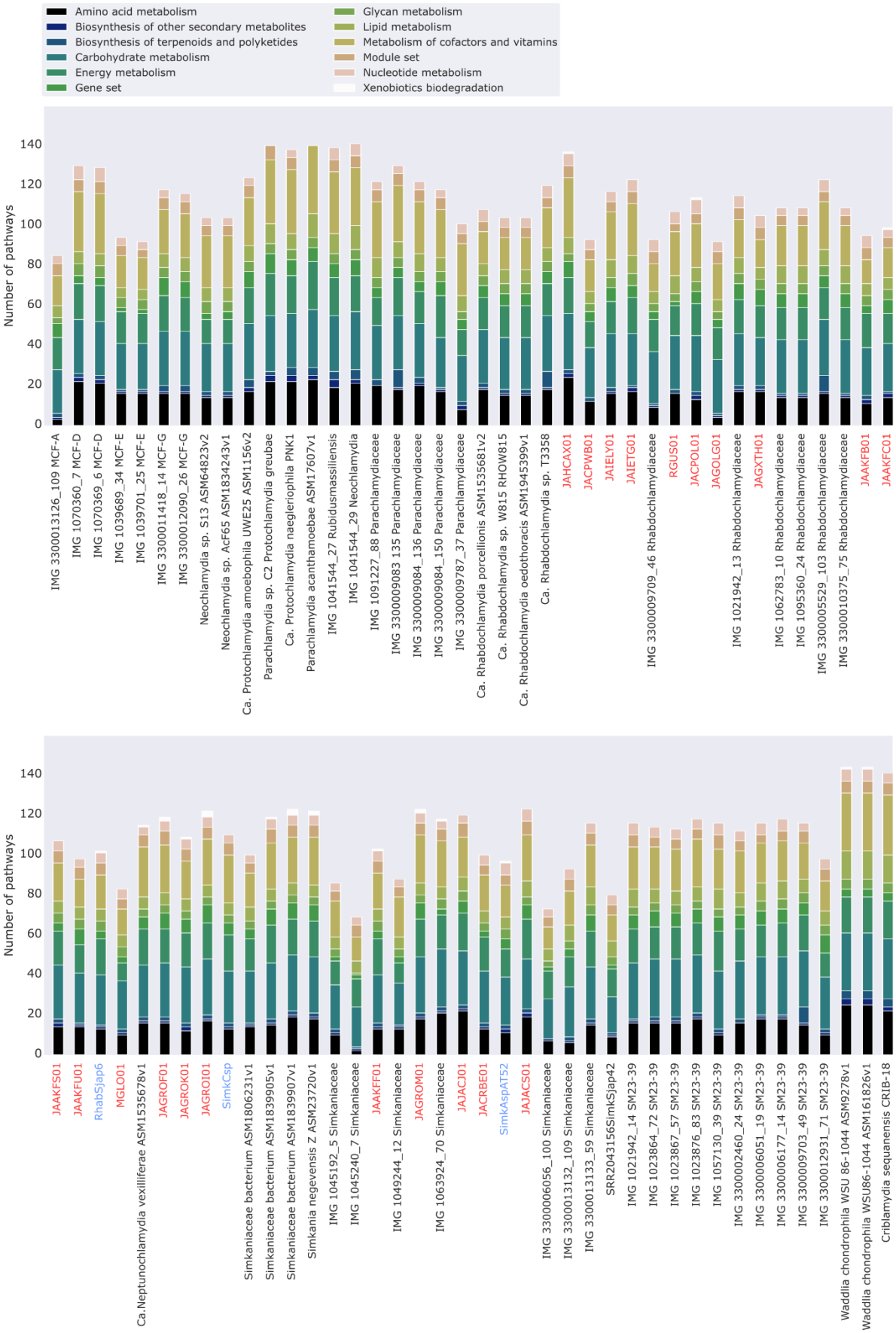


Figure 4.4. The number of each type of metabolic pathway found in each Chlamydiota genome. Red indicates environmental MAGs; blue indicates MAGs assembled in this chapter.

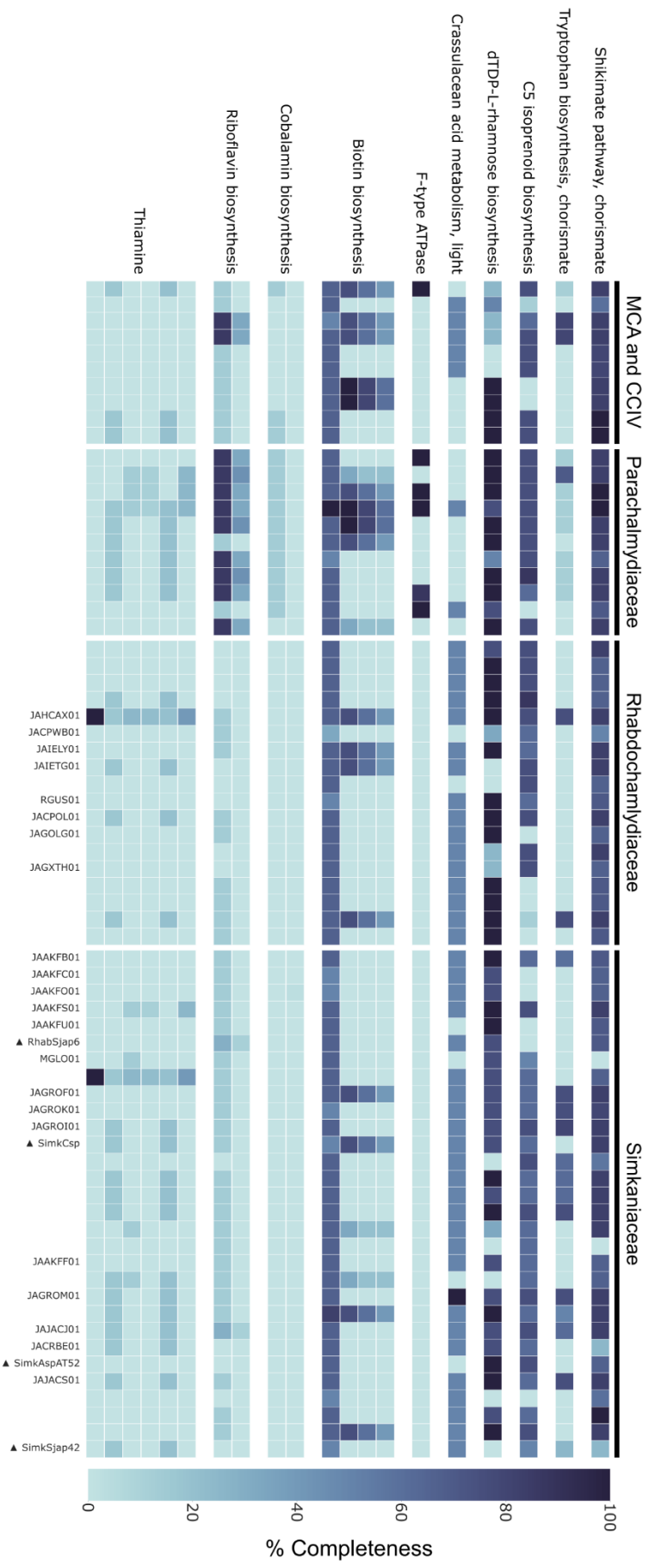


Figure 4.5. Heatmap of the completeness of KEGG metabolic pathways of interest across Parachlamydiales. New genomes are indicated by . Full metabolic pathway completeness for all genomes and all other pathways are available in Appendix D.1. New genomes assembled in this study are indicated by ▲.

4.4.5 Final conclusions

Chlamydiota are hyper diverse and under described. I have added to the known phylogenies of the Parachlamydiales families: Simkaniaceae and Rhabdochlamydiaceae. In addition, I clarify the status of two additional Rhabdochlamydiaceae and one Simkaniaceae genera, for which I propose the names '*Ca*. Sacchlamydia', '*Ca.* Acherochlamydia', and '*Ca.* Amphrikania'. The metabolic potential of Chlamydiota is likewise highly diverse, with several clusters of genes that could be associated with both defensive and nutritional symbiosis or otherwise associated with pathogenicity. Closer study of the metabolic interactions between Parachlamydiales and their microeukaryotic hosts are required to elucidate if and how they affect each other.

Chapter 5. '*Candidatus* Tisiphia' is a widespread symbiont in the mosquito *Anopheles plumbeus*

5.1 Abstract

Symbiotic bacteria alter host biology in numerous ways, including the insect's ability to reproduce or vector disease. Deployment of symbionts in the control of vector borne disease has focussed on *Wolbachia* interactions with *Aedes* but has been hampered in *Anopheles* by a lack of compatible symbioses. Previously, PCR screening noted the presence of symbiotic bacteria '*Ca.* Tisiphia' (== torix *Rickettsia*) in *Anopheles plumbeus* from the UK, an aggressive biter and potential secondary vector of Malaria and West Nile virus in Europe. In this chapter, I first screen *An. plumbeus* samples collected over a ten-year period across Germany using PCR and compare incidence to climate databases to explore the possibility of environmental influence on infection. I find 95% infection rate that does not apparently fluctuate with time, precipitation, temperature, or forest type. I then use FISH imaging to establish that the bacteria are localised to the oocytes, secondary follicles, and lateral ducts of the ovaries, indicating that this is a maternally inherited agent. Finally, I assemble a high-quality draft genome of '*Ca.* Tisiphia' from *An. plumbeus* using Illumina and PacBio reads to explore its affiliation with other strains and potential metabolism. This analysis establishes the *A. plumbeus* symbiont as closely related to one found in *Culicoides* midges and shows similar patterns of metabolic potential. Moving forward, *An. plumbeus* has historically been cultured and provides a viable avenue of symbiosis research in anopheline mosquitoes which to date only have one other proven infection of a heritable symbiotic bacteria. The system also provides future opportunity to study the impacts of '*Ca.* Tisiphia' on natural and transinfected hosts, especially in relation to reproductive fitness and vector efficiency.

5.2 Introduction

Bacterial symbionts in insects form vital components of their host's biology, ecology, and evolution. They are known to influence how insects reproduce, how they respond to environmental stress, and their interaction with pathogens and parasites (Dunbar *et al.*, 2007; Himler *et al.*, 2011; Vega, Arribére and Castro-vazquez, 2012; Hendry, Hunter and Baltrus, 2014; Xie *et al.*, 2014; Hayashi *et al.*, 2016). Several species of symbionts are

vertically inherited from one generation to the next, usually through the maternal germline, and may become intrinsically linked with their host physiology, metabolism, and development (Buchner, 1965; Zchori-Fein, Borad and Harari, 2006; Moran, McCutcheon and Nakabachi, 2008; Kremer *et al.*, 2009; Giorgini *et al.*, 2010). Most importantly, symbionts have been deployed in the control of vector populations and vector competence (Hoffmann *et al.*, 2011).

Success in symbiont-mediated disease control have been restricted to species from the genus *Aedes*. Transinfection with *Wolbachia* from a drosophilid fly have been successfully used to alter vector competence and lower risk of catching Dengue Fever from *Aedes aeygpti* in endemic areas (Hoffmann *et al.*, 2011; Walker *et al.*, 2011; Pereira *et al.*, 2018). However, important vectors like *Anopheles* are rarely naturally infected with *Wolbachia*, and species within the group are commonly unreceptive to artificial *Wolbachia* infections (Hughes *et al.*, 2014). In *Anopheles* mosquitoes there is a single well-established case of natural *Wolbachia* infection (Walker *et al.*, 2021). Therefore, it is desirable to find potential alternatives that are either more capable of surviving transinfection or alter vector competence in the native host species.

We previously detected the symbiont '*Ca.* Tisiphia' (== Torix group Rickettsia) in *Anopheles plumbeus* in the UK (Pilgrim *et al.*, 2021). *An. plumbeus* is broadly distributed across Europe and is an indiscriminate biter. It is also capable of transmitting West Nile virus and malaria, although these diseases do not natively occur in the majority of its known range and competence has only been tested in the laboratory setting (Bueno-Marí and Jiménez-Peydró, 2011; Dekoninck *et al.*, 2011; Schaffner *et al.*, 2012). It has been highlighted as a species that could act as a secondary vector for tropical disease as changing climates causes the northward spread of tropical diseases and their associated primary hosts like *Aedes albopictus* (Schaffner *et al.*, 2012; Heym *et al.*, 2017).

'Ca. Tisiphia' are theorised to be commonly associated with hosts deriving from wet or aquatic environments and may originate from symbionts of freshwater ciliates (Driscoll *et al.*, 2013; Schrallhammer *et al.*, 2013; Kang *et al.*, 2014). Infection with *'Ca.* Tisiphia' occurs in a broad range of invertebrates from annelids to gastropods to arthropods (Pilgrim *et al.*, 2021), as well as algae (Hollants *et al.*, 2013) and amoebae (Dyková *et al.*, 2013). Their relatives in *Rickettsia* are capable of nutritional symbioses, protecting against

fungal infections, and reproductive manipulation (see Table 1.2). However, the known effects of '*Ca*. Tisiphia' itself are limited to an association with increased host size in Torix leeches, and weak impacts on fecundity in *Cimex lectularius* bedbugs (Kikuchi and Fukatsu, 2005; Thongprem *et al.*, 2020). There is no observed congruence of host and symbiont phylogeny, indicating that host shifts occur commonly and that long standing associations with species are rare. External influence such as temperature or natural enemies can also influence the prevalence of symbionts in host populations (Cass *et al.*, 2016; Corbin *et al.*, 2017; Leclair *et al.*, 2017). The most likely scenario for *'Ca.* Tisiphia' is that it has a similar life history to *Rickettsia*, and that its impacts have not been observed because a) they are context dependant to their environment or b) tests of phenotype have not been carried out.

Here I use PCR assays to establish the extent of '*Ca.* Tisiphia' infection in *An. plumbeus* mosquitoes across Germany collected through Citizen Science initiatives, and assess potential associations with temperature, precipitation, and forest type. I also sequence and assemble draft genomes for the symbiont '*Ca.* Tisiphia' and provide evidence of vertical transmission of the symbiont through the maternal germline through FISH imaging. The symbiont genome is examined through bioinformatics approaches to establish potential nutritional or protective symbioses.

5.3 Methods

5.3.1 Collection of Anopheles plumbeus

Two hundred and fifty-five *An. plumbeus* specimens from 2012-2021 were collected across Germany by Doreen Werner and citizen volunteers as part of the mosquito atlas (Mückenatlas) project (Werner *et al.*, 2014). These were stored in 70% ethanol or dry (see supplementary materials for storage and exact geographic information). Post hoc analysis indicated storage method did not affect detection of symbionts by PCR assay.

Specimens were also collected by Helen Davison as larvae and raised to adults in water collected from their larval pools. These specimens were either killed by flash freezing in liquid nitrogen prior to genomic DNA extraction, or in 4% paraformaldehyde solution prior for fluorescence imaging.

5.3.2 DNA extraction and PCR screening of Anopheles plumbeus for 'Ca. Tisiphia'

Promega Wizard® Genomic DNA Purification kit was used for DNA preparation. DNA quality was then checked with a combination of HCO/C1J primers HCO_2198 (5'-TAA ACT TCA GGG TGA CCA AAA AAT CA-3')/CIJ_1718 (5'-GGA GGA TTT GGA AAT TGA TTA GT-3') (Folmer *et al.*, 1994; Hajibabaei *et al.*, 2005; Siozios *et al.*, 2020). *Ca.* Tisiphia presence was assessed with a PCR assay amplifying the 320-bp region of the 17 kDa omp gene Ri17kD_F (5'-TCTGGCATGAATAAACAAGG-3')/Ri17kD_R (5'-ACTCACGACAATATTGCCC-3') (Pilgrim et al., 2017). PCR conditions used were as follows: 95 °C for 5 min, followed by 35 cycles of denaturation (94 °C, 30 s), annealing (54 °C, 30 s), extension (72 °C, 120 s).

A selection of '*Ca.* Tisiphia' amplicons from positive samples across time and space were Sanger sequenced through Eurofins barcode service and identity confirmed by comparing them to the NCBI database via BLAST homology searches.

5.3.3 Association of symbiont prevalence with geographic and climatic information

Annual average monthly temperature and precipitation data were retrieved for each sample's coordinate and year from TerraClim (Abatzoglou *et al.*, 2018) which has a spatial resolution of ~4-km (1/24th degree). Forest cover data was retrieved from Copernicus land datasets for 2018 (European Union, 2018) and raster data for forest type extracted in QGIS 3.16 (QGIS.org, 2020) within a 3km radius of each sample location. *Anopheles plumbeus* has historically been recorded to have a maximum flight range of up to 13km (Becker et al., 2010). However, this is based on one single study from 1925 and is not verified by other sources. As such I chose an estimated range of 3km based on the average flight ranges other Anopheline mosquitos (Becker et al., 2010; Verdonschot & Besse-Lototskaya, 2014). Scikit-learn's standard scaler (Pedregosa *et al.*, 2011) was applied to data before performing a generalised linear model with a binomial logit link function on data with the following formula:

$$Infected \sim tasmin + tasmax + precip + forest\_ratio$$

All statistics and geographic inferences were carried out in Python with the packages Statsmodel and Scikit-learn (Rossum and Drake, 2009; Seabold and Perktold, 2010; Pedregosa *et al.*, 2011). QGIS 3.16 was used to produce maps and extract raster data for forest types before passing it to python for analysis (QGIS.org, 2020). All other figures

were produced with Matplotlib and Seaborn (Hunter, 2007; Waskom and Seaborn development team, 2020).

5.3.4 Fluorescence in situ microscopy (FISH)

Reproductive organs of a single female and a single male were dissected and incubated in cold 4% paraformaldehyde for 3 hours, agitated gently every 30 minutes, then washed with cold PBS for 5 minutes two times. Tissue was stained with Hoescht ds33342 30 minutes at room temperature, then hybridised overnight at room temperature with hybridisation buffer (5X SSC, 0.01% SDS, 30% formamide) and 5'-CCATCATCCCCTACTACA-(ATTO 633)-3' oligonucleotide probe specific to *Ca*. Tisiphia 16S rRNA (Pilgrim *et al.*, 2017). Hybridised tissue was washed in wash buffer (5X SSC, 0.01% SDS) at 48°C for 60 minutes with gentle shaking every 20 minutes. Samples were then mounted in Vectashield. Images were taken with a ZEIS LSM 880 confocal microscope through ZEIS Zen black, and final images were annotated in Inkscape Ver 1.2 (Inkscape Project, 2020).

5.3.5 De novo sequencing, assembly, and annotation.

A combination of short and long read sequencing was used to construct scaffolds for the *'Ca.* Tisiphia' genome. For short reads, Iridian Genomes extracted and processed DNA of one male for illumina sequencing deposited under bioproject accession PRJNA694375. The short reads of *An. plumbeus* was cleaned with trimmomatic 0.36 (Bolger, Lohse and Usadel, 2014) and quality checked with FASTQ (Babraham Bioinformatics, 2019). For long reads, genomic DNA from one male was extracted with Qiagen Genomic-tip for ultra-low PacBio sequencing carried out by the Centre for Genomic Research, University of Liverpool. Long read sequences are deposited under bioproject accession number PRJNA901697.

A combination of long and short reads were used to assemble as complete a genome for '*Ca.* Tisiphia' as possible. First, the '*Ca*. Tisiphia' genome was identified in the illumina short reads and assembled through Minmap2, MEGAHIT and MetaBAT2 as per the pipeline used in Chapter 2. Second, PacBio HiFi long read sequences were assembled using Flye 2.9. 1-b1780 with the '-meta' flag to improve sensitivity for low coverage reads. Third, the long read assembly was queried against a local blast database of *Rickettsia* and '*Ca*. Tisiphia' genomes, (including the illumina assembly from the first step) to identify sequences belonging to this strain of '*Ca*. Tisiphia'. Lastly, the long read assembly was

polished with the Illumina reads using Polypolish (Wick and Holt, 2022) with default settings to give 23 final scaffolds.

5.3.6 Phylogeny and metabolic predictions

Annotation of the final assembly was carried out with InterProScan v5 (Jones *et al.*, 2014). Metabolic pathway prediction for presence and completion was carried out through Anvi'o 7 using KEGG kofams and COG20 (Aramaki *et al.*, 2020; Eren *et al.*, 2021; Galperin *et al.*, 2021). NRPS pathways were investigated with AntiSMASH 6.0 (Blin *et al.*, 2021).

The '*Ca*. Tisiphia' strain for *An. plumbeus* was compared to the other existing '*Ca*. Tisiphia' genomes through Anvi'o 7. A core genome consisting of 205 gene clusters that contain a total of 3280 genes was found through Anvio-7. Phylogenies were estimated from single copy gene clusters with IQTREE 2.1.4 using Model Finder Plus and with 1000 ultrafast bootstraps and SH-aLRT support (Kalyaanamoorthy *et al.*, 2017; Hoang *et al.*, 2018; Minh *et al.*, 2020).

5.4 Results and Discussion

5.4.1 Distribution and predicted environment

'Ca. Tisiphia' was observed to infect *An. plumbeus* across all sites examined in Germany, with 95% specimens positive on PCR assay. The few negative specimens were found to mostly occur in the southeast of the country (Figure 5.1). The infection seems to be stable and there is no evidence of frequency change over time, with samples from all years spanning 2012 to 2022 displaying similar rates of infection (Figure 5.3, Appendix E.1, Appendix figure A.11).

There is no clear evidence of an influence on '*Ca*. Tisiphia' infection rates in *An. plumbeus* caused by average minimum or maximum temperature, precipitation or forest types (Figure 5.2, and Figure 5.4). While there appears to be a significant effect of precipitation on the number of uninfected individuals, this could be an artifact of increased water availability leading to more mosquitoes and thus a higher chance of detecting rarer uninfected individuals (Appendix figure A.11). No variation is unsurprising as it appears to be a very high prevalence infection. I also acknowledge that using climate databases to retroactively find data is not as accurate as field measurements. However, results agree with previous field observations of *Rickettsia* infection in *Acyrthosiphon pisum* in Japan

(Tsuchida *et al.*, 2002). I chose to use the high resolution TerraClim database, but this may still mask small differences in microenvironments and is limited to mostly abiotic data. I encourage future symbioses research to consider environmental measurements to describe the ecology of these organisms more comprehensively.

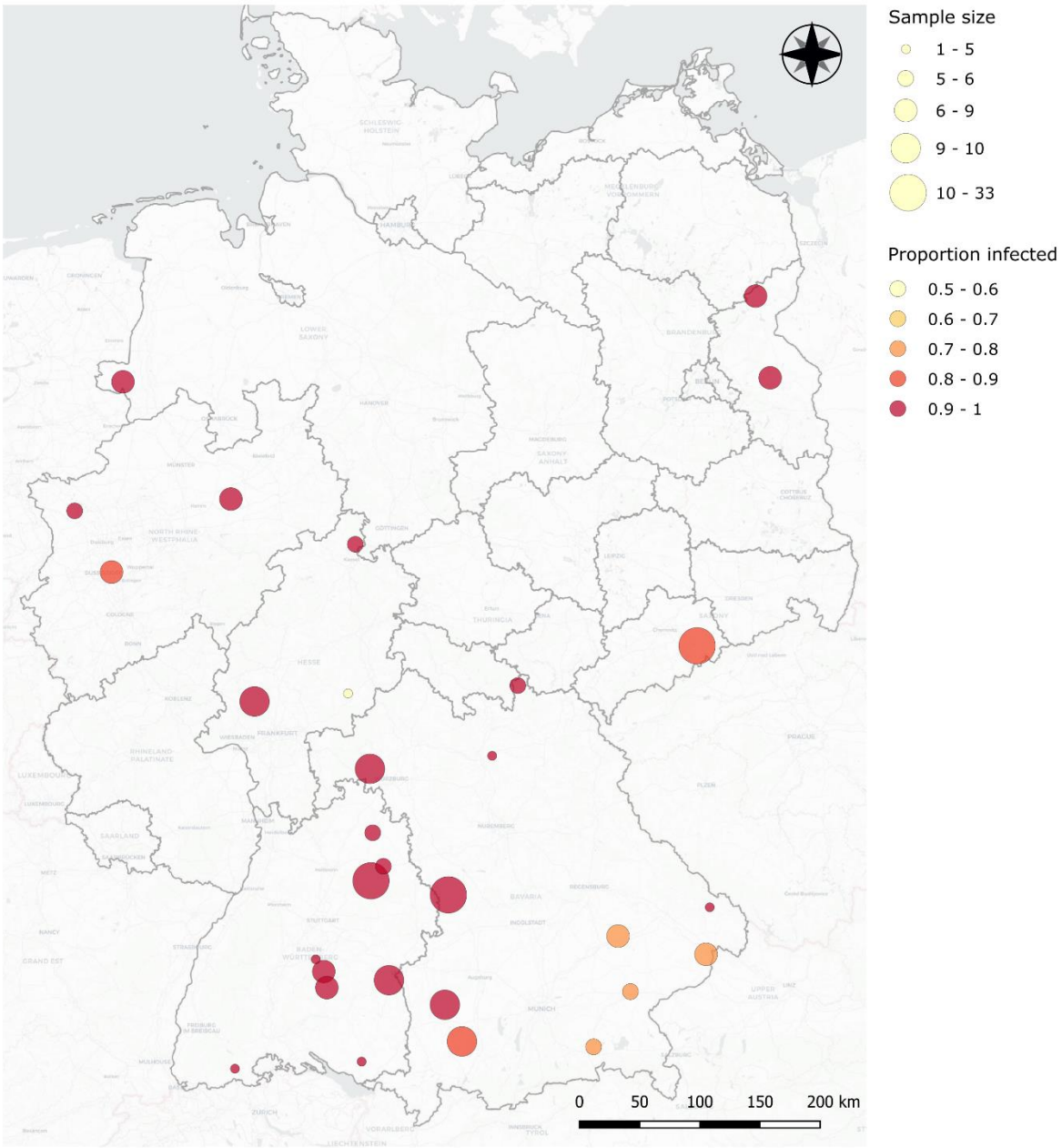


Figure 5.1. Map of '*Ca*. Tisiphia' infection rates across Germany where the size of the circle represents the number of individuals sampled and the colour indicates the proportion of '*Ca*. Tisiphia' infected individuals. Red = 90-100% infection to light yellow = 50-60% infection.

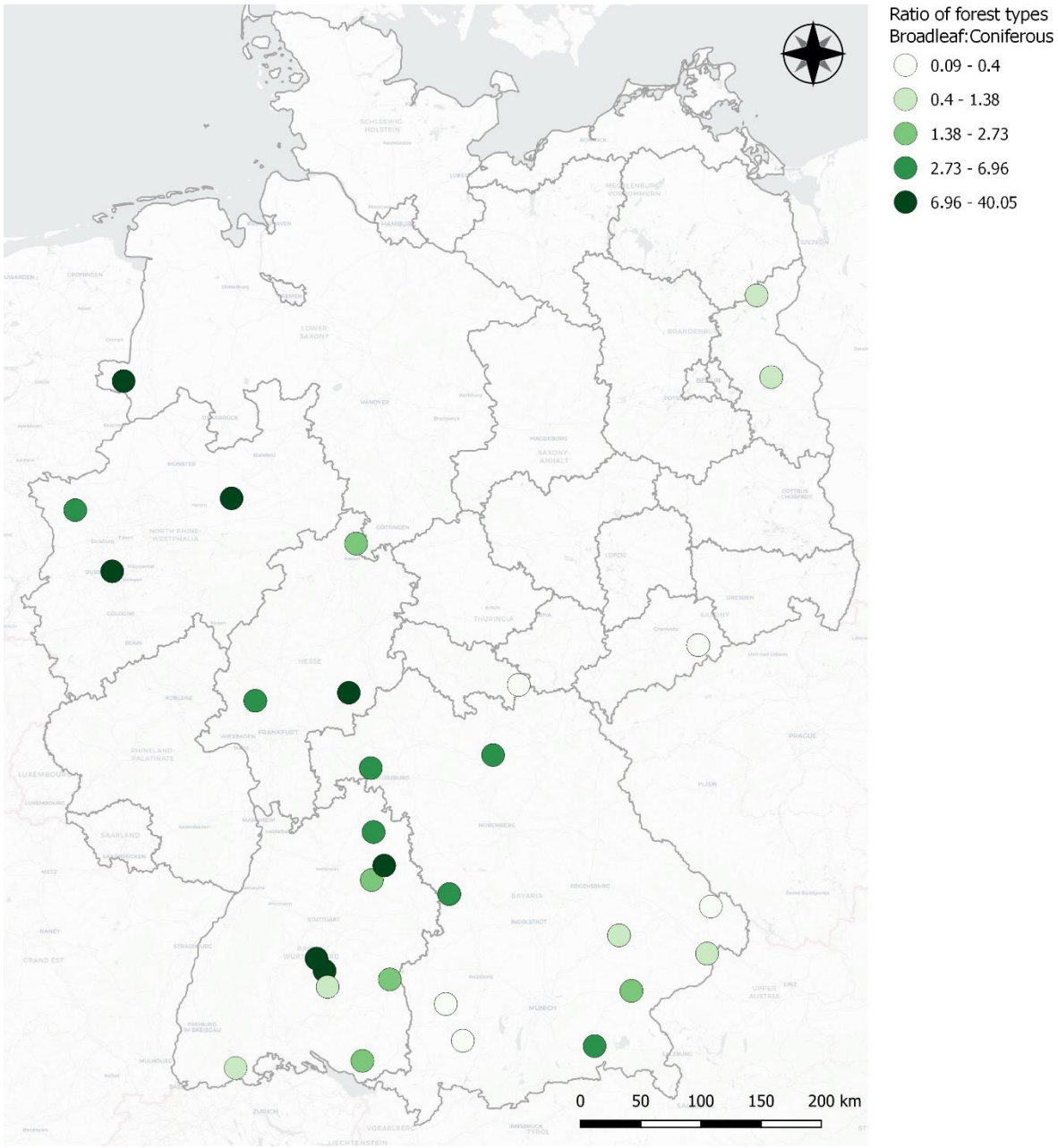


Figure 5.2. The ratio of broadleaf to coniferous forest in a 3km radius of each collection site. Darker green indicates more broadleaf, lighter green indicates closer to equal proportions.

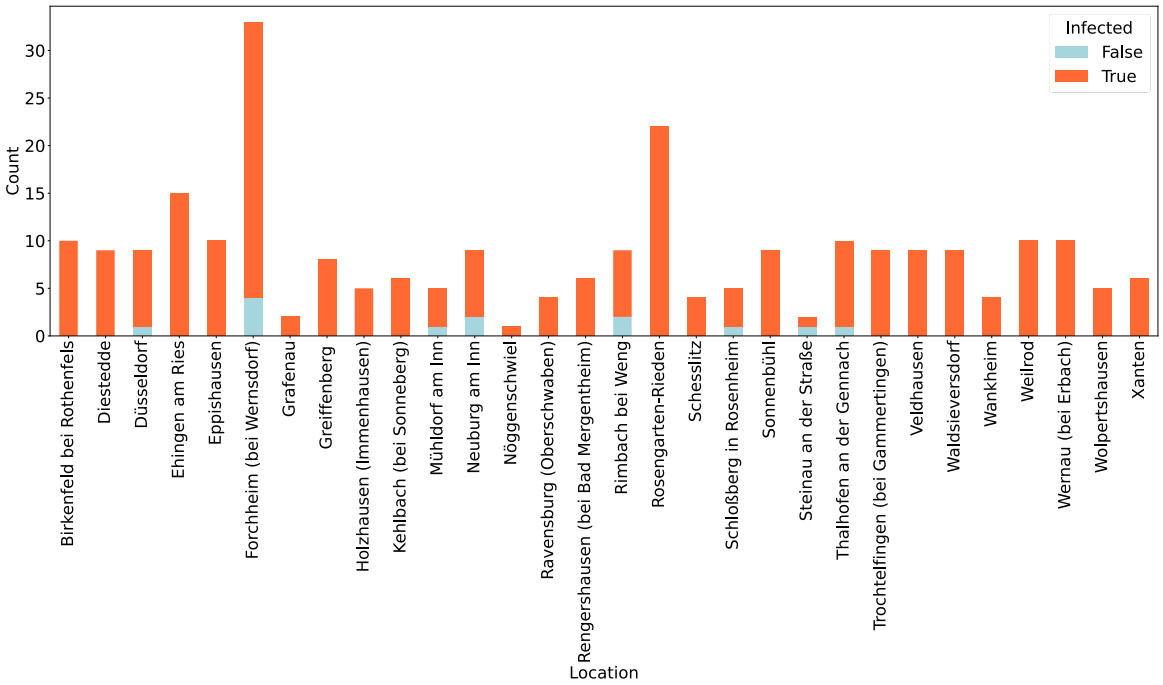


Figure 5.3. '*Ca*. Tisiphia' infection rates by site. Positive infections are shown in orange, negative infections are shown in light blue.

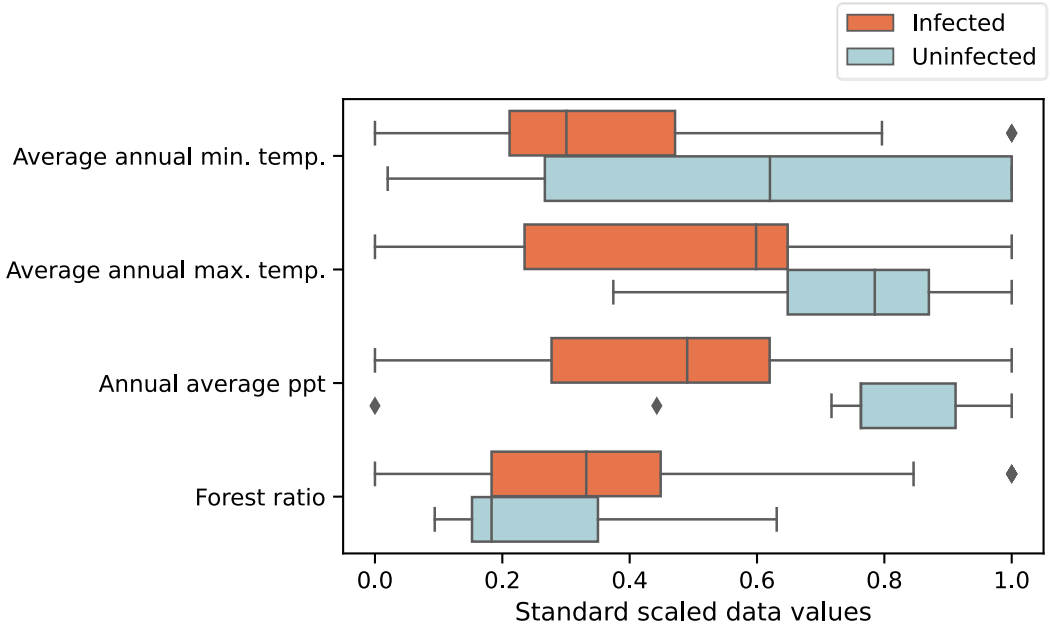


Figure 5.4. Standardised and scaled environmental data comparing Uninfected (N=13) and Infected (N=237) by environmental variable.

5.4.2 Phylogeny and metabolism

The bacteria sequenced from *An. plumbeus* is most closely related to a '*Ca*. Tisiphia' found in the midge *Culicoides newsteadi* (Figure 5.5). General features of both genomes are consistent with other '*Ca*. Tisiphia' (Table 5.1 and see Table 2.1 where it is labelled as Torix *Rickettsia*); TsAplum has a single full set of rRNAs (16S, 5S and 23S) and GC content is ~33%. It also has several repeat domains (Table 5.1) which are associated with protein-protein interactions and are prevalent in *Wolbachia* symbionts (Siozios *et al.*, 2013; Rice, Sheehan and Newton, 2017).

Table 5.1. Summary of the genome assembly for TsAplum.

Strain Name	TsAplum
Symbiont genome accession	n/a
Host	*Anopheles plumbeus*
Raw reads accession	SRR22298143
Total nucleotides	1,622,210
Contigs	31
GC content	32.82%
N50	62798
Number of CDS	1701
Avg. CDS length (bp)	788
Coding density	82.57%
rRNAs	1 x 5S, 1 x 16S, 1 x 23S
tRNAs	31
ORFs with Ankyrin repeat domains	3
ORFs with Leucine rich repeats	1
ORFs with Tetratricopeptide repeats	6

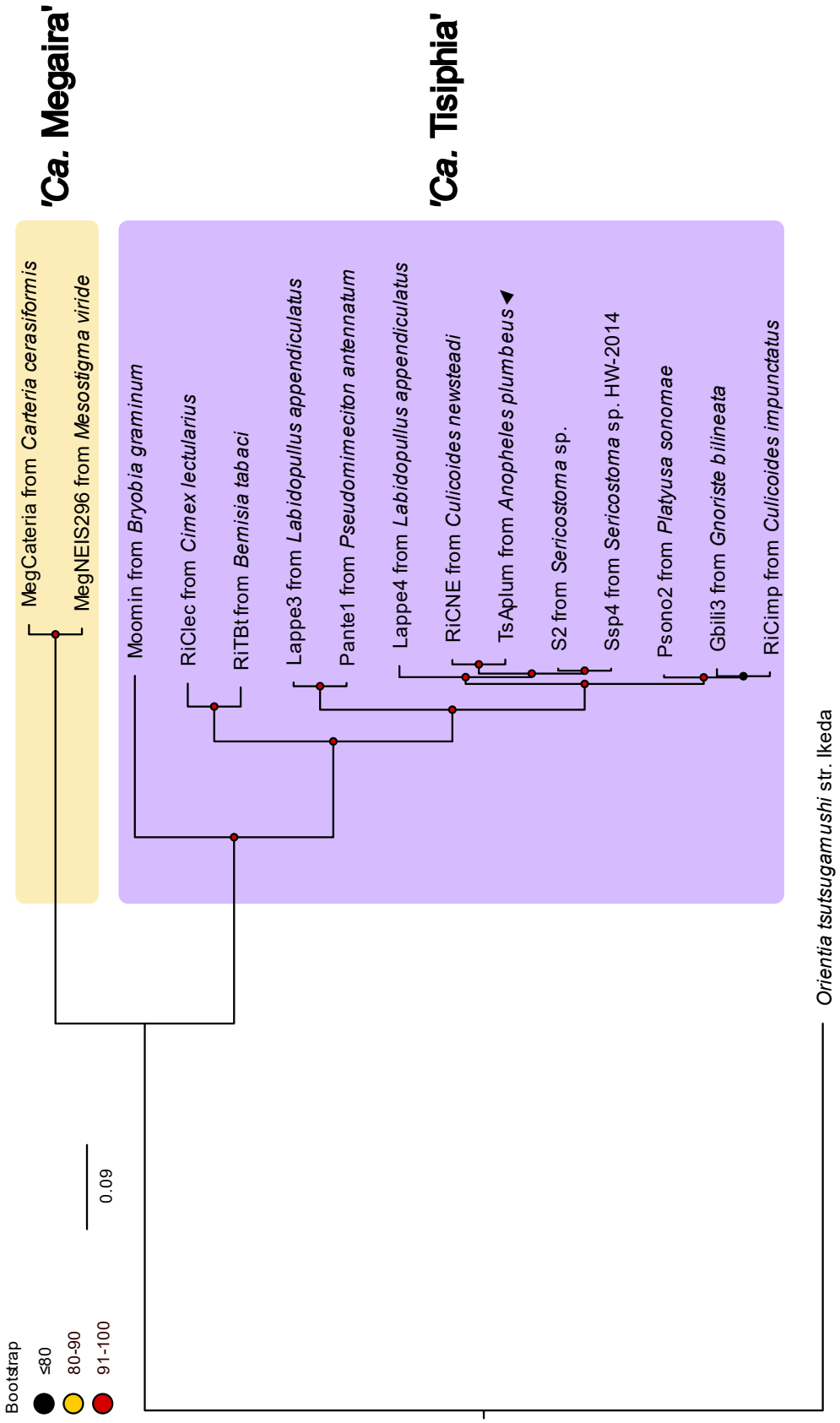


Figure 5.5. Genome wide phylogeny of '*Ca*. Tisiphia' and '*Ca*. Megaira'. Maximum likelihood (ML) phylogeny constructed from 205 single copy gene clusters that contain a total of 3280 genes. New genomes are indicated by ◄ and bootstrap values based on 1000 replicates are indicated with coloured circles (red = 91-100, yellow = 81-90, black <= 80).

Overall '*Ca*. Tisiphia' found in *An. plumbeus* mirrors the metabolic potential found in other members of its genus (Figure 5.6 and 5.7). It does not have any obvious metabolic pathway that would contribute to nutritional symbiosis such as B vitamin production nor any NRPS/PKS system for small molecule synthesis (Appendix E.1.). It does have several toxin/anti-toxin systems as well as secretion pathways Tat, Sec, VirB (Type IV), all of which are essential in various symbiont-host interactions (Masui, Sasaki and Ishikawa, 2000; Meloni *et al.*, 2003; Wu *et al.*, 2004; Dale and Moran, 2006; Tseng, Tyler and Setubal, 2009). Additionally, it has a number of ORFs containing ankyrin and leucine rich repeats which are thought to be important in interactions with cognate eukaryotic proteins (Siozios *et al.*, 2013; Rice, Sheehan and Newton, 2017). Thus, the genome itself, whilst firmly placing the symbiont in the context of the genus and identifying relatedness to other strains, does not raise obvious hypotheses about the impact of infection on the host. Phenotype studies are required to properly assess the influence of this bacteria on its host. Key studies would address the factors driving the spread of the symbiont into the population (testing for beneficial aspects of infection, cytoplasmic incompatibility, and paternal inheritance) and impacts on viral infection and transmission outcomes.

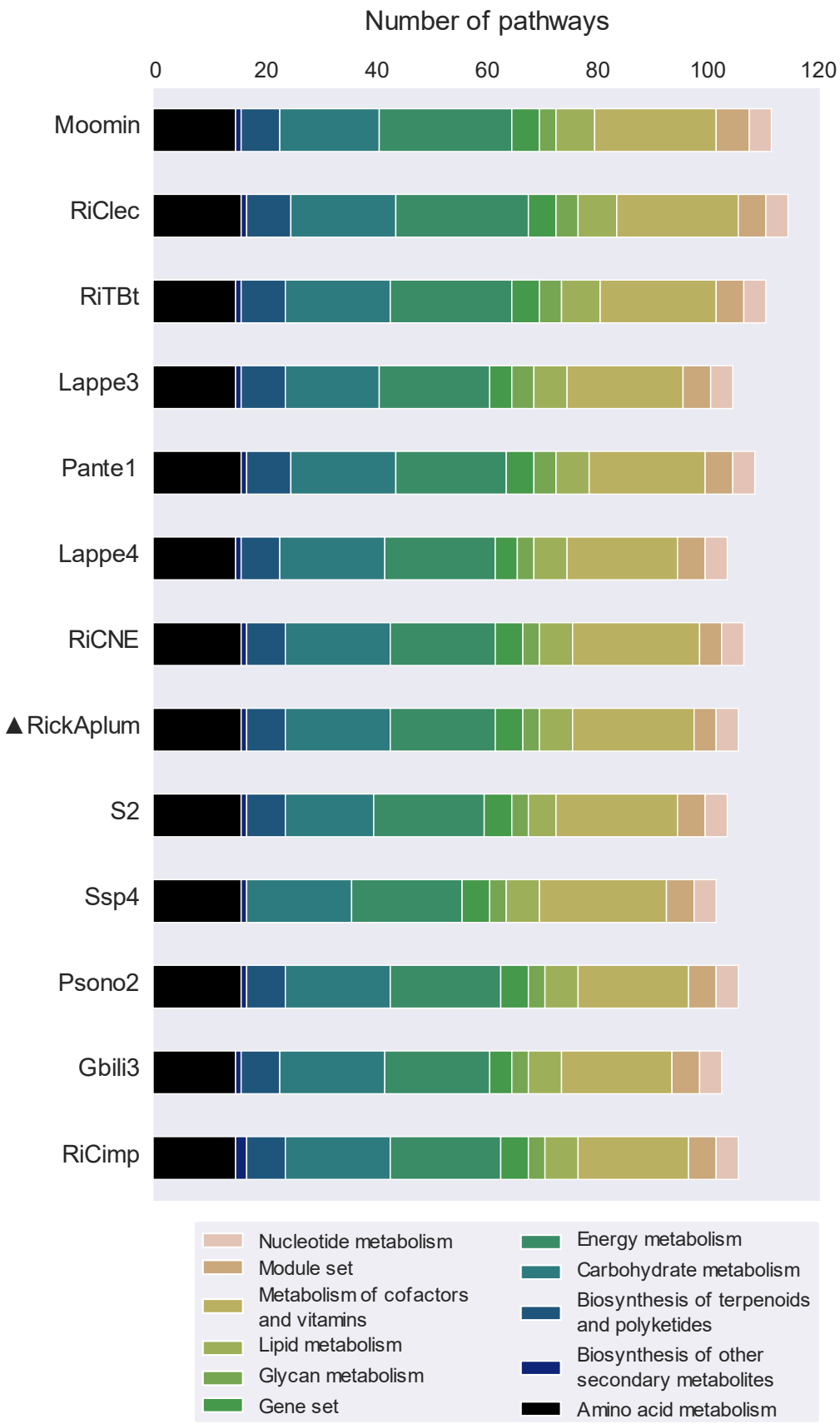


Figure 5.6. KEGG module distribution in '*Ca*. Tisiphia'. The number of pathways found per genome annotated by KEGG module category for '*Ca*. Tisiphia'. Full metadata can be found in Appendix E.1. ▲ indicates the genome assembled in this chapter.

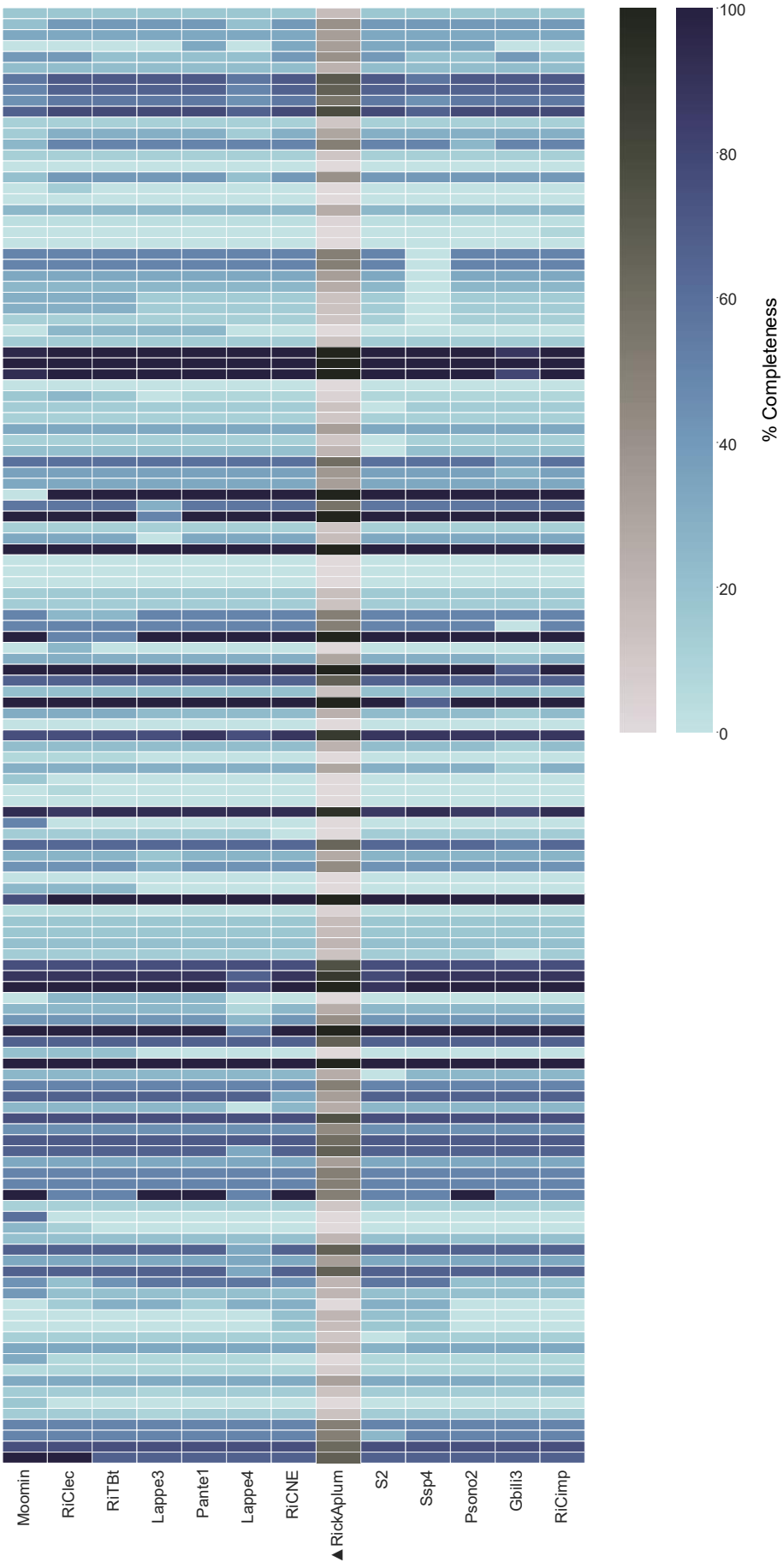


Figure 5.7. Predicted completeness of KEGG kofam metabolic pathways across '*Ca*. Tisiphia'. The genome assembled in this chapter is coloured grey and indicated with ▲. Full metadata can be found in Appendix E.1.

5.4.3 FISH imaging

'Ca. Tisiphia' is observed in oocytes and oviduct branches but was not detected in testes (Figure 5.8 versus Figure 5.9). Localisation and clear polarity of the infection in ovaries strongly suggest that this is a maternally inherited infection (Figure 5.8). The bacteria cluster around the oocyte of the primary follicles as well as in the lateral ducts and secondary follicles. In *Drosophila melanogaster*, *Wolbachia* is similarly polarised to one end of the primary follicles to the oocyte (Ferree *et al.*, 2005), and in *Proechinophthirus fluctus*, their endosymbionts *Sodalis* appears to use the lateral oviducts to access the ovaries (Boyd *et al.*, 2016).

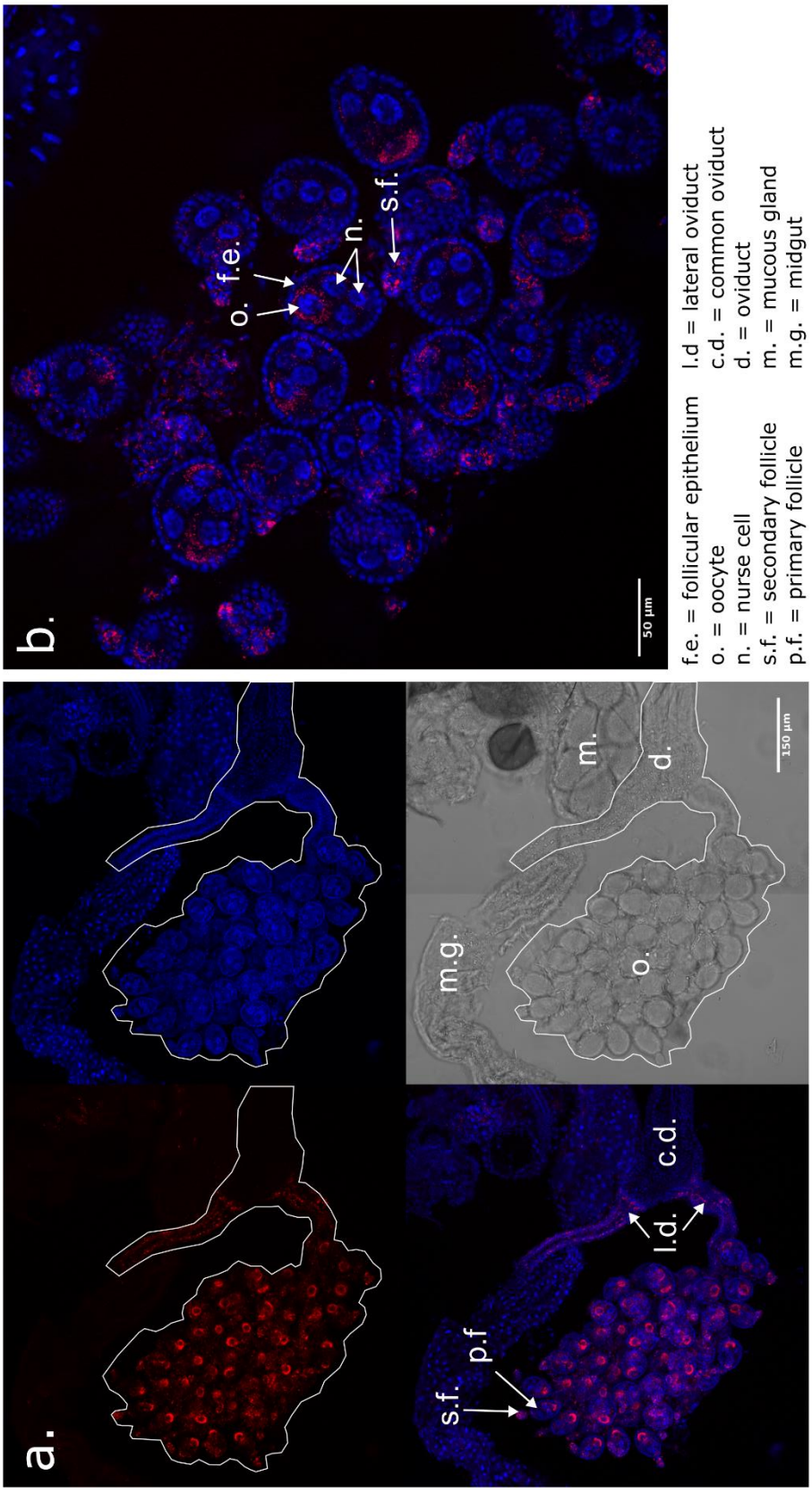


Figure 5.8. Fluorescence in situ microscopy images of *Anopheles plumbeus* ovaries infected with '*Ca*. Tisiphia'. Red shows '*Ca*. Tisiphia' stained with ATTO-633, blue are host nuclei stained with Hoechst blue dye. Panels show a) the whole female reproductive organ outlined in white and a breakdown of each light channel and b) a close up of the ovaries showing localised infection within the primary and secondary follicles. White bars indicate a) 150 micrometres and b) 50 micrometres.

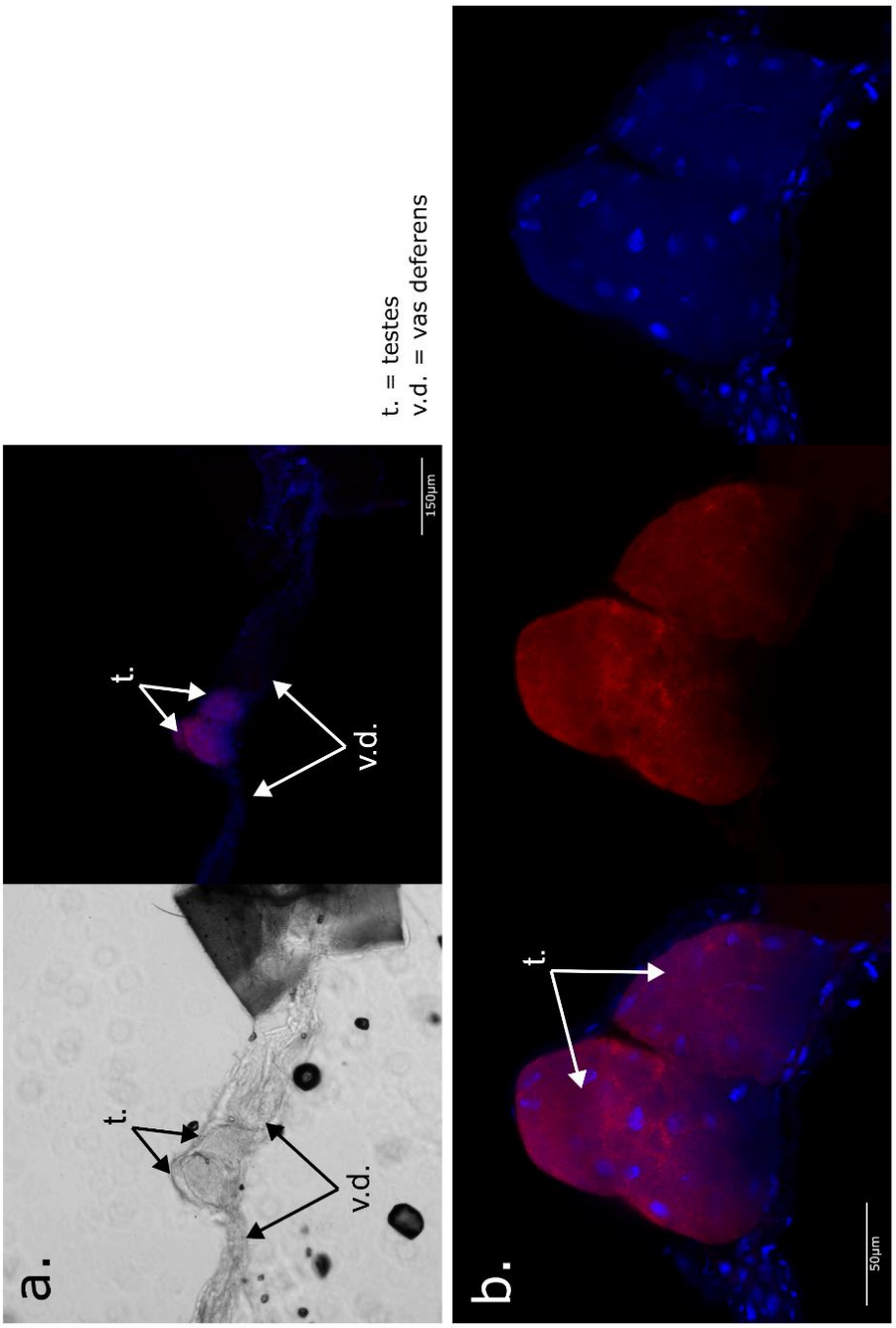


Figure 5.9. Fluorescence in situ microscopy images of *Anopheles plumbeus* testes. Blue are host nuclei stained with Hoechst 33342, Red is ATTO-633 auto-fluorescence in the testes not '*Ca*. Tisiphia' staining. White bars indicate a) 150 micrometres and b) 50 micrometres.

5.4.4 Final conclusions

An. plumbeus and its '*Ca.* Tisiphia' make a good potential model for symbioses in *Anopheles* mosquitoes as well as '*Ca*. Tisiphia' infection more generally. Outside of *An. plumbeus*, there is only a single well substantiated case of *Wolbachia* and no other symbiont in Anopheline mosquitoes (Walker *et al.*, 2021). The infection in *An. plumbeus* is clearly evidenced, likely heritable, and occurs in a species that has seen previous success as a laboratory colony (Kotter, 2005). Beyond this, *An. plumbeus* is a species of interest with the capability to carry west nile virus and malaria (Dekoninck *et al.*, 2011; Schaffner *et al.*, 2012). '*Ca*. Tisiphia' in *An. plumbeus* provides a viable avenue for symbiont-mediated vector modification and control to be tested in anopheline species. It is also an example of a temporally and spatially stable infection of non-pathogenic Rickettsiaceae and a good foil to fluctuating systems like Belli *Rickettsia* in *Bemisia tabaci* (Bockoven *et al.*, 2020).

Future work should also establish the effects of this symbiont in transinfection in alternative hosts alongside the native *An. plumbeus* host. Other symbionts like *Wolbachia* are known to produce functionally interesting phenotypes related to vector competence when transferred from the original host into other, naïve species (Moreira *et al.*, 2009). Alongside this, impacts on host function and physiology, and potential means of spread into natural populations would need to be assessed. A first step to establishing transinfection would be to isolate the '*Ca.* Tisiphia' infection into cell culture, which would also represent an important community resource for onward study.

In summary, '*Ca*. Tisiphia' is found in 95% of *An. plumbeus* individuals from Germany and forms a well-established, stable, and heritable infection that persists across space and time. Metabolic potential is typical of similar symbiotic bacteria species, and I find no evidence of large-scale environmental factors influencing rates of infection. However, '*Ca.* Tisiphia' and *An. plumbeus* provide a unique opportunity to study the effects of a native symbiont infection in anopheline mosquitoes, as well as explore its potential use for disease mitigation in other species that cannot be infected with currently used symbionts.

Chapter 6. Discussion

6.1 Synthesis

Intracellular bacteria are found widely in both invertebrates and microeukaryotes. In recent decades we have come to appreciate the impact these organisms can have on their hosts, from mutually benefit to ongoing antagonistic evolutionary arms races. Most of our knowledge comes from a limited number of model case studies, which is unfortunate when it is considered that these biotic interactions have as much impact on an organism's evolution, behaviour, and ecology as pathogens and parasites. For instance, *Buchnera* and a plethora of secondary facultative symbionts that infect aphids and shape the insect's diet, reproduction, and ability to defend itself against natural enemies (Buchner, 1965; Oliver *et al.*, 2003, 2008, 2010; Dunbar *et al.*, 2007; Łukasik *et al.*, 2013; Feng *et al.*, 2019).

The bacterial orders where most members are symbionts, like Rickettsiales and Chlamydiota, have huge unexplored diversity outside of pathogenic representatives. However, most of that information comes from environmental metagenomic studies or PCR screens alone, with the former lacking host association data and the latter any biological information. In many cases, infection is detached from any ecological context. The original hosts of most Chlamydiota for instance are entirely unknown, including species that are pathogenic to vertebrates (Horn, 2008; Pillonel, Bertelli and Greub, 2018; Halter *et al.*, 2022). Rickettsiales are often screened for, so their putative hosts are better described (Guo *et al.*, 2016; Thongprem *et al.*, 2020; Pilgrim *et al.*, 2021), but very few studies attempt to associate any real-world environmental data with them (Tsuchida *et al.*, 2002). This leaves Rickettsiales' niche almost entirely undescribed outside of their host and, sporadically, their effects in laboratory conditions.

As discussed in Chapter 1, I believe it is unlikely for widespread, heritable infections like those observed in the family Rickettsiaceae to have no effect on their host or vice versa. Rates of infection by *Rickettsia*, '*Ca.* Tisiphia', and '*Ca.* Megaira' vary within species over time and space. Symbiont often seem to either become fixed or lost in populations, which leads me to believe that these bacteria have context dependant interactions with their host and may only be maintained under the correct environmental pressures, as seen in other species (Oliver *et al.*, 2008; Corbin *et al.*, 2017). In particular, *Rickettsia* infections

are common in the natural environment, but are often lost when the host is taken into the laboratory and the symbiont not directly selected (Zélé *et al.*, 2020). Intermediate levels of infection do exist, for instance protective phenotypes, which in the absence of a threat can be almost as costly as the threat itself (Sumida *et al.*, 2017). Most knowledge of *Rickettsia,* '*Ca.* Tisiphia', and '*Ca*. Megaira' distribution is based only on 16S rRNA PCR results, for which no microscopy has been done, or on single individuals that are not representative of populations (Küchler, Kehl and Dettner, 2009; Guo *et al.*, 2016; Lanzoni *et al.*, 2019; Thongprem *et al.*, 2020; Pilgrim *et al.*, 2021). We currently have no way of predicting the rates of false positives and negatives for PCR screens across host species or the rates of cobiont infection. However, considering the cosmopolitan nature of their infection and the phenotypes seen in within *Rickettsia* and their relatives *Wolbachia*, I believe it is unlikely that they persist without impact across the many hundreds of organisms they infect. However, before we can properly begin investigating the effects of these more obscure symbionts, we must first understand what species are found where, how they've evolved, and what their potential is as models for studying symbioses. In thesis I have used both broad and fine scale methods to begin probing the distributions, phylogenies, and potential phenotypes of both Chlamydiota and Rickettsiales.

Using bioinformatic approaches, in Chapter 2 I expand the available genomic resources available for more obscure symbiont groups. In particular, I report the first genomes for '*Ca.* Megaira', and I find that the former Torix group *Rickettsia* is sufficiently diverged from the genus *Rickettsia* to be considered its own genus, '*Ca.* Tisiphia'. Infections were described across several orders of invertebrates mirroring the diversity found previous studies (Pilgrim *et al.*, 2021). Rickettsiaceae seem to be particularly common in beetles (Coleoptera), though this could be reflective of the biases within invertebrate genomes deposited on the NCBI databases more than real distributions. Two different species of *Ca*. Tisiphia were also found in the stonefly genus *Sericostoma*, which could hint at some relationship between '*Ca*. Tisiphia' and *Sericostoma.* In general, this chapter establishes the broad genomic diversity of this clade, alongside diversity of hosts. It helps clarify the evolution of more obscure groups in the *Rickettsia* and '*Ca*. Megaira' clades, constructing both complete and draft genomes, previously represented largely by 16S rRNA sequence

alone. I also found the potential for nutritional symbiosis in the form of biotin synthesis in one novel genome for the Rhyzobius group *Rickettsia*.

While intracellular symbionts are arguably better studied in invertebrates, it is likely that both Rickettsiales and Chlamydiota originated in microeukaryotes (Driscoll *et al.*, 2013; Schrallhammer *et al.*, 2013). Therefore, in Chapter 3, now armed with better defined Rickettsiaceae phylogenies, I extract further '*Ca.* Megaira' genomes from Algae and Ciliates and identify existing unclassified environmental MAGs that belong to this genus. I find that they follow similar patterns in repeat regions as seen in *Wolbachia* (Siozios *et al.*, 2013; Rice, Sheehan and Newton, 2017), and that they have the potential for protective symbiosis based on their metabolic potential. No other Rickettsiales were extracted from algae and ciliate SRA data in this way, but the available genomes for known '*Ca*. Megaira' have been substantially expanded. These data also indicate this genus is hyperdiverse and should in reality be a bacterial Family, and that we probably do not yet appreciate the diversity within it.

I find that Chlamydiota was equally as common as Rickettsiales in Algae and Ciliate SRA samples from NCBI, and more common in environmental MAGs. All the genomes I assemble or find belong to the relatively newly described Parachlamydiales (Gupta *et al.*, 2015). The available phylogenomic information for the Parachlamydiales is less substantial than Rickettsiaceae. Many Parachlamydiales are not associated with a host, and most are not identified beyond family. As for Chapter 3, Chapter 4 characterises several novel potential host organisms for Chlamydiota amongst microeukaryotes and allow a new genera to be defined that should aid future phylogenomic efforts. Parachlamydiales have the potential to be protective symbionts (Arthofer *et al.*, 2022). Additionally, annotation suggests that several possess CRISPRcas systems. They might also be involved with nutritional symbiosis through B vitamin pathways or plant associated Phylloquinone pathways. However, it is uncertain from metabolic prediction alone whether these pathways are maintained for the environmental stage of Parachlamydiales biphasic lifestyle (König *et al.*, 2017) or are involved host interactions as symbionts or as pathogens.

Chapters 2, 3 and 4 begin to explore previously overlooked infections of commonly symbiotic bacterial groups in ways that should enable future research on phenotype and

ecology. All three are broad scale approaches that leverage a variety of short-read and long-read genomic data. Short-read data will be more error prone and less accurate than newer methods as fragmented reads make it hard to delineate repetitive regions on genomes (sequencing methods reviewed in Giani *et al.*, 2020). Long-read sequencing is generally better for assembling complete symbiont genomes than short-read sequencing because whole genomes can be characterised in a single contig (as done in Chapter 2 and 3), but it is significantly more expensive and there are still issues with the reliability of the technology. Additionally, all data submitted to online databases is variable in quality and type. Quality thresholds and curation of bacterial assemblies are needed enable trustworthy comparisons between any new or old assemblies used.

While a broad approach is useful for defining evolutionary relationships and establishing potential study systems, it does not allow for as much nuance as a single organism model might allow. It does not show the real-world relationship between host and putative symbiont. Relationships change from broad to fine scale. In biogeography for instance, species models vary drastically depending on the scale of the study (Götzenberger *et al.*, 2012; Mod *et al.*, 2020). Local scale studies can show the exact opposite patterns to global scale ones, so it is always important to explore more specific case studies. In the case of Chapters 2, 3 and 4, the assembled genomes are real, but we cannot be sure that any of them are true heritable symbiont infections of the SRA source organism until they are examined in directed studies. This an especially important consideration for microeukaryotes where it is often difficult to achieve truly pure cultures for genome sequencing.

In contrast to previous chapters, Chapter 5 provides real world context for endosymbiont infection. I examine '*Ca*. Tisiphia' infection in *An. plumbeus* mosquitoes, which have previously only been described by a limited number of marker sequences for two individuals from a UK population (Pilgrim *et al.*, 2021). I confirmed that the infection is fixed in the population throughout Germany on a spatial and temporal scale, and that it is inherited through the maternal germline. I also found no clear relation between infection rates and environmental factors (temperature, precipitation, or forest type) recovered from climate and land use databases. This infection is nearly fixed in all the

populations tested, making it unsurprising that no environmental effects were discovered with these methods.

The hosts of the genomes I have assembled in this thesis include species that are medically and economically important. By default, that makes them desirable as model study systems as there is drive, motivation and funding to better understand how these species interact with the environment:

- Brown algae like *Nemacystus decipiens* (ito-mozuku) and *Saccharina japonica* (kombu) are an important staples of edible seaweed agriculture. In particular, *S. japonica* is cultivated in China, Japan, Korea, Russia and France; it is used to produce alginates and is one of the most consumed seaweeds in East Asia (Ye *et al.*, 2015).
- The '*Ca*. Tisiphia' infection in the mosquito *Anopheles plumbeus* provides a novel opportunity to study a native symbiont in anopheline mosquitoes. *Wolbachia* symbionts have previously been found to interfere with the vector competence of mosquitoes, but this was an artificial infection and only works for *Aedes* species (Walker *et al.*, 2011).
- Microalgae have 1/10th the biomass of terrestrial plants and can be 10-50 times more efficient at carbon sequestration (Batista *et al.*, 2015; Onyeaka *et al.*, 2021). There is great interest in using algae in carbon capture technology and bioenergy (Onyeaka *et al.*, 2021), Understanding the basic biology and interactions of various algae is vital in creating efficient and long-lasting systems. The interference of symbionts could influence how well, or poorly, different algae or combinations of algae perform.

Examining diversity of symbionts with genomic data provides its own set of challenges. Whole genome sequences give a clearer picture of relationships between bacteria are, compared to highly conserved gene regions which may underestimate diversity. However, making phylogenies with 16S rRNA is still generally more accessible than multigene trees and is the standard across most bacterial taxonomy. Unfortunately, 16S rRNA phylogenies cannot always corroborate whole genome phylogenies because binning algorithms often do not assemble marker sequences for MAGs (Yuan *et al.*, 2015). This difference means

that the most common marker sequences like 16S rRNA – often the only marker sequence for rare bacteria – form parallel phylogenies with MAGs, as we see in Chapter 3. The 16S rRNA phylogeny for '*Ca.* Megaira' would suggest that individuals are far more closely related compared to multigene phylogenies (Figure 3.1 versus Figure 3.2). In a eukaryotic species, the deep divergence shown by '*Ca.* Megaira' would justify the erection of a new phylum. The disparity between 16S rRNA and whole genome phylogeny results in MAGs remaining unclassified in databases because they cannot be resolved with traditional identification methods.

One other difficulty is that hosts often cannot be resolved from environmental data or for pooled single celled organisms. In general, I have demonstrated a massively unexplored diversity in multiple genera and families of symbionts ('*Ca*. Tisiphia', '*Ca*. Megaira', Rhabdochlamydiaceae and Simkaniaceae). However, bacterial taxonomy limits my ability to properly resolve many of these groups. Several are from environmental data for which I cannot resolve a single host, so a meaningful genera name cannot be usefully assigned. Single cell genomics would have to be used to properly confirm infection to particular host species. Even for multicellular hosts, microscopy needs to be more widely adopted so the evidence of infection is more than a band on a gel.

In addition to difficulties with parallel phylogenies, the cladistic rules for identifying obscure bacteria are simultaneously extremely limiting and not limiting enough. On the one hand genera and species cannot be officially named without culturing and profiling which is impossible for many obligate intracellular bacteria. On the other hand, '*Candidatus'* families, genera and species can be named by anyone without limit and are often not updated on online databases because they are not deemed "official". Unfortunately, *Candidatus* names have the potential to overlap. For instance in Parachlamydiales, one clade could be classified under '*Ca*. Limichlamydiaceae' or '*Ca*. Anoxychlamydiales' depending on which paper you reference (Pillonel, Bertelli and Greub, 2018; Dharamshi *et al.*, 2020). They rely on authors to be aware of all current literature about their chosen bacteria which might span disciplines from microbiology to geology (Huang *et al.*, 2021; Sabuda *et al.*, 2021), as well as current MAGs or markers, which may not even be named when they are deposited on online databases (e.g. GTDB, LPSN and NCBI). Shortfalls in nomenclature leaves us with intensely paraphyletic families

of bacteria, as I observe for Rickettsiaceae, Rhabdochlamydiaceae and Simkaniaceae in this thesis. The naming system needs urgent revision to properly accommodate the continuously expanding repertoire of bacteria, especially those that are not culturable. Over 2/3rds of the bacteria in GTDB only have place holder names that are not informative or memorable (Pallen, Telatin and Oren, 2021). While efforts are being made to automatically generate Latin names with programmes like Gan (Pallen, Telatin and Oren, 2021), these methods have yet to be fully adopted. In my opinion, a genus of bacterialy symbionts should be supported by at least two genomes sharing >65% AAI similarity, or be associated with a host in order to be named. Good quality MAGs are a reliable and relatively non-fussy way of identifying bacteria, and – I would argue – a far better metric than traditional chemical and morphological assessments that still take priority.

6.2 Final perspectives

This thesis shows the massive diversity of both bacterial endosymbionts and their putative hosts. Many symbionts are only known by their 16S rRNA sequences and have no reference for how they interact with their host or the environment. I clarify the evolution of '*Ca*. Megaira', '*Ca*. Tisiphia' and some Parachlamydiales. I also highlight the importance of examining symbionts with relation to their real-world environments and their hosts.

Results across chapters show that infections with obscure symbionts are unlikely to exist without impact on their hosts. Bioinformatic analysis predicts potential for nutritional symbiosis and defensive symbiosis in several new species. In addition, I have shown that '*Ca*. Tisiphia' can be maternally inherited in *Anopheles plumbeus* mosquitos. Because of this, they are inextricably linked with their host's survival and reproduction. '*Ca*. Tisiphia' and, by extension, *Rickettsia* could provide potential research alternatives to *Wolbachia* in Anopheline vector species where transinfections with symbionts have been unsuccessful.

Overall, this thesis helps elucidate the evolution of several understudied endosymbiont species as well as some potentially important host-symbiont interactions. I show that the level of diversity in symbionts is likely underestimated and lay the groundwork for future research to explore new study systems. Obscure bacterial endosymbionts are globally common and are unlikely to exist without impact on their hosts ecology and evolution.

References

Abatzoglou, J.T. *et al.* (2018) 'TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015', *Scientific Data*, 5(1), p. 170191. Available at: https://doi.org/10.1038/sdata.2017.191.

Aguin-Pombo, D. *et al.* (2021) 'Parthenogenesis and sex-ratio distorting bacteria in *Empoasca* (Hemiptera: Cicadellidae) leafhoppers', *Annals of the Entomological Society of America* [Preprint]. Available at: https://doi.org/10.1093/AESA/SAAB025.

Altinli, M. *et al.* (2018) '*Wolbachia* diversity and cytoplasmic incompatibility patterns in *Culex pipiens* populations in Turkey', *Parasites & Vectors*, 11(1), p. 198. Available at: https://doi.org/10.1186/s13071-018-2777-9.

Andersson, S.G.E. *et al.* (1998) 'The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria', *Nature*, 396(6707), pp. 133–140. Available at: https://doi.org/10.1038/24094.

Angelakis, E. and Raoult, D. (2017) '*Rickettsia* and *Rickettsia*-like organisms', in *Infectious Diseases*. Elsevier, pp. 1666-1675.e1. Available at: https://doi.org/10.1016/b978-0-7020-6285-8.00187-8.

Aramaki, T. *et al.* (2020) 'KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold', *Bioinformatics*. Edited by A. Valencia, 36(7), pp. 2251–2252. Available at: https://doi.org/10.1093/bioinformatics/btz859.

Aravind, L. *et al.* (2015) 'The natural history of ADP-Ribosyltransferases and the ADP-ribosylation system', *Curr Top Microbiol Immunol.*, 384, pp. 3–32. Available at: https://doi.org/10.1007/82_2014_414.

Arthofer, P. *et al.* (2022) 'Defensive symbiosis against giant viruses in amoebae', *Proceedings of the National Academy of Sciences*, 119(36). Available at: https://doi.org/10.1073/pnas.2205856119.

Babraham Bioinformatics (2019) 'Babraham Bioinformatics - FastQC A quality control tool for high throughput sequence data'. Babraham Bioinformatics. Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (Accessed: 20 December 2022).

Bastian, M., Heymann, S. and Jacomy, M. (2009) 'Gephi : An open source software for exploring and manipulating networks visualization and exploration of large graphs', *Proceedings of the International AAAI Conference on Web and Social Media*, 3(1), pp. 361–362.

Bastidas, R.J. *et al.* (2013) 'Chlamydial intracellular survival strategies', *Cold Spring Harbor Perspectives in Medicine*, 3(5), pp. a010256–a010256. Available at: https://doi.org/10.1101/cshperspect.a010256.

Batista, A.P. *et al.* (2015) 'Combining urban wastewater treatment with biohydrogen production – An integrated microalgae-based approach', *Bioresource Technology*, 184, pp. 230–235. Available at: https://doi.org/10.1016/j.biortech.2014.10.064.

Bayramova, F., Jacquier, N. and Greub, G. (2018) 'Insight in the biology of *Chlamydia*-related bacteria', *Microbes and Infection*, 20(7–8), pp. 432–440. Available at: https://doi.org/10.1016/j.micinf.2017.11.008.

van der Beek, S.L. *et al.* (2019) 'Streptococcal dTDP-L-rhamnose biosynthesis enzymes: functional characterization and lead compound identification', *Molecular Microbiology*, 111(4), pp. 951–964. Available at: https://doi.org/10.1111/mmi.14197.

Bertelli, C. *et al.* (2016) 'CRISPR system acquisition and evolution of an obligate intracellular *Chlamydia*-related bacterium', *Genome Biology and Evolution*, 8(8), pp. 2376–2386. Available at: https://doi.org/10.1093/GBE/EVW138.

Biggs, H.M. *et al.* (2016) 'Diagnosis and management of tickborne rickettsial diseases: rocky mountain spotted fever and other spotted fever group Rickettsioses, Ehrlichioses, and Anaplasmosis — United States', *MMWR. Recommendations and Reports*, 65(2), pp. 1–44. Available at: https://doi.org/10.15585/mmwr.rr6502a1.

Blin, K. *et al.* (2021) 'antiSMASH 6.0: improving cluster detection and comparison capabilities', *Nucleic Acids Research*, 49(W1), pp. W29–W35. Available at: https://doi.org/10.1093/nar/gkab335.

Blow, F. (2017) *Variation in the structure and function of invertebrate-associated bacterial communities*. University of Liverpool. Available at: https://doi.org/10.17638/03009325.

Blow, F. *et al.* (2020) 'B-vitamin nutrition in the pea aphid-*Buchnera* symbiosis', *Journal of Insect Physiology*, 126, p. 104092. Available at: https://doi.org/10.1016/j.jinsphys.2020.104092.

Bockoven, A.A. *et al.* (2020) 'What goes up might come down: the spectacular spread of an endosymbiont is followed by its decline a decade later', *Microbial Ecology*, 79(2), pp. 482–494. Available at: https://doi.org/10.1007/s00248-019-01417-4.

Bodnar, J.L. *et al.* (2018) 'The folA gene from the *Rickettsia* endosymbiont of *Ixodes pacificus* encodes a functional dihydrofolate reductase enzyme', *Ticks and Tick-borne Diseases*, 9(3), pp. 443–449. Available at: https://doi.org/10.1016/j.ttbdis.2017.12.013.

Boettcher, K.J., Ruby, E.G. and McFall-Ngai, M.J. (1996) 'Bioluminescence in the symbiotic squid *Euprymna scolopes* is controlled by a daily biological rhythm', *Journal of Comparative Physiology A*, 179(1), pp. 65–73. Available at: https://doi.org/10.1007/BF00193435.

Bolger, A.M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: a flexible trimmer for Illumina sequence data', *Bioinformatics (Oxford, England)*, 30(15), pp. 2114–2120. Available at: https://doi.org/10.1093/bioinformatics/btu170.

Boyd, B.M. *et al.* (2016) 'Two bacterial genera, *Sodalis* and *Rickettsia*, associated with the seal louse *Proechinophthirus fluctus* (Phthiraptera: Anoplura)', *Applied and Environmental Microbiology*, 82(11), pp. 3185–3197. Available at: https://doi.org/10.1128/AEM.00282-16/SUPPL_FILE/ZAM999117155SO1.PDF.

Breitwieser, F.P., Lu, J. and Salzberg, S.L. (2017) 'A review of methods and databases for metagenomic classification and assembly', *Briefings in Bioinformatics* [Preprint]. Available at: https://doi.org/10.1093/bib/bbx120.

Bressan, A. (2014) 'Emergence and evolution of *Arsenophonus* bacteria as insect-vectored plant pathogens', *Infection, Genetics and Evolution*, 22, pp. 81–90. Available at: https://doi.org/10.1016/j.meegid.2014.01.004.

Brinkman, F.S.L. *et al.* (2002) 'Evidence that plant-like genes in chlamydia species reflect an ancestral relationship between Chlamydiaceae, *Cyanobacteria*, and the Chloroplast', *Genome Research*, 12(8), pp. 1159–1167. Available at: https://doi.org/10.1101/gr.341802.

Broad Institute (2013) 'DISCOVAR'. Available at: https://software.broadinstitute.org/software/discovar/blog/?page_id=14.

Bruen, T.C., Philippe, H. and Bryant, D. (2006) 'A simple and robust statistical test for detecting the presence of recombination', *Genetics*, 172(4), pp. 2665–2681. Available at: https://doi.org/10.1534/genetics.105.048975.

van Bruggen, J.J.A., Stumm, C.K. and Vogels, G.D. (1983) 'Symbiosis of methanogenic bacteria and sapropelic protozoa', *Archives of Microbiology*, 136(2), pp. 89–95. Available at: https://doi.org/10.1007/BF00404779.

Brumin, M., Kontsedalov, S. and Ghanim, M. (2011) '*Rickettsia* influences thermotolerance in the whitefly *Bemisia tabaci* B biotype', *Insect Science*, 18(1), pp. 57–66. Available at: https://doi.org/10.1111/j.1744-7917.2010.01396.x.

Buchner, P. (1965) *Endosymbiosis of animals with plant microorganisms*. New York: Interscience Publishers.

Bueno-Marí, R. and Jiménez-Peydró, R. (2011) '*Anopheles plumbeus* Stephens, 1828: a neglected malaria vector in Europe', *Malaria Reports*, 1(1), p. 2. Available at: https://doi.org/10.4081/malaria.2011.e2.

Bushnell, B. (2015) *BBMap*, *https://sourceforge.net/projects/bbmap/*.

Camacho, C. *et al.* (2009) 'BLAST+: Architecture and applications', *BMC Bioinformatics*, 10. Available at: https://doi.org/10.1186/1471-2105-10-421.

Carrier, T.J. *et al.* (2022) 'Symbiont transmission in marine sponges: reproduction, development, and metamorphosis', *BMC Biology*, 20(1), p. 100. Available at: https://doi.org/10.1186/s12915-022-01291-6.

Cass, B.N. *et al.* (2016) 'Conditional fitness benefits of the *Rickettsia* bacterial symbiont in an insect pest', *Oecologia*, 180(1), pp. 169–179. Available at: https://doi.org/10.1007/s00442-015-3436-x.

Castelli, M. *et al.* (2019) '*Deianiraea*, an extracellular bacterium associated with the ciliate *Paramecium*, suggests an alternative scenario for the evolution of Rickettsiales', *The ISME Journal*, 13(9), pp. 2280–2294. Available at: https://doi.org/10.1038/s41396-019-0433-9.

Castelli, M. *et al.* (2021) '"*Candidatus* Sarmatiella mevalonica" endosymbiont of the ciliate *Paramecium* provides insights on evolutionary plasticity among Rickettsiales', *Environmental Microbiology*, pp. 1462-2920.15396. Available at: https://doi.org/10.1111/1462-2920.15396.

Charlat, S., Hurst, G.D.D. and Merçot, H. (2003) 'Evolutionary consequences of *Wolbachia* infections', *Trends in Genetics*, 19(4), pp. 217–223. Available at: https://doi.org/10.1016/S0168-9525(03)00024-6.

Chaumeil, P.-A. *et al.* (2020) 'GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database', *Bioinformatics*, 36(6), pp. 1925–1927. Available at: https://doi.org/10.1093/BIOINFORMATICS/BTZ848.

Chen, Yuxin *et al.* (2018) 'SOAPnuke: A MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data', *GigaScience*, 7(1), pp. 1–6. Available at: https://doi.org/10.1093/gigascience/gix120.

Chen, Z. *et al.* (2022) 'Indicator species drive the key ecological functions of microbiota in a river impacted by acid mine drainage generated by rare earth elements mining in South China', *Environmental Microbiology*, 24(2), pp. 919–937. Available at: https://doi.org/10.1111/1462-2920.15501.

Chiel, E., Zchori-Fein, E., *et al.* (2009) 'Almost there: transmission routes of bacterial symbionts between trophic levels', *PLoS ONE*. Edited by J.E. Stajich, 4(3), p. e4767. Available at: https://doi.org/10.1371/journal.pone.0004767.

Chiel, E., Inbar, M., *et al.* (2009) 'Assessments of fitness effects by the facultative symbiont *Rickettsia* in the sweetpotato whitefly (Hemiptera: Aleyrodidae)', *Annals of the Entomological Society of America*, 102(3), pp. 413–418. Available at: https://doi.org/10.1603/008.102.0309.

Choi, J.Y. *et al.* (1997) 'Evidence for symbiont-induced alteration of a host's gene expression: irreversible loss of SAM synthetase from *Amoeba proteus*', *The Journal of Eukaryotic Microbiology*, 44(5), pp. 412–419. Available at: https://doi.org/10.1111/j.1550-7408.1997.tb05717.x.

Christodoulou, D. *et al.* (2018) 'Reserve flux capacity in the pentose phosphate pathway enables *Escherichia coli*'s rapid response to oxidative stress', *Cell Systems*, 6, pp. 569–578. Available at: https://doi.org/10.1016/j.cels.2018.04.009.

Clay, K., Holah, J. and Rudgers, J.A. (2005) 'Herbivores cause a rapid increase in hereditary symbiosis and alter plant community composition', *Proceedings of the National Academy of Sciences*, 102(35), pp. 12465–12470. Available at: https://doi.org/10.1073/pnas.0503059102.

Collingro, A., Köstlbacher, S. and Horn, M. (2020) 'Chlamydiae in the Environment', *Trends in Microbiology*, 28(11), pp. 877–888. Available at: https://doi.org/10.1016/j.tim.2020.05.020.

Comandatore, F. *et al.* (2015) 'Supergroup C *Wolbachia* , mutualist symbionts of filarial nematodes, have a distinct genome structure', *Open Biology*, 5(12), p. 150099. Available at: https://doi.org/10.1098/rsob.150099.

Corbin, C. *et al.* (2017) 'Heritable symbionts in a world of varying temperature', *Heredity*, 118(1), pp. 10–20. Available at: https://doi.org/10.1038/hdy.2016.71.

Couvin, D. *et al.* (2018) 'CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins', *Nucleic Acids Research*, 46(W1), pp. W246–W251. Available at: https://doi.org/10.1093/nar/gky425.

Dale, C. and Moran, N.A. (2006) 'molecular interactions between bacterial symbionts and their hosts', *Cell*, 126(3), pp. 453–465. Available at: https://doi.org/10.1016/j.cell.2006.07.014.

Dally, M. *et al.* (2020) 'Cellular localization of two *rickettsia* symbionts in the digestive system and within the ovaries of the mirid bug, *Macrolophous pygmaeus*', *Insects*, 11(8), p. 530. Available at: https://doi.org/10.3390/insects11080530.

Dangeard, P. (1896) 'Contribution à l'étude des Acrasiées', *Botaniste*, 5, pp. 1–20.

Daniels, E.W. and Breyer, E.P. (1967) 'Ultrastructure of the giant amoeba *Pelomyxa palustris*', *The Journal of Protozoology*, 14(1), pp. 167–179. Available at: https://doi.org/10.1111/j.1550-7408.1967.tb01463.x.

Daniels, R., Vanderleyden, J. and Michiels, J. (2004) 'Quorum sensing and swarming migration in bacteria', *FEMS Microbiology Reviews*, 28(3), pp. 261–289. Available at: https://doi.org/10.1016/j.femsre.2003.09.004.

Davison, H.R. *et al.* (2022) 'Genomic diversity across the *Rickettsia* and "*Candidatus* Megaira" genera and proposal of genus status for the Torix group', *Nature Communications*, 13(1), p. 2630. Available at: https://doi.org/10.1038/s41467-022-30385-6.

Davison, H.R. (2022) 'VibrantStarling/Code-used-to-extract-bacterial-genomes-from-invertebrate-genomes: SRA-dive v1.0.0'. Available at: https://doi.org/10.5281/zenodo.6396821.

De Coster, W. *et al.* (2018) 'NanoPack: Visualizing and processing long-read sequencing data', *Bioinformatics*, 34(15), pp. 2666–2669. Available at: https://doi.org/10.1093/bioinformatics/bty149.

Dekoninck, W. *et al.* (2011) 'Human-induced expanded distribution of *Anopheles plumbeus*, experimental vector of west nile virus and a potential vector of human malaria in belgium', *Journal of Medical Entomology*, 48(4), pp. 924–928. Available at: https://doi.org/10.1603/ME10235.

Deveau, H., Garneau, J.E. and Moineau, S. (2010) 'CRISPR/Cas system and its role in phage-bacteria interactions', *Annual Review of Microbiology*, 64(1), pp. 475–493. Available at: https://doi.org/10.1146/annurev.micro.112408.134123.

Dharamshi, J.E. *et al.* (2020) 'Marine sediments illuminate Chlamydiae diversity and evolution', *Current Biology*, 30(6), pp. 1032-1048.e7. Available at: https://doi.org/10.1016/j.cub.2020.02.016.

Dietrich, M.R. *et al.* (2020) 'How to choose your research organism', *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 80, p. 101227. Available at: https://doi.org/10.1016/j.shpsc.2019.101227.

Dosselaere, F. and Vanderleyden, J. (2001) 'A metabolic node in action: Chorismate-utilizing enzymes in microorganisms', *Critical Reviews in Microbiology*, 27(2), pp. 75–131. Available at: https://doi.org/10.1080/20014091096710.

Doudoumis, V. *et al.* (2017) 'Challenging the *Wigglesworthia , Sodalis , Wolbachia* symbiosis dogma in tsetse flies: *Spiroplasma* is present in both laboratory and natural populations', *Scientific Reports 2017 7:1*, 7(1), pp. 1–13. Available at: https://doi.org/10.1038/s41598-017-04740-3.

Douglas, A.E. (2011) 'Lessons from studying insect symbioses', *Cell Host and Microbe*, 10(4), pp. 359–367. Available at: https://doi.org/10.1016/j.chom.2011.09.001.

Douglas, A.E. (2017) 'The B vitamin nutrition of insects: the contributions of diet, microbiome and horizontally acquired genes', *Current Opinion in Insect Science*, 23, pp. 65–69. Available at: https://doi.org/10.1016/j.cois.2017.07.012.

Draghi, A. *et al.* (2004) 'Characterization of '*Candidatus* Piscichlamydia salmonis' (Order Chlamydiales), a *Chlamydia*-Like bacterium associated with epitheliocystis in farmed atlantic salmon (*Salmo salar*)', *Journal of Clinical Microbiology*, 42(11), pp. 5286–5297. Available at: https://doi.org/10.1128/JCM.42.11.5286-5297.2004.

Drew, G.C. *et al.* (2021) 'Transitions in symbiosis: evidence for environmental acquisition and social transmission within a clade of heritable symbionts', *The ISME Journal*, 15(10), pp. 2956–2968. Available at: https://doi.org/10.1038/s41396-021-00977-z.

Driscoll, T. *et al.* (2013) 'Bacterial DNA sifted from the *Trichoplax adhaerens* (Animalia: Placozoa) genome project reveals a putative rickettsial endosymbiont', *Genome Biology and Evolution*, 5(4), pp. 621–645. Available at: https://doi.org/10.1093/GBE/EVT036.

Driscoll, T. *et al.* (2017) 'Wholly *Rickettsia*! Reconstructed Metabolic Profile of the Quintessential Bacterial Parasite of Eukaryotic Cells.', *mBio*, 8(5). Available at: https://doi.org/10.1128/mBio.00859-17.

Du, Y., Maslov, D.A. and Chang, K.P. (1994) 'Monophyletic origin of beta-division proteobacterial endosymbionts and their coevolution with insect trypanosomatid protozoa *Blastocrithidia culicis* and *Crithidia* spp.', *Proceedings of the National Academy of Sciences*, 91(18), pp. 8437–8441. Available at: https://doi.org/10.1073/pnas.91.18.8437.

Dunbar, H.E. *et al.* (2007) 'Aphid thermal tolerance is governed by a point mutation in bacterial symbionts', *PLoS Biology*, 5(5), pp. 1006–1015. Available at: https://doi.org/10.1371/journal.pbio.0050096.

Duncan, A.B. *et al.* (2010) 'Parasite-mediated protection against osmotic stress for *Paramecium caudatum* infected by *Holospora undulata* is host genotype specific', *FEMS Microbiology Ecology*, 74(2), pp. 353–360. Available at: https://doi.org/10.1111/j.1574-6941.2010.00952.x.

Duplouy, A. and O'Neill, S.L. (2010) 'Male-killing *Wolbachia* in the butterfly *Hypolimnas bolina*', in P. Pontarotti (ed.) *Evolutionary Biology -- Concepts, Molecular and Morphological Evolution: 13th Meeting 2009*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 209–227. Available at: https://doi.org/10.1007/978-3-642-12340-5_13.

Duron, O. *et al.* (2008) 'The diversity of reproductive parasites among arthropods: *Wolbachia* do not walk alone', *BMC Biology*, 6(1), p. 27. Available at: https://doi.org/10.1186/1741-7007-6-27.

Dyková, I. *et al.* (2013) '*Nuclearia pattersoni* sp. n. (Filosea), a new species of amphizoic amoeba isolated from gills of roach (*Rutilus rutilus*), and its rickettsial endosymbiont', *Folia Parasitologica*, 50(3), pp. 161–170. Available at: https://doi.org/10.14411/fp.2003.030.

Eddy, S.R. (2011) 'Accelerated profile HMM searches', *PLoS Computational Biology*. Edited by W.R. Pearson, 7(10), p. e1002195. Available at: https://doi.org/10.1371/journal.pcbi.1002195.

Eddy, S.R. (2018) 'HMMER 3.2.1'. Howard Hughes Medical Institute. Available at: http://hmmer.org/.

Edgar, R.C. (2004) 'MUSCLE: A multiple sequence alignment method with reduced time and space complexity', *BMC Bioinformatics*, 5(1), pp. 1–19. Available at: https://doi.org/10.1186/1471-2105-5-113.

Engelstädter, J. and Hurst, G.D.D. (2009) 'The ecology and evolution of microbes that manipulate host reproduction', *Annual Review of Ecology, Evolution, and Systematics*, 40(1), pp. 127–149. Available at: https://doi.org/10.1146/annurev.ecolsys.110308.120206.

Eren, A.M. *et al.* (2021) 'Community-led, integrated, reproducible multi-omics with Anvi'o', *Nature Microbiology*, 6(1), pp. 3–6. Available at: https://doi.org/10.1038/s41564-020-00834-3.

European Union (2018) *Copernicus Land Monitoring Service*, *European Environment Agency (EEA)*. Available at: https://land.copernicus.eu/pan-european/high-resolution-layers/forests/forest-type-1/status-maps/forest-type-2018 (Accessed: 31 August 2022).

Feng, H. *et al.* (2019) 'Trading amino acids at the aphid–*Buchnera* symbiotic interface', *Proceedings of the National Academy of Sciences of the United States of America*, 116(32), pp. 16003–16011. Available at: https://doi.org/10.1073/pnas.1906223116.

Feng, L., Shou, Q. and Butcher, R.A. (2016) 'Identification of a dTDP-rhamnose biosynthetic pathway that oscillates with the molting cycle in *Caenorhabditis elegans*', *Biochemical Journal*, 473(11), pp. 1507–1521. Available at: https://doi.org/10.1042/BCJ20160142.

Ferree, P.M. *et al.* (2005) '*Wolbachia* utilizes host microtubules and dynein for anterior localization in the *Drosophila* oocyte', *PLOS Pathogens*, 1(2), p. e14. Available at: https://doi.org/10.1371/journal.ppat.0010014.

Fisher, D.J. *et al.* (2012) 'Uptake of biotin by *Chlamydia* Spp. through the use of a bacterial transporter (BioY) and a host-cell transporter (SMVT)', *PLoS ONE*, 7(9), p. e46052. Available at: https://doi.org/10.1371/journal.pone.0046052.

Flissi, A. *et al.* (2019) 'Norine: update of the nonribosomal peptide resource', *Nucleic Acids Research*, 48(D1), pp. D465–D469. Available at: https://doi.org/10.1093/nar/gkz1000.

Flórez, L.V. *et al.* (2017) 'Antibiotic-producing symbionts dynamically transition between plant pathogenicity and insect-defensive mutualism', *Nature Communications*, 8(1), p. 15172. Available at: https://doi.org/10.1038/ncomms15172.

Folmer, O. *et al.* (1994) 'DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates', *Molecular Marine Biology and Biotechnology*, 3(5), pp. 294–299.

French, K.E. (2017) 'Engineering mycorrhizal symbioses to alter plant metabolism and improve crop health', *Frontiers in Microbiology*, 8. Available at: https://doi.org/10.3389/fmicb.2017.01403.

Fry, A.J., Palmer, M.R. and Rand, D.M. (2004) 'Variable fitness effects of *Wolbachia* infection in *Drosophila melanogaster*', *Heredity*, 93(4), pp. 379–389. Available at: https://doi.org/10.1038/sj.hdy.6800514.

Fujishima, M. and Kodama, Y. (2012) 'Endosymbionts in *Paramecium*', *European Journal of Protistology*, 48(2), pp. 124–137. Available at: https://doi.org/10.1016/j.ejop.2011.10.002.

Fuxelius, H.-H. *et al.* (2007) 'The genomic and metabolic diversity of *Rickettsia*', *Research in Microbiology*, 158(10), pp. 745–753. Available at: https://doi.org/10.1016/j.resmic.2007.09.008.

Galindo, L.J. *et al.* (2019) 'Combined cultivation and single-cell approaches to the phylogenomics of nucleariid amoebae, close relatives of fungi', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1786), p. 20190094. Available at: https://doi.org/10.1098/rstb.2019.0094.

Galperin, M.Y. *et al.* (2021) 'COG database update: focus on microbial diversity, model organisms, and widespread pathogens', *Nucleic Acids Research*, 49(D1), pp. D274–D281. Available at: https://doi.org/10.1093/nar/gkaa1018.

Giani, A.M. *et al.* (2020) 'Long walk to genomics: History and current approaches to genome sequencing and assembly', *Computational and Structural Biotechnology Journal*, 18, pp. 9–19. Available at: https://doi.org/10.1016/j.csbj.2019.11.002.

Giannotti, D. *et al.* (2022) 'The "Other" Rickettsiales: an overview of the Family "*Candidatus* Midichloriaceae"', *Applied and Environmental Microbiology*, 88(6), pp. e02432-21. Available at: https://doi.org/10.1128/aem.02432-21.

Gilchrist, C.L.M. and Chooi, Y.-H. (2021) 'Clinker & clustermap.js: automatic generation of gene cluster comparison figures', *Bioinformatics*. Edited by P. Robinson, 37(16), pp. 2473–2475. Available at: https://doi.org/10.1093/bioinformatics/btab007.

Gill, A.C., Darby, A.C. and Makepeace, B.L. (2014) 'Iron necessity: The secret of *Wolbachia's* success?', *PLoS Neglected Tropical Diseases*, 8(10), p. e3224. Available at: https://doi.org/10.1371/journal.pntd.0003224.

Gillespie, J.J. *et al.* (2007) 'Plasmids and rickettsial evolution: insight from *Rickettsia felis*.', *Plos one*, 2(3), pp. e266–e266. Available at: https://doi.org/10.1371/JOURNAL.PONE.0000266.

Gillespie, J.J. *et al.* (2012) 'A *Rickettsia* genome overrun by mobile genetic elements provides insight into the acquisition of genes characteristic of an obligate intracellular lifestyle', *Journal of Bacteriology*, 194(2), pp. 376–394. Available at: https://doi.org/10.1128/JB.06244-11.

Gillespie, J.J. *et al.* (2015) 'Genomic diversification in strains of *Rickettsia felis* isolated from different Arthropods', *Genome Biology and Evolution*, 7(1), pp. 35–56. Available at: https://doi.org/10.1093/gbe/evu262.

Gillespie, J.J. *et al.* (2018) 'A tangled web: Origins of reproductive parasitism', *Genome Biology and Evolution*. Edited by B. Eric, 10(9), pp. 2292–2309. Available at: https://doi.org/10.1093/gbe/evy159.

Giorgini, M. *et al.* (2010) '*Rickettsia* symbionts cause parthenogenetic reproduction in the parasitoid wasp *Pnigalio soemius* (hymenoptera: Eulophidae)', *Applied and Environmental Microbiology*, 76(8), pp. 2589–2599. Available at: https://doi.org/10.1128/AEM.03154-09.

Goffredi, S.K. *et al.* (2020) 'Methanotrophic bacterial symbionts fuel dense populations of deep-sea feather duster worms (Sabellida, Annelida) and extend the spatial influence of methane seepage', *Science Advances*, 6(14), p. eaay8562. Available at: https://doi.org/10.1126/sciadv.aay8562.

Gogarten, J.P. and Townsend, J.P. (2005) 'Horizontal gene transfer, genome innovation and evolution', *Nature Reviews Microbiology*, 3(9), pp. 679–687. Available at: https://doi.org/10.1038/nrmicro1204.

Goh, K.M. *et al.* (2019) 'Current status and potential applications of underexplored Prokaryotes', *Microorganisms*, 7(10), p. 468. Available at: https://doi.org/10.3390/microorganisms7100468.

Götzenberger, L. *et al.* (2012) 'Ecological assembly rules in plant communities—approaches, patterns and prospects', *Biological Reviews*, 87(1), pp. 111–127. Available at: https://doi.org/10.1111/j.1469-185X.2011.00187.x.

Grosser, K. *et al.* (2018) 'More than the "Killer Trait": Infection with the bacterial endosymbiont *Caedibacter taeniospiralis* causes transcriptomic modulation in *Paramecium* host', *Genome Biology and Evolution*, 10(2), pp. 646–656. Available at: https://doi.org/10.1093/gbe/evy024.

Gruber, F., Lipozenčić, J. and Kehler, T. (2015) 'History of venereal diseases from antiquity to the renaissance', *Acta Dermatovenerologica Croatica*, 23(1), pp. 1–11.

Gruber-Vodicka, H.R., Seah, B.K.B. and Pruesse, E. (2020) 'phyloFlash: Rapid small-subunit rRNA profiling and targeted assembly from metagenomes', *mSystems*. Edited by M. Arumugam, 5(5). Available at: https://doi.org/10.1128/mSystems.00920-20.

Guillotte, M.L. *et al.* (2021) 'Lipid a structural divergence in *rickettsia* pathogens', *mSphere*, 6(3). Available at: https://doi.org/10.1128/MSPHERE.00184-21/SUPPL_FILE/MSPHERE.00184-21-SF004.PDF.

Guindon, S. *et al.* (2010) 'New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0', *Systematic Biology*, 59(3), pp. 307–321. Available at: https://doi.org/10.1093/sysbio/syq010.

Guo, W.P. *et al.* (2016) 'Extensive genetic diversity of Rickettsiales bacteria in multiple mosquito species', *Scientific Reports*, 6. Available at: https://doi.org/10.1038/srep38770.

Gupta, R.S. *et al.* (2015) 'A phylogenomic and molecular markers based analysis of the phylum Chlamydiae: proposal to divide the class Chlamydiia into two orders, Chlamydiales and Parachlamydiales ord. nov., and emended description of the class

Chlamydiia', *Antonie van Leeuwenhoek*, 108(3), pp. 765–781. Available at: https://doi.org/10.1007/s10482-015-0532-1.

Gutiérrez, G. *et al.* (2017) 'Identification of *Pelomyxa palustris* Endosymbionts', *Protist*, 168(4), pp. 408–424. Available at: https://doi.org/10.1016/j.protis.2017.06.001.

Hackstein, J.H.P. and Vogels, G.D. (1997) 'Endosymbiotic interactions in anaerobic protozoa', *Antonie van Leeuwenhoek*, 71(1/2), pp. 151–158. Available at: https://doi.org/10.1023/A:1000154526395.

Hagen, R. *et al.* (2018) 'Conjugative transposons and their cargo genes vary across natural populations of *Rickettsia buchneri* infecting the tick *Ixodes scapularis*', *Genome Biology and Evolution*, 10(12), pp. 3218–3229. Available at: https://doi.org/10.1093/GBE/EVY247.

Hagimori, T. *et al.* (2006) 'The first finding of a *Rickettsia* bacterium associated with parthenogenesis induction among insects', *Current Microbiology*, 52(2), pp. 97–101. Available at: https://doi.org/10.1007/s00284-005-0092-0.

Hajibabaei, M. *et al.* (2005) 'Critical factors for assembling a high volume of DNA barcodes', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1462), pp. 1959–1967. Available at: https://doi.org/10.1098/rstb.2005.1727.

Halter, T. *et al.* (2022) 'Ecology and evolution of chlamydial symbionts of arthropods', *ISME Communications*, 2(1), p. 45. Available at: https://doi.org/10.1038/s43705-022-00124-5.

Hamberger, B. *et al.* (2006) 'Comparative genomics of the shikimate pathway in *Arabidopsis, Populus trichocarpa and Oryza sativa*: Shikimate pathway gene family structure and identification of candidates for missing links in phenylalanine biosynthesis', in *Recent Advances in Phytochemistry*. Elsevier, pp. 85–113. Available at: https://doi.org/10.1016/S0079-9920(06)80038-9.

Hansen, R.D.E. *et al.* (2011) 'A worm's best friend: recruitment of neutrophils by *Wolbachia* confounds eosinophil degranulation against the filarial nematode *Onchocerca ochengi*', *Proceedings of the Royal Society B: Biological Sciences*, 278(1716), pp. 2293–2302. Available at: https://doi.org/10.1098/rspb.2010.2367.

Hanski, I. *et al.* (2012) 'Environmental biodiversity, human microbiota, and allergy are interrelated', *Proceedings of the National Academy of Sciences*, 109(21), pp. 8334–8339.

Hawkins, J.P., Ordonez, P.A. and Oresnik, I.J. (2018) 'Characterization of mutations that affect the nonoxidative pentose phosphate pathway in *Sinorhizobium meliloti*', *Journal of Bacteriology*. Edited by A. Becker, 200(2). Available at: https://doi.org/10.1128/JB.00436-17.

Hayashi, M. *et al.* (2016) 'A Nightmare for males? a maternally transmitted male-killing bacterium and strong female bias in a green lacewing population', *PLOS ONE*. Edited by

K. Bourtzis, 11(6), p. e0155794. Available at: https://doi.org/10.1371/journal.pone.0155794.

Hegemann, J.D. *et al.* (2015) 'Lasso peptides: An intriguing class of bacterial natural products.', *Accounts of chemical research*, 48(7), pp. 1909–19. Available at: https://doi.org/10.1021/acs.accounts.5b00156.

Hendry, T.A., Hunter, M.S. and Baltrus, D.A. (2014) 'The facultative symbiont *Rickettsia* protects an invasive whitefly against entomopathogenic *Pseudomonas syringae* strains', *Applied and Environmental Microbiology*, 80(23), pp. 7161–7168. Available at: https://doi.org/10.1128/AEM.02447-14.

Heym, E.C. *et al.* (2017) '*Anopheles plumbeus* (Diptera: Culicidae) in Germany: updated geographic distribution and public health impact of a nuisance and vector mosquito', *Tropical Medicine & International Health*, 22(1), pp. 103–112. Available at: https://doi.org/10.1111/tmi.12805.

Himler, A.G. *et al.* (2011) 'rapid spread of a bacterial symbiont in an invasive whitefly is driven by fitness benefits and female bias', *Science*, 332(6026), pp. 254–256. Available at: https://doi.org/10.1126/science.1199410.

Hoang, D.T. *et al.* (2018) 'UFBoot2: Improving the Ultrafast Bootstrap approximation', *Molecular Biology and Evolution*, 35(2), pp. 518–522. Available at: https://doi.org/10.5281/zenodo.854445.

van Hoek, A.H.A.M. *et al.* (2000) 'Multiple acquisition of methanogenic archaeal symbionts by anaerobic ciliates', *Molecular Biology and Evolution*, 17(2), pp. 251–258. Available at: https://doi.org/10.1093/oxfordjournals.molbev.a026304.

Hoffmann, A.A. *et al.* (2011) 'Successful establishment of *Wolbachia* in *Aedes* populations to suppress dengue transmission', *Nature*, 476(7361), pp. 454–457. Available at: https://doi.org/10.1038/nature10356.

Hollants, J. *et al.* (2013) 'Permanent residents or temporary lodgers: Characterizing intracellular bacterial communities in the siphonous green alga *Bryopsis*', *Proceedings of the Royal Society B: Biological Sciences*, 280(1754), pp. 1–8. Available at: https://doi.org/10.1098/rspb.2012.2659.

Horn, M. *et al.* (2004) 'Illuminating the evolutionary history of Chlamydiae', *Science*, 304(5671), pp. 728–730. Available at: https://doi.org/10.1126/science.1096330.

Horn, M. (2008) 'Chlamydiae as symbionts in eukaryotes', *Annual Review of Microbiology*, 62(1), pp. 113–131. Available at: https://doi.org/10.1146/annurev.micro.62.081307.162818.

Hosokawa, T. *et al.* (2010) '*Wolbachia* as a bacteriocyte-associated nutritional mutualist', *Proceedings of the National Academy of Sciences of the United States of America*, 107(2), pp. 769–774. Available at: https://doi.org/10.1073/pnas.0911476107.

Hotopp, J.C.D. *et al.* (2007) 'Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes', *Science*, 317(5845), pp. 1753–1756. Available at: https://doi.org/10.1126/science.1142490.

Hrček, J., McLean, A.H.C. and Godfray, H.C.J. (2016) 'Symbionts modify interactions between insects and natural enemies in the field', *Journal of Animal Ecology*, 85(6), pp. 1605–1612. Available at: https://doi.org/10.1111/1365-2656.12586.

Huang, J. and Gogarten, J. (2007) 'Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids?', *Genome Biology*, 8(6), p. R99. Available at: https://doi.org/10.1186/gb-2007-8-6-r99.

Huang, Y. *et al.* (2021) 'Acesulfame aerobic biodegradation by enriched consortia and *Chelatococcus* spp.: Kinetics, transformation products, and genomic characterization', *Water Research*, 202, p. 117454. Available at: https://doi.org/10.1016/j.watres.2021.117454.

Hughes, G.L. *et al.* (2014) 'Native microbiome impedes vertical transmission of *Wolbachia* in *Anopheles* mosquitoes', *Proceedings of the National Academy of Sciences*, 111(34), pp. 12498–12503. Available at: https://doi.org/10.1073/pnas.1408888111.

Hunter, J.D. (2007) 'Matplotlib: A 2D Graphics Environment', *Computing in Science & Engineering*, 9(3), pp. 90–95. Available at: https://doi.org/10.1109/MCSE.2007.55.

Hurst, G.D.D. *et al.* (1994) 'The effect of infection with male-killing *Rickettsia* on the demography of female *Adalia bipunctata* L. (two spot ladybird)', *Heredity*, 73(3), pp. 309–316. Available at: https://doi.org/10.1038/hdy.1994.138.

Hurst, G.D.D. *et al.* (1999) 'Invasion of one insect species, *Adalia bipunctata*, by two different male-killing bacteria', *Insect Molecular Biology*, 8(1), pp. 133–139. Available at: https://doi.org/10.1046/j.1365-2583.1999.810133.x.

Hurst, G.D.D. (2017) 'Extended genomes: symbiosis and evolution.', *Interface focus*, 7(5), p. 20170001. Available at: https://doi.org/10.1098/rsfs.2017.0001.

Illumina (2011) 'CASAVA'.

Inkscape Project (2020) 'Inkscape'. Available at: https://inkscape.org.

Jaenike, J. (2007) 'Fighting back against male-killers', *Trends in Ecology & Evolution*, 22(4), pp. 167–169. Available at: https://doi.org/10.1016/j.tree.2007.01.008.

Jain, C. *et al.* (2018) 'High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries', *Nature Communications*, 9(1), p. 5114. Available at: https://doi.org/10.1038/s41467-018-07641-9.

Jeon, K.W. and Ahn, T.I. (1978) 'Temperature sensitivity: A cell character determined by obligate endosymbionts in amoebas', *Science*, 202(4368), pp. 635–637. Available at: https://doi.org/10.1126/science.202.4368.635.

Jeon, K.W. and Lorch, I.J. (1967) 'Unusual intra-cellular bacterial infection in large, free-living amoebae', *Experimental Cell Research*, 48(1), pp. 236–240. Available at: https://doi.org/10.1016/0014-4827(67)90313-8.

Jiang, N. *et al.* (2021) 'Rhamnose in plants - from biosynthesis to diverse functions', *Plant Science*, 302, p. 110687. Available at: https://doi.org/10.1016/j.plantsci.2020.110687.

Jiggins, F.M., Hurst, G.D.D. and Majerus, M.E.N. (2000) 'Sex-ratio-distorting *Wolbachia* causes sex-role reversal in its butterfly host', *Proceedings of the Royal Society B: Biological Sciences*, 267(1438), pp. 69–73. Available at: https://doi.org/10.1098/rspb.2000.0968.

Jofré, E., Lagares, A. and Mori, G. (2004) 'Disruption of dTDP-rhamnose biosynthesis modifies lipopolysaccharide core, exopolysaccharide production, and root colonization in *Azospirillum brasilense*', *FEMS Microbiology Letters*, 231(2), pp. 267–275. Available at: https://doi.org/10.1016/S0378-1097(04)00003-5.

Jones, P. *et al.* (2014) 'InterProScan 5: genome-scale protein function classification', *Bioinformatics*, 30(9), pp. 1236–1240. Available at: https://doi.org/10.1093/bioinformatics/btu031.

Jørgensen, C.B. (2001) 'August Krogh and Claude Bernard on basic principles in experimental physiology', *BioScience*, 51(1), p. 59. Available at: https://doi.org/10.1641/0006-3568(2001)051[0059:AKACBO]2.0.CO;2.

Joshi, N. and Fass, J. (2011) 'Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files'. Available at: https://github.com/najoshi/sickle.

Kaech, H. and Vorburger, C. (2021) 'Horizontal transmission of the heritable protective endosymbiont *Hamiltonella defensa* depends on titre and haplotype', *Frontiers in Microbiology*, 11, p. 628755. Available at: https://doi.org/10.3389/fmicb.2020.628755.

Kahane, S. *et al.* (1993) 'Description and partial characterization of a new *Chlamydia*-like microorganism', *FEMS Microbiology Letters*, 109(2–3), pp. 329–333. Available at: https://doi.org/10.1111/j.1574-6968.1993.tb06189.x.

Kalyaanamoorthy, S. *et al.* (2017) 'ModelFinder: Fast model selection for accurate phylogenetic estimates', *Nature Methods*, 14(6), pp. 587–589. Available at: https://doi.org/10.1038/nmeth.4285.

Kang, D.D. *et al.* (2019) 'MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies', *PeerJ*, 2019(7), p. e7359. Available at: https://doi.org/10.7717/peerj.7359.

Kang, Y.J. *et al.* (2014) 'Extensive diversity of Rickettsiales bacteria in two species of ticks from China and the evolution of the Rickettsiales', *BMC Evolutionary Biology*, 14(1), pp. 1–12. Available at: https://doi.org/10.1186/S12862-014-0167-2/FIGURES/4.

Kantor, R.S., Miller, S.E. and Nelson, K.L. (2019) 'The water microbiome through a pilot scale advanced treatment facility for direct potable reuse', *Frontiers in Microbiology*, 10. Available at: https://doi.org/10.3389/fmicb.2019.00993.

Karimi, E. *et al.* (2019) 'Genomic blueprints of sponge-prokaryote symbiosis are shared by low abundant and cultivatable Alphaproteobacteria', *Scientific Reports*, 9(1), p. 1999. Available at: https://doi.org/10.1038/s41598-019-38737-x.

Kau, A.L. *et al.* (2011) 'Human nutrition, the gut microbiome and the immune system', *Nature*, 474, pp. 327–336.

Kawafune, K. *et al.* (2015) 'Two different rickettsial bacteria invading *Volvox carteri*', *PLOS ONE*, 10(2), p. e0116192. Available at: https://doi.org/10.1371/JOURNAL.PONE.0116192.

Kikuchi, Y. and Fukatsu, T. (2005) '*Rickettsia* infection in natural leech populations', *Microbial Ecology*, 49(2), pp. 265–271. Available at: https://doi.org/10.1007/s00248-004-0140-5.

Kjeldsen, K.U. *et al.* (2010) 'Two types of endosymbiotic bacteria in the enigmatic marine worm *Xenoturbella bocki*', *Applied and Environmental Microbiology*, 76(8), pp. 2657–2662. Available at: https://doi.org/10.1128/AEM.01092-09.

Klimaszewski, J. *et al.* (2013) 'Molecular and microscopic analysis of the gut contents of abundant rove beetle species (Coleoptera, Staphylinidae) in the boreal balsam fir forest of Quebec, Canada', *ZooKeys*, 353, pp. 1–24. Available at: https://doi.org/10.3897/zookeys.353.5991.

Kliot, A. *et al.* (2014) 'Implication of the bacterial endosymbiont *Rickettsia* spp. in interactions of the whitefly *Bemisia tabaci* with tomato yellow leaf curl virus', *Journal of Virology*, 88(10), pp. 5652–5660. Available at: https://doi.org/10.1128/JVI.00071-14.

Kochert, G. and Olson, L.W. (1970) 'Endosymbiotic bacteria in *Volvox carteri*', *Transactions of the American Microscopical Society*, 89(4), p. 475. Available at: https://doi.org/10.2307/3224556.

Kolde, R. (2019) 'pheatmap: Pretty Heatmaps. R package'.

Kolmogorov, M. *et al.* (2019) 'Assembly of long, error-prone reads using repeat graphs', *Nature Biotechnology*, 37(5), pp. 540–546. Available at: https://doi.org/10.1038/s41587-019-0072-8.

Kondo, N. *et al.* (2002) 'Genome fragment of *Wolbachia* endosymbiont transferred to X chromosome of host insect', *Proceedings of the National Academy of Sciences*, 99(22), pp. 14280–14285. Available at: https://doi.org/10.1073/PNAS.222228199.

König, L. *et al.* (2017) 'Biphasic metabolism and host interaction of a Chlamydial symbiont', *mSystems*. Edited by A.D. Kent, 2(3). Available at: https://doi.org/10.1128/mSystems.00202-16.

Konstantinidis, K.T., Rosselló-Móra, R. and Amann, R. (2017) 'Uncultivated microbes in need of their own taxonomy', *The ISME Journal*, 11(11), pp. 2399–2406. Available at: https://doi.org/10.1038/ismej.2017.113.

Kontsedalov, S. *et al.* (2008) 'The presence of *Rickettsia* is associated with increased susceptibility of *Bemisia tabaci* (Homoptera: Aleyrodidae) to insecticides', *Pest Management Science*, 64(8), pp. 789–792. Available at: https://doi.org/10.1002/ps.1595.

Koren, S. *et al.* (2017) 'Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation', *Genome Research*, 27(5), pp. 722–736. Available at: https://doi.org/10.1101/GR.215087.116.

Köstlbacher, S. *et al.* (2021) 'Pangenomics reveals alternative environmental lifestyles among chlamydiae', *Nature Communications*, 12(1), p. 4021. Available at: https://doi.org/10.1038/s41467-021-24294-3.

Kotter, H. (2005) *Bionomie und Verbreitung der autochthonen Fiebermücke Anopheles plumbeus (Culicidae) und ihrer Vektorkompetenz für Plasmodium falciparum, Erreger der Malaria tropica*. Available at: https://doi.org/10.11588/heidok.00006104.

Krassowski, M., Arts, M., and CyrilLagger (2020) 'ComplexUpset'. Available at: https://doi.org/10.5281/zenodo.3700590.

Krebs, H.A. (1975) 'The August Krogh Principle: "For many problems there is an animal on which it can be most conveniently studied"', *The Journal of Experimental Zoology*, 194(1), pp. 221–226. Available at: https://doi.org/10.1002/jez.1401940115.

Kremer, N. *et al.* (2009) '*Wolbachia* interferes with ferritin expression and iron metabolism in insects', *PLoS Pathogens*, 5(10), p. e1000630. Available at: https://doi.org/10.1371/journal.ppat.1000630.

Küchler, S.M., Kehl, S. and Dettner, K. (2009) 'Characterization and localization of *Rickettsia* sp. in water beetles of genus *Deronectes* (Coleoptera: Dytiscidae)', *FEMS Microbiology Ecology*, 68(2), pp. 201–211. Available at: https://doi.org/10.1111/j.1574-6941.2009.00665.x.

Kumar, S. *et al.* (2013) 'Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots', *Frontiers in Genetics*, 4. Available at: https://doi.org/10.3389/fgene.2013.00237.

Kurtz, S. *et al.* (2004) 'Versatile and open software for comparing large genomes', *Genome Biology 2004 5:2*, 5(2), pp. 1–9. Available at: https://doi.org/10.1186/GB-2004-5-2-R12.

Lagkouvardos, I. *et al.* (2014) 'Integrating metagenomic and amplicon databases to resolve the phylogenetic and ecological diversity of the Chlamydiae', *The ISME Journal*, 8(1), pp. 115–125. Available at: https://doi.org/10.1038/ismej.2013.142.

Langmead, B. and Salzberg, S.L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nature Methods 2012 9:4*, 9(4), pp. 357–359. Available at: https://doi.org/10.1038/nmeth.1923.

Lanzoni, O. *et al.* (2019) 'Diversity and environmental distribution of the cosmopolitan endosymbiont "*Candidatus* Megaira"', *Scientific reports*, 9(1), p. 1179. Available at: https://doi.org/10.1038/s41598-018-37629-w.

Lartillot, N. *et al.* (2013) 'PhyloBayes MPI: Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment', *Systematic Biology*, 62(4), pp. 611–615. Available at: https://doi.org/10.1093/sysbio/syt022.

Lawson, E.T. *et al.* (2001) '*Rickettsia* associated with male-killing in a Buprestid Beetle', *Genetics*, 86, pp. 497–505.

Leclair, M. *et al.* (2017) 'Consequences of coinfection with protective symbionts on the host phenotype and symbiont titres in the pea aphid system', *Insect Science*, 24(5), pp. 798–808. Available at: https://doi.org/10.1111/1744-7917.12380.

Lefoulon, E. *et al.* (2016) 'Breakdown of coevolution between symbiotic bacteria *Wolbachia* and their filarial hosts', *PeerJ*, 4, p. e1840. Available at: https://doi.org/10.7717/peerj.1840.

Li, D. *et al.* (2015) 'MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph', *Bioinformatics*, 31(10), pp. 1674–1676. Available at: https://doi.org/10.1093/bioinformatics/btv033.

Li, H. *et al.* (2009) 'The sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), pp. 2078–2079. Available at: https://doi.org/10.1093/bioinformatics/btp352.

Li, H. (2018) 'Minimap2: pairwise alignment for nucleotide sequences', *Bioinformatics*. Edited by I. Birol, 34(18), pp. 3094–3100. Available at: https://doi.org/10.1093/bioinformatics/bty191.

Li, Y.H. *et al.* (2017) 'Plant-mediated horizontal transmission of *Rickettsia* endosymbiont between different whitefly species', *FEMS microbiology ecology*, 93(12), pp. 1–9. Available at: https://doi.org/10.1093/femsec/fix138.

Lo, N. (2003) 'Evidence for cocladogenesis between diverse Dictyopteran lineages and their intracellular endosymbionts', *Molecular Biology and Evolution*, 20(6), pp. 907–913. Available at: https://doi.org/10.1093/molbev/msg097.

Łukasik, P. *et al.* (2013) 'Protection against a fungal pathogen conferred by the aphid facultative endosymbionts *Rickettsia* and *Spiroplasma* is expressed in multiple host genotypes and species and is not influenced by co-infection with another symbiont', *Journal of Evolutionary Biology*, 26(12), pp. 2654–2661. Available at: https://doi.org/10.1111/jeb.12260.

Ma, Y., Pan, F. and McNeil, M. (2002) 'Formation of dTDP-rhamnose is essential for growth of *Mycobacteria*', *Journal of Bacteriology*, 184(12), pp. 3392–3395. Available at: https://doi.org/10.1128/JB.184.12.3392-3395.2002.

Manzano-Marín, A. *et al.* (2020) 'Serial horizontal transfer of vitamin-biosynthetic genes enables the establishment of new nutritional symbionts in aphids' di-symbiotic systems', *The ISME Journal*, 14(1), pp. 259–273. Available at: https://doi.org/10.1038/s41396-019-0533-6.

Martijn, J. *et al.* (2015) 'Single-cell genomics of a rare environmental alphaproteobacterium provides unique insights into Rickettsiaceae evolution', *The ISME Journal*, 9(11), pp. 2373–2385. Available at: https://doi.org/10.1038/ismej.2015.46.

Martin, M. (2011) 'Cutadapt removes adapter sequences from high-throughput sequencing reads', *EMBnet.journal*, 17(1), pp. 10–12. Available at: https://doi.org/10.14806/EJ.17.1.200.

Massey, J.H. and Newton, I.L.G. (2022) 'Diversity and function of arthropod endosymbiont toxins', *Trends in Microbiology*, 30(2), pp. 185–198. Available at: https://doi.org/10.1016/j.tim.2021.06.008.

Masui, S., Sasaki, T. and Ishikawa, H. (2000) 'Genes for the Type IV secretion system in an intracellular symbiont, *Wolbachia*, a causative agent of various sexual alterations in arthropods', *Journal of Bacteriology*, 182(22), pp. 6529–6531.

Matz, M.V. (2018) 'Fantastic beasts and how to sequence them: ecological genomics for obscure model organisms', *Trends in Genetics*, 34(2), pp. 121–132. Available at: https://doi.org/10.1016/j.tig.2017.11.002.

McDaniel, E.A. *et al.* (2021) 'Genome-resolved metagenomics of a photosynthetic bioreactor performing biological nutrient removal', *Microbiology Resource Announcements*, 10(18). Available at: https://doi.org/10.1128/MRA.00244-21.

McLean, A.H.C. *et al.* (2018) 'Consequences of symbiont co-infections for insect host phenotypes', *The Journal of Animal Ecology*, 87(2), pp. 478–488. Available at: https://doi.org/10.1111/1365-2656.12705.

Mediannikov, O. *et al.* (2012) 'New *Rickettsia* sp. in tsetse flies from Senegal', *Comparative Immunology, Microbiology and Infectious Diseases*, 35(2), pp. 145–150. Available at: https://doi.org/10.1016/j.cimid.2011.12.011.

Meloni, S. *et al.* (2003) 'The twin-arginine translocation (Tat) system is essential for Rhizobium–legume symbiosis', *Molecular Microbiology*, 48(5), pp. 1195–1207. Available at: https://doi.org/10.1046/j.1365-2958.2003.03510.x.

Minh, B.Q. *et al.* (2020) 'IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era', *Molecular Biology and Evolution*. Edited by E. Teeling, 37(5), pp. 1530–1534. Available at: https://doi.org/10.1093/molbev/msaa015.

Mod, H.K. *et al.* (2020) 'Scale dependence of ecological assembly rules: Insights from empirical datasets and joint species distribution modelling', *Journal of Ecology*, 108(5), pp. 1967–1977. Available at: https://doi.org/10.1111/1365-2745.13434.

Montenegro, H. *et al.* (2005) 'Male-killing *Spiroplasma* naturally infecting *Drosophila melanogaster*', *Insect Molecular Biology*, 14(3), pp. 281–287. Available at: https://doi.org/10.1111/j.1365-2583.2005.00558.x.

Moran, N.A. (2007) *Symbiosis as an adaptive process and source of phenotypic complexity*, *in the light of evolution: Volume I: Adaptation and complex design*. National Academies Press (US). Available at: https://www.ncbi.nlm.nih.gov/books/NBK254296/ (Accessed: 15 December 2022).

Moran, N.A., McCutcheon, J.P. and Nakabachi, A. (2008) 'Genomics and evolution of heritable bacterial symbionts', *Annual Review of Genetics*, 42(1), pp. 165–190. Available at: https://doi.org/10.1146/annurev.genet.41.110306.130119.

Moreira, L.A. *et al.* (2009) 'A *Wolbachia* symbiont in *Aedes aegypti* limits infection with dengue, chikungunya, and *Plasmodium*', *Cell*, 139(7), pp. 1268–1278. Available at: https://doi.org/10.1016/j.cell.2009.11.042.

Munson, M.A., Baumann, P. and Kinsey, M.G. (1991) '*Buchnera* gen. nov. and *Buchnera aphidicola* sp. nov., a Taxon Consisting of the Mycetocyte-Associated, Primary Endosymbionts of Aphids', *International Journal of Systematic Bacteriology*, 41(4), pp. 566–568. Available at: https://doi.org/10.1099/00207713-41-4-566.

Murray, G.G.R. *et al.* (2016) 'The phylogeny of *Rickettsia* using different evolutionary signatures: how tree-like is bacterial evolution?', *Systematic Biology*, 65(2), pp. 265–279. Available at: https://doi.org/10.1093/sysbio/syv084.

Nayfach, S. *et al.* (2021) 'A genomic catalog of Earth's microbiomes', *Nature Biotechnology*, 39(4), pp. 499–509. Available at: https://doi.org/10.1038/s41587-020-0718-6.

Nguyen, L.-T. *et al.* (2015) 'IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies', *Molecular Biology and Evolution*, 32(1), pp. 268–274. Available at: https://doi.org/10.1093/molbev/msu300.

Nicholson, W.L. and Paddock, C.D. (2017) *Rickettsial (Spotted & Typhus Fevers) & related infections, including Anaplasmosis & Ehrlichiosis*, *Centers for Disease Control and Prevention*. Available at: https://wwwnc.cdc.gov/travel/yellowbook/2018/infectious-diseases-related-to-travel/rickettsial-spotted-and-typhus-fevers-and-related-infections-including-anaplasmosis-and-ehrlichiosis.

Nozaki, H. *et al.* (1989) '*Pleodorina japonica* sp. nov. (*Volvocales, Chlorophyta*) with bacteria-like endosymbionts', *Phycologia*, 28(2), pp. 252–267. Available at: https://doi.org/10.2216/i0031-8884-28-2-252.1.

Nurk, S. *et al.* (2013) 'assembling genomes and mini-metagenomes from highly chimeric reads', *lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*, 7821 LNBI, pp. 158–170. Available at: https://doi.org/10.1007/978-3-642-37195-0_13.

Nylund, A. *et al.* (2018) 'Genotyping of '*Candidatus* Syngnamydia salmonis' (chlamydiales; Simkaniaceae) co-cultured in *Paramoeba perurans* (amoebozoa; Paramoebidae)', *Archives of Microbiology*, 200(6), pp. 859–867. Available at: https://doi.org/10.1007/s00203-018-1488-0.

Okonechnikov, K., Conesa, A. and García-Alcalde, F. (2016) 'Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data', *Bioinformatics*, 32(2), p. btv566. Available at: https://doi.org/10.1093/bioinformatics/btv566.

Oliver, K.M. *et al.* (2003) 'Facultative bacterial symbionts in aphids confer resistance to parasitic wasps', *Proceedings of the National Academy of Sciences*, 100(4), pp. 1803–1807. Available at: https://doi.org/10.1073/pnas.0335320100.

Oliver, K.M. *et al.* (2008) 'Population dynamics of defensive symbionts in aphids', *Proceedings of the Royal Society of Biology*, 275, pp. 293–299. Available at: https://doi.org/10.1098/rspb.2007.1192.

Oliver, K.M. *et al.* (2010) 'Facultative symbionts in aphids and the horizontal transfer of ecologically important traits', *Annual Review of Entomology*, 55(1), pp. 247–266. Available at: https://doi.org/10.1146/annurev-ento-112408-085305.

Onyeaka, H. *et al.* (2021) 'Minimizing carbon footprint via microalgae as a biological capture', *Carbon Capture Science & Technology*, 1, p. 100007. Available at: https://doi.org/10.1016/j.ccst.2021.100007.

Ortiz, M. *et al.* (2021) 'Multiple energy sources and metabolic strategies sustain microbial diversity in Antarctic desert soils', *Proceedings of the National Academy of Sciences*, 118(45). Available at: https://doi.org/10.1073/pnas.2025322118.

Pallen, M.J., Telatin, A. and Oren, A. (2021) 'The next million names for archaea and bacteria', *Trends in Microbiology*, 29(4), pp. 289–298. Available at: https://doi.org/10.1016/j.tim.2020.10.009.

Pannebakker, B.A. *et al.* (2007) 'Parasitic inhibition of cell death facilitates symbiosis', *Proceedings of the National Academy of Sciences*, 104(1), pp. 213–215. Available at: https://doi.org/10.1073/pnas.0607845104.

Parks, D.H. *et al.* (2015) 'CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes', *Genome Research*, 25(7), pp. 1043–1055. Available at: https://doi.org/10.1101/gr.186072.114.

Pasqualetti, C. *et al.* (2020) 'The obligate symbiont "*Candidatus* Megaira polyxenophila" has variable effects on the growth of different host species', *Frontiers in Microbiology*, 11, p. 1425. Available at: https://doi.org/10.3389/fmicb.2020.01425.

Pedregosa, F. *et al.* (2011) 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research*, 12(85), pp. 2825–2830.

Penard, E. (1902) *Faune rhizopodique du bassin du Léman*. Genève, H. Kündig. Available at: https://archive.org/details/faunerhizopodiqu00pena/page/138/mode/2up?q=Pelomyxa (Accessed: 12 May 2020).

Pereira, T.N. *et al.* (2018) '*Wolbachia* significantly impacts the vector competence of Aedes aegypti for Mayaro virus', *Scientific Reports*, 8(1), pp. 1–9. Available at: https://doi.org/10.1038/s41598-018-25236-8.

Perlman, S.J., Hunter, M.S. and Zchori-Fein, E. (2006) 'The emerging diversity of *Rickettsia*', *Proceedings of the Royal Society B: Biological Sciences*, 273(1598), pp. 2097–2106. Available at: https://doi.org/10.1098/rspb.2006.3541.

Perotti, M.A. *et al.* (2006) '*Rickettsia* as obligate and mycetomic bacteria.', *The FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 20(13), pp. 2372–2374. Available at: https://doi.org/10.1096/fj.06-5870fje.

Pilgrim, J. *et al.* (2017) 'Torix group *Rickettsia* are widespread in Culicoides biting midges (Diptera: Ceratopogonidae), reach high frequency and carry unique genomic features', *Environmental Microbiology*, 19(10), pp. 4238–4255. Available at: https://doi.org/10.1111/1462-2920.13887.

Pilgrim, J. *et al.* (2021) 'Torix *Rickettsia* are widespread in arthropods and reflect a neglected symbiosis', *GigaScience*, 10(3), pp. 1–19. Available at: https://doi.org/10.1093/gigascience/giab021.

Pillonel, T., Bertelli, C. and Greub, G. (2018) 'Environmental metagenomic assemblies reveal seven new highly divergent Chlamydial lineages and hallmarks of a conserved intracellular lifestyle', *Frontiers in Microbiology*, 9. Available at: https://doi.org/10.3389/fmicb.2018.00079.

Poltronieri, P. and Čerekovic, N. (2018) 'Roles of nicotinamide adenine dinucleotide (NAD+) in biological systems', *Challenges*, 9(1), p. 3. Available at: https://doi.org/10.3390/challe9010003.

Pritchard, L. *et al.* (2016) 'Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens', *Analytical Methods*, 8(1), pp. 12–24. Available at: https://doi.org/10.1039/C5AY02550H.

Prowazek, S. von S. (1912) 'Chlamydzoan', in *Handbuch der pathogenen Protozoen Vol. 1*. Leipzig: J. A. Barth, pp. 119–121. Available at: https://doi.org/10.5962/bhl.title.10198.

Pukall, R., TschÃ¤pe, H. and Smalla, K. (1996) 'Monitoring the spread of broad host and narrow host range plasmids in soil microcosms', *FEMS Microbiology Ecology*, 20(1), pp. 53–66. Available at: https://doi.org/10.1111/j.1574-6941.1996.tb00304.x.

QGIS.org (2020) 'QGIS 3.16'. Hannover: QGIS Association. Available at: http://www.qgis.org/.

R Core Team (2020) 'R: A Language and Environment for Statistical Computing'. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.r-project.org/.

Reed, J.W. and Walker, G.C. (1991) 'The exoD gene of *Rhizobium meliloti* encodes a novel function needed for alfalfa nodule invasion', *Journal of Bacteriology*, 173(2), pp. 664–677. Available at: https://doi.org/10.1128/jb.173.2.664-677.1991.

Reynolds, L.A. *et al.* (2019) 'Suppression of *Wolbachia*-mediated male-killing in the butterfly *Hypolimnas bolina* involves a single genomic region', *PeerJ*, 7, p. e7677. Available at: https://doi.org/10.7717/peerj.7677.

Rice, D.W., Sheehan, K.B. and Newton, I.L.G. (2017) 'Large-scale identification of *Wolbachia pipientis* effectors', *Genome Biology and Evolution*, 9(7), pp. 1925–1937. Available at: https://doi.org/10.1093/gbe/evx139.

Rodriguez-R, L.M. *et al.* (2020) 'Iterative subtractive binning of freshwater chronoseries metagenomes identifies over 400 novel species and their ecologic preferences', *Environmental Microbiology*, 22(8), pp. 3394–3412. Available at: https://doi.org/10.1111/1462-2920.15112.

Rodriguez-R, L.M. *et al.* (2021) 'Reply to: "Re-evaluating the evidence for a universal genetic boundary among microbial species"', *Nature Communications*, 12(1), p. 4060. Available at: https://doi.org/10.1038/s41467-021-24129-1.

Rodriguez-R, L.M. and Konstantinidis, K.T. (2016) 'The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes'. Available at: https://doi.org/10.7287/PEERJ.PREPRINTS.1900V1.

Rossum, G.V. and Drake, F.L. (2009) *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Sabaneyeva, E. *et al.* (2018) 'Host and symbiont intraspecific variability : The case of *Paramecium calkinsi* and "*Candidatus* Trichorickettsia mobilis"', *European Journal of Protistology*, 62, pp. 79–94. Available at: https://doi.org/10.1016/j.ejop.2017.12.002.

Sabuda, M.C. *et al.* (2021) 'Biogeochemical gradients in a serpentinization-influenced aquifer: Implications for gas exchange between the subsurface and atmosphere', *Journal of Geophysical Research: Biogeosciences*, 126(8), p. e2020JG006209. Available at: https://doi.org/10.1029/2020JG006209.

Sahlin, K. *et al.* (2014) 'BESST - Efficient scaffolding of large fragmented assemblies', *BMC Bioinformatics*, 15(1), p. 281. Available at: https://doi.org/10.1186/1471-2105-15-281.

Sakurai, M. *et al.* (2005) '*Rickettsia* symbiont in the pea aphid *Acyrthosiphon pisum*: novel cellular tropism, effect on host fitness, and interaction with the essential symbiont

Buchnera', *Environmental Microbiology*, 71(7), pp. 4069–4075. Available at: https://doi.org/10.1128/aem.71.7.4069-4075.2005.

Salem, H. *et al.* (2020) 'Symbiont digestive range reflects host plant breadth in herbivorous beetles', *Current Biology*, 30(15), pp. 2875-2886.e4. Available at: https://doi.org/10.1016/j.cub.2020.05.043.

Sangwan, N., Xia, F. and Gilbert, J.A. (2016) 'Recovering complete and draft population genomes from metagenome datasets', *Microbiome*, 4(1), p. 8. Available at: https://doi.org/10.1186/s40168-016-0154-5.

Sapp, J. (2002) 'Paul Buchner (1886–1978) and hereditary symbiosis in insects', *International Microbiology*, 5(3), pp. 145–150. Available at: https://doi.org/10.1007/s10123-002-0079-7.

Schaffner, F. *et al.* (2012) '*Anopheles plumbeus (*Diptera: Culicidae) in Europe: A mere nuisance mosquito or potential malaria vector?', *Malaria Journal*, 11, p. 393. Available at: https://doi.org/10.1186/1475-2875-11-393.

Schardl, C.L., Leuchtmann, A. and Spiering, M.J. (2004) 'Symbioses of grasses with seedborne fungal endophytes', *Annual Review of Plant Biology*, 55(1), pp. 315–340. Available at: https://doi.org/10.1146/annurev.arplant.55.031903.141735.

Schneider, D. *et al.* (2020) 'Metagenomes of wastewater at different treatment stages in central Germany', *Microbiology Resource Announcements*, 9(15). Available at: https://doi.org/10.1128/MRA.00201-20.

Scholz, M. *et al.* (2020) 'Large scale genome reconstructions illuminate *Wolbachia* evolution', *Nature Communications*, 11, p. 5235. Available at: https://doi.org/10.1038/s41467-020-19016-0.

Schrallhammer, M. (2010) 'The killer trait of *Paramecium* and its causative agents'.

Schrallhammer, M. *et al.* (2013) '"*Candidatus* Megaira polyxenophila" gen. nov., sp. nov.: Considerations on evolutionary history, host range and shift of early divergent *Rickettsiae*', *PLoS ONE*. Edited by S.A. Ralph, 8(8), p. e72581. Available at: https://doi.org/10.1371/journal.pone.0072581.

Schrallhammer, M., Castelli, M. and Petroni, G. (2018) 'Phylogenetic relationships among endosymbiotic R-body producer: Bacteria providing their host the killer trait', *Systematic and Applied Microbiology*, 41(3), pp. 213–220. Available at: https://doi.org/10.1016/j.syapm.2018.01.005.

von der Schulenburg, J.H.G. *et al.* (2001) 'Incidence of male-killing *Rickettsia* spp. (*α-Proteobacteria*) in the ten-spot ladybird beetle *Adalia decempunctata L.* (Coleoptera: Coccinellidae)', *Applied and Environmental Microbiology*, 67(1), pp. 270–277. Available at: https://doi.org/10.1128/AEM.67.1.270–277.2001.

Schulz, F. *et al.* (2016) 'A Rickettsiales symbiont of amoebae with ancient features', *Environmental Microbiology*, 18(8), pp. 2326–2342. Available at: https://doi.org/10.1111/1462-2920.12881.

Seabold, S. and Perktold, J. (2010) 'Statsmodels: Econometric and Statistical Modeling with Python', *PROC. OF THE 9th PYTHON IN SCIENCE CONF*. Available at: http://statsmodels.sourceforge.net/ (Accessed: 11 February 2021).

Seemann, T. (2014) 'Prokka: rapid prokaryotic genome annotation', *Bioinformatics*, 30(14), pp. 2068–2069. Available at: https://doi.org/10.1093/bioinformatics/btu153.

Shanmugabalaji, V. *et al.* (2022) 'Plastoglobules: A hub of lipid metabolism in the chloroplast', in *Advances in Botanical Research*. Elsevier, pp. 91–119. Available at: https://doi.org/10.1016/bs.abr.2021.09.002.

Shen, W. *et al.* (2016) 'SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation', *PLOS ONE*, 11(10), p. e0163962. Available at: https://doi.org/10.1371/JOURNAL.PONE.0163962.

Siozios, S. *et al.* (2013) 'The diversity and evolution of *Wolbachia* ankyrin repeat domain genes', *PLoS ONE*, 8(2), p. e55390. Available at: https://doi.org/10.1371/journal.pone.0055390.

Siozios, S. *et al.* (2020) 'DNA barcoding reveals incorrect labelling of insects sold as food in the UK', *PeerJ*, 8, p. e8496. Available at: https://doi.org/10.7717/peerj.8496.

Siozios, S. (2022) 'SioStef/panplots'. Available at: https://doi.org/10.5281/zenodo.6408803.

Smee, M.R., Raines, S.A. and Ferrari, J. (2021) 'Genetic identity and genotype × genotype interactions between symbionts outweigh species level effects in an insect microbiome', *The ISME Journal*, 15(9), pp. 2537–2546. Available at: https://doi.org/10.1038/s41396-021-00943-9.

Sonneborn, T.M. (1943) 'Gene and cytoplasm. I. the determination and inheritance of the killer character in variety 4 of *Paramecium aurelia*', in *Proceedings of the National Academy of Sciences*, pp. 329–338. Available at: https://doi.org/10.1073/pnas.29.11.329.

Stouthamer, C.M., Kelly, S. and Hunter, M.S. (2018) 'Enrichment of low-density symbiont DNA from minute insects', *Journal of Microbiological Methods*, 151, pp. 16–19. Available at: https://doi.org/10.1016/j.mimet.2018.05.013.

Stouthamer, R., Breeuwer, J.A.J. and Hurst, G.D.D. (1999) '*Wolbachia* p*ipientis* : Microbial manipulator of arthropod reproduction', *Annual Review of Microbiology*, 53(1), pp. 71–102. Available at: https://doi.org/10.1146/annurev.micro.53.1.71.

Sumida, Y. *et al.* (2017) '*Wolbachia* induces costs to life-history and reproductive traits in the moth, *Ephestia kuehniella*', *Journal of Stored Products Research*, 71, pp. 93–98. Available at: https://doi.org/10.1016/j.jspr.2017.02.003.

Tamura, A. *et al.* (1995) 'Classification of *Rickettsia tsutsugamushi* in a new genus, *Orientia* gen. nov., as *Orientia tsutsugamushi* comb. nov', *International journal of systematic bacteriology*, 45(3), pp. 589–591. Available at: https://doi.org/10.1099/00207713-45-3-589.

Tandon, P., Jin, Q. and Huang, L. (2017) 'A promising approach to enhance microalgae productivity by exogenous supply of vitamins', *Microbial Cell Factories*, 16(1), p. 219. Available at: https://doi.org/10.1186/s12934-017-0834-2.

Tatusova, T. *et al.* (2016) 'NCBI prokaryotic genome annotation pipeline', *Nucleic Acids Research*, 44(14), pp. 6614–6624. Available at: https://doi.org/10.1093/nar/gkw569.

Thongprem, P. *et al.* (2020) 'Incidence and Diversity of Torix *Rickettsia*–Odonata Symbioses', *Microbial Ecology* [Preprint]. Available at: https://doi.org/10.1007/s00248-020-01568-9.

Thongprem, P. *et al.* (2020) 'Transmission, tropism, and biological impacts of torix *Rickettsia* in the common bed bug *Cimex lectularius* (Hemiptera: Cimicidae)', *Frontiers in Microbiology*, 11. Available at: https://doi.org/10.3389/fmicb.2020.608763.

Todorova, A.K. *et al.* (1995) 'Nostocyclamide: A new macrocyclic, thiazole-containing allelochemical from *Nostoc* sp. 31 (*Cyanobacteria*)', *The Journal of Organic Chemistry*, 60(24), pp. 7891–7895. Available at: https://doi.org/10.1021/jo00129a032.

Tran, P.Q. *et al.* (2021) 'Depth-discrete metagenomics reveals the roles of microbes in biogeochemical cycling in the tropical freshwater Lake Tanganyika', *The ISME Journal*, 15(7), pp. 1971–1986. Available at: https://doi.org/10.1038/s41396-021-00898-x.

Tseng, T.-T., Tyler, B.M. and Setubal, J.C. (2009) 'Protein secretion systems in bacterial-host associations, and their description in the Gene Ontology', *BMC Microbiology*, 9(1), p. S2. Available at: https://doi.org/10.1186/1471-2180-9-S1-S2.

Tsuchida, T. *et al.* (2002) 'Diversity and geographic distribution of secondary endosymbiotic bacteria in natural populations of the pea aphid, *Acyrthosiphon pisum'*, *Molecular Ecology*, 11(10), pp. 2123–2135. Available at: https://doi.org/10.1046/j.1365-294X.2002.01606.x.

Tully, B.J. *et al.* (2018) 'A dynamic microbial community with high functional redundancy inhabits the cold, oxic subseafloor aquifer', *The ISME Journal*, 12(1), pp. 1–16. Available at: https://doi.org/10.1038/ismej.2017.187.

Turnbaugh, P.J. *et al.* (2009) 'A core gut microbiome in obese and lean twins', *Nature*, 457(7228), pp. 480–484.

Tvedte, E.S. *et al.* (2019) 'Genome of the parasitoid wasp *Diachasma alloeum*, an emerging model for ecological speciation and transitions to asexual reproduction', *Genome Biology and Evolution*. Edited by J. Gonzalez, 11(10), pp. 2767–2773. Available at: https://doi.org/10.1093/gbe/evz205.

Van Ham, R.C.H.J. *et al.* (2003) 'Reductive genome evolution in *Buchnera aphidicola*', *Proceedings of the National Academy of Sciences of the United States of America*, 100(2), pp. 581–586. Available at: https://doi.org/10.1073/PNAS.0235981100.

Vannini, C. *et al.* (2013) 'A new obligate bacterial symbiont colonizing the ciliate *Euplotes* in brackish and freshwater : " *Candidatus* Protistobacter heckmanni "', 70, pp. 233–243. Available at: https://doi.org/10.3354/ame01657.

Vannini, C. *et al.* (2014) 'Flagellar movement in two bacteria of the family *Rickettsiaceae*: a re-evaluation of motility in an evolutionary perspective', *PLoS ONE*. Edited by C.A. Brissette, 9(2), p. e87718. Available at: https://doi.org/10.1371/journal.pone.0087718.

Vega, I.A., Arribére, M.A. and Castro-vazquez, A. (2012) 'Apple snails and their endosymbionts bioconcentrate heavy metals and uranium from contaminated drinking water', *Environmental Science and Pollution Research*, 19, pp. 3307–3316. Available at: https://doi.org/10.1007/s11356-012-0848-6.

Větrovský, T. *et al.* (2019) 'A meta-analysis of global fungal distribution reveals climate-driven patterns', *Nature Communications*, 10(1), p. 5142. Available at: https://doi.org/10.1038/s41467-019-13164-8.

Vosloo, S. *et al.* (2021) 'Evaluating *de Novo* assembly and binning strategies for time series drinking water metagenomes', *Microbiology Spectrum*, 9(3). Available at: https://doi.org/10.1128/Spectrum.01434-21.

Walker, B.J. *et al.* (2014) 'Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement', *PLoS ONE*. Edited by J. Wang, 9(11), p. e112963. Available at: https://doi.org/10.1371/journal.pone.0112963.

Walker, T. *et al.* (2011) 'The wMel *Wolbachia* strain blocks dengue and invades caged *Aedes aegypti* populations', *Nature*, 476(7361), pp. 450–453. Available at: https://doi.org/10.1038/nature10355.

Walker, T. *et al.* (2021) 'Stable high-density and maternally inherited *Wolbachia* infections in *Anopheles moucheti* and *Anopheles demeilloni* mosquitoes', *Current Biology*, 31(11), pp. 2310-2320.e5. Available at: https://doi.org/10.1016/J.CUB.2021.03.056.

Wang, H.-L. *et al.* (2020) 'A newly recorded *Rickettsia* of the Torix group is a recent intruder and an endosymbiont in the whitefly *Bemisia tabaci*', *Environmental Microbiology*, 22(4), pp. 1207–1221. Available at: https://doi.org/10.1111/1462-2920.14927.

Waskom, M. and Seaborn development team (2020) 'mwaskom/seaborn'. Zenodo. Available at: https://doi.org/10.5281/zenodo.592845.

Watanabe, K. *et al.* (2016) 'Ciliate *Paramecium* is a natural reservoir of *Legionella pneumophila*', *Nature Publishing Group*, pp. 1–15. Available at: https://doi.org/10.1038/srep24322.

Watts, T. *et al.* (2009) 'Variable incidence of *Spiroplasma* infections in natural populations of *Drosophila* species', *PLoS ONE*, 4(5), p. e5703. Available at: https://doi.org/10.1371/journal.pone.0005703.

Wei, G. *et al.* (2019) 'Terpene biosynthesis in red algae Is catalyzed by microbial type but not typical plant terpene synthases', *Plant Physiology*, 179(2), pp. 382–390. Available at: https://doi.org/10.1104/pp.18.01413.

Weinert, L.A. *et al.* (2009) 'Evolution and diversity of *Rickettsia* bacteria', *BMC Biology*, 7(1), p. 6. Available at: https://doi.org/10.1186/1741-7007-7-6.

Weinert, L.A. *et al.* (2015) 'The incidence of bacterial endosymbionts in terrestrial arthropods', *Proceedings of the Royal Society B: Biological Sciences*, 282(1807), p. 20150249. Available at: https://doi.org/10.1098/rspb.2015.0249.

Weldon, S.R., Russell, J.A. and Oliver, K.M. (2020) 'More Is not always better: coinfections with defensive symbionts generate highly variable outcomes', *Applied and Environmental Microbiology*, 86(5), pp. e02537-19. Available at: https://doi.org/10.1128/AEM.02537-19.

Wenski, S.L., Thiengmag, S. and Helfrich, E.J.N. (2022) 'Complex peptide natural products: Biosynthetic principles, challenges and opportunities for pathway engineering', *Synthetic and Systems Biotechnology*, 7(1), pp. 631–647. Available at: https://doi.org/10.1016/j.synbio.2022.01.007.

Werner, D. *et al.* (2014) 'The citizen science project "Mückenatlas" supports mosquito (Diptera, Culicidae) monitoring in Germany.', *Proceedings of the 8th International Conference on Urban Pests, 20-23 July 2014, Zurich, Switzerland*, pp. 119–124.

Werren, J.H. *et al.* (1994) 'Rickettsial relative associated with male killing in the ladybird beetle (*Adalia bipunctata*)', *Journal of Bacteriology*, 176(2), pp. 388–394. Available at: https://doi.org/10.1128/jb.176.2.388-394.1994.

Wick, R.R. and Holt, K.E. (2022) 'Polypolish: Short-read polishing of long-read bacterial genome assemblies', *PLoS computational biology*, 18(1), p. e1009802. Available at: https://doi.org/10.1371/journal.pcbi.1009802.

Wickham, H. (2016) 'ggplot2: Elegant graphics for data analysis'. Springer-Verlag New York. Available at: https://ggplot2.tidyverse.org (Accessed: 19 July 2021).

Wood, D.E., Lu, J. and Langmead, B. (2019) 'Improved metagenomic analysis with Kraken 2', *Genome Biology*, 20(1), p. 257. Available at: https://doi.org/10.1186/s13059-019-1891-0.

Wu, M. *et al.* (2004) 'Phylogenomics of the reproductive parasite *Wolbachia pipientis* wMel: A streamlined genome overrun by mobile genetic elements', *PLoS Biology*, 2(3), p. e69. Available at: https://doi.org/10.1371/journal.pbio.0020069.

Wu, T. *et al.* (2022) 'Local adaptation to hosts and parasitoids shape *Hamiltonella defensa* genotypes across aphid species', *Proceedings of the Royal Society B: Biological Sciences*, 289(1985), p. 20221269. Available at: https://doi.org/10.1098/rspb.2022.1269.

Xie, J. *et al.* (2011) 'Effect of the *Drosophila* endosymbiont *Spiroplasma* on parasitoid wasp development and on the reproductive fitness of wasp-attacked fly survivors.', *Evolutionary ecology*, 53(5), pp. 1065–1079. Available at: https://doi.org/10.1007/s10682-010-9453-7.

Xie, J. *et al.* (2014) 'Male killing *Spiroplasma* protects *Drosophila melanogaster* against two parasitoid wasps', *Heredity*, 112(4), pp. 399–408. Available at: https://doi.org/10.1038/hdy.2013.118.

Xie, J., Vilchez, I. and Mateos, M. (2010) '*Spiroplasma* bacteria enhance survival of *Drosophila hydei* attacked by the parasitic wasp *Leptopilina heterotoma*', *PLoS ONE*, 5(8), p. e12149. Available at: https://doi.org/10.1371/journal.pone.0012149.

Yan, P. *et al.* (2019) 'Microbial diversity in the tick *Argas japonicus* (Acari: Argasidae) with a focus on *Rickettsia* pathogens', *Medical and Veterinary Entomology*, 33(3), pp. 327–335. Available at: https://doi.org/10.1111/mve.12373.

Yancey, C.E. *et al.* (2022) 'Metagenomic and metatranscriptomic insights into population diversity of *Microcystis* blooms: Spatial and temporal dynamics of *mcy* genotypes, including a partial operon that can be abundant and expressed', *Applied and Environmental Microbiology*, 88(9). Available at: https://doi.org/10.1128/aem.02464-21.

Ye, N. *et al.* (2015) '*Saccharina* genomes provide novel insight into kelp biology', *Nature Communications*, 6, p. 6986. Available at: https://doi.org/10.1038/ncomms7986.

Yong, E. (2016) *I contain multitudes: The microbes within us and a grander view of life*. Illustrated edition. New York, NY: Ecco Press.

Yuan, C. *et al.* (2015) 'Reconstructing 16S rRNA genes in metagenomic data', *Bioinformatics*, 31(12), pp. i35–i43. Available at: https://doi.org/10.1093/bioinformatics/btv231.

Yurchenko, T. *et al.* (2018) 'A gene transfer event suggests a long-term partnership between eustigmatophyte algae and a novel lineage of endosymbiotic bacteria', *The ISME Journal*, 12(9), pp. 2163–2175. Available at: https://doi.org/10.1038/s41396-018-0177-y.

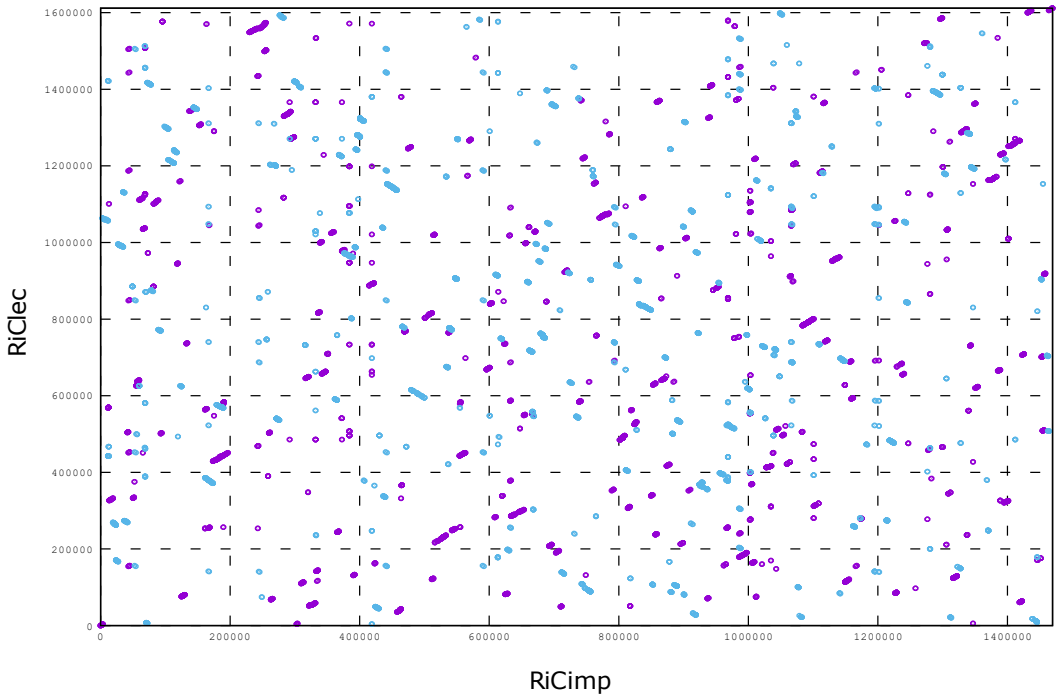
Zchori-Fein, E., Borad, C. and Harari, A.R. (2006) 'Oogenesis in the date stone beetle, *Coccotrypes dactyliperda*, depends on symbiotic bacteria', *Physiological Entomology*, 31(2), pp. 164–169. Available at: https://doi.org/10.1111/j.1365-3032.2006.00504.x.

Zélé, F. *et al.* (2012) 'Infection with *Wolbachia* protects mosquitoes against *Plasmodium*-induced mortality in a natural system', *Journal of Evolutionary Biology*, 25(7), pp. 1243–1252. Available at: https://doi.org/10.1111/j.1420-9101.2012.02519.x.

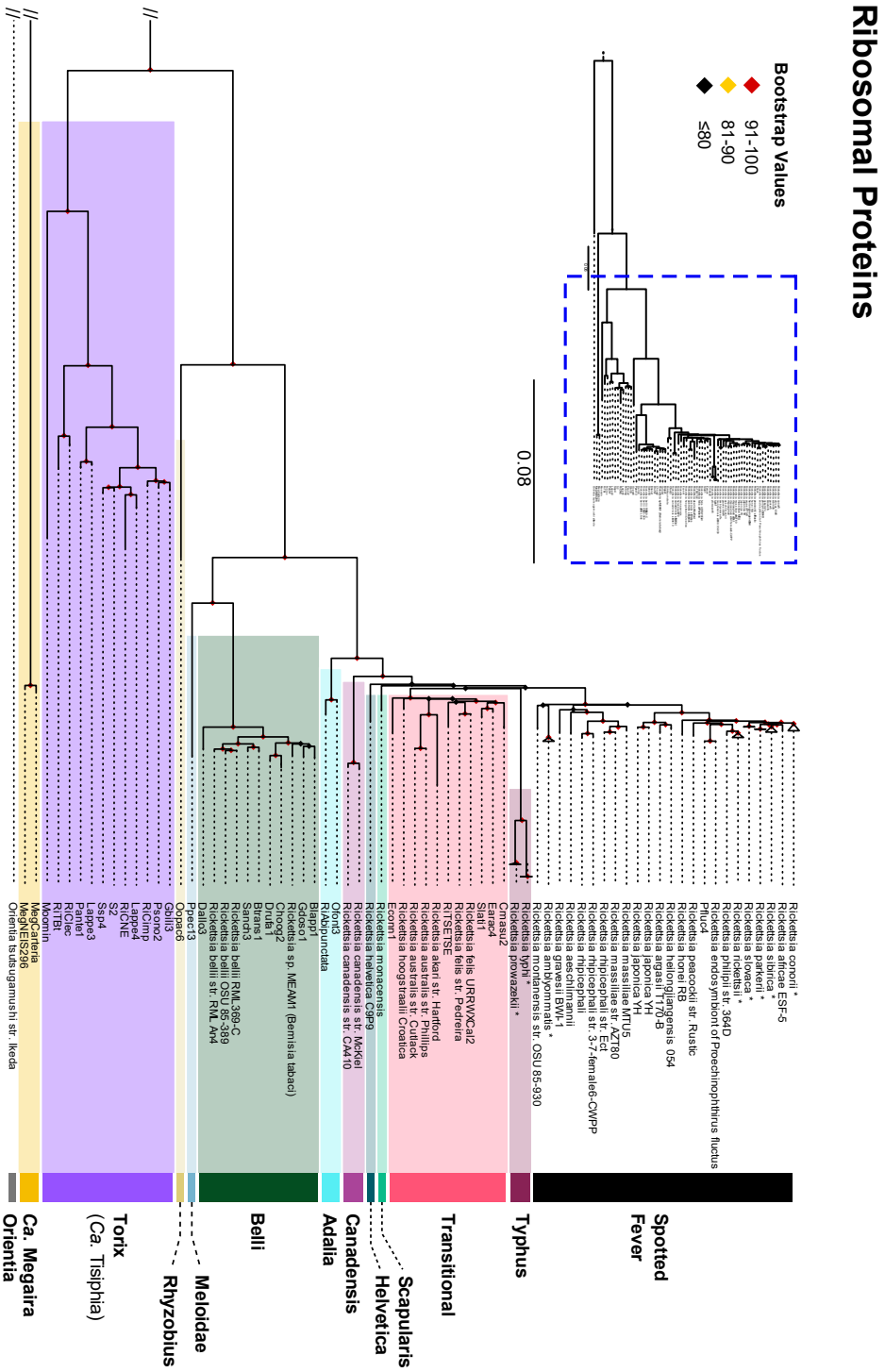
Zélé, F. *et al.* (2020) 'Endosymbiont diversity in natural populations of *Tetranychus* mites is rapidly lost under laboratory conditions', *Heredity*, 124(4), pp. 603–617. Available at: https://doi.org/10.1038/s41437-020-0297-9.

Appendices

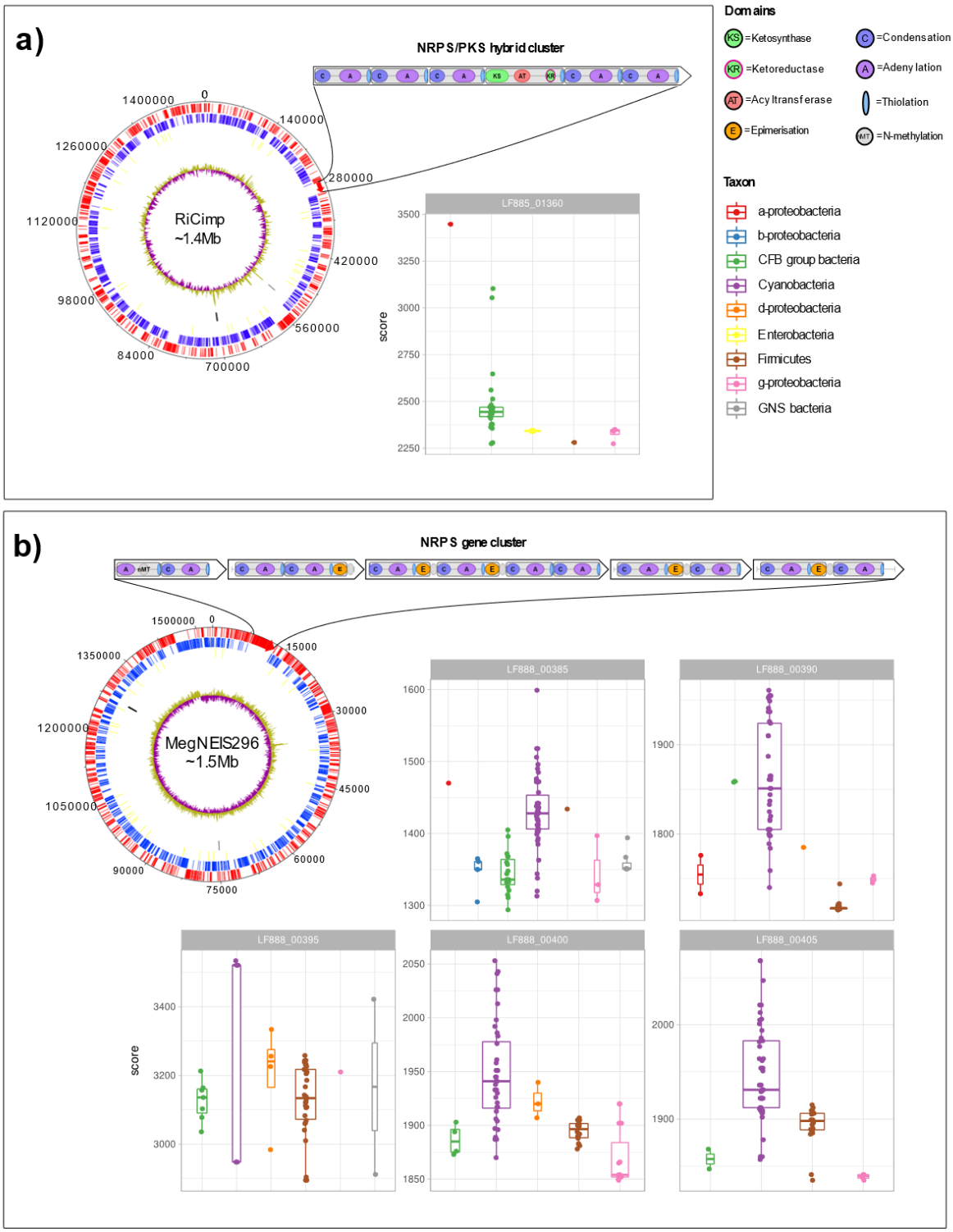
Appendix A. Additional Figures



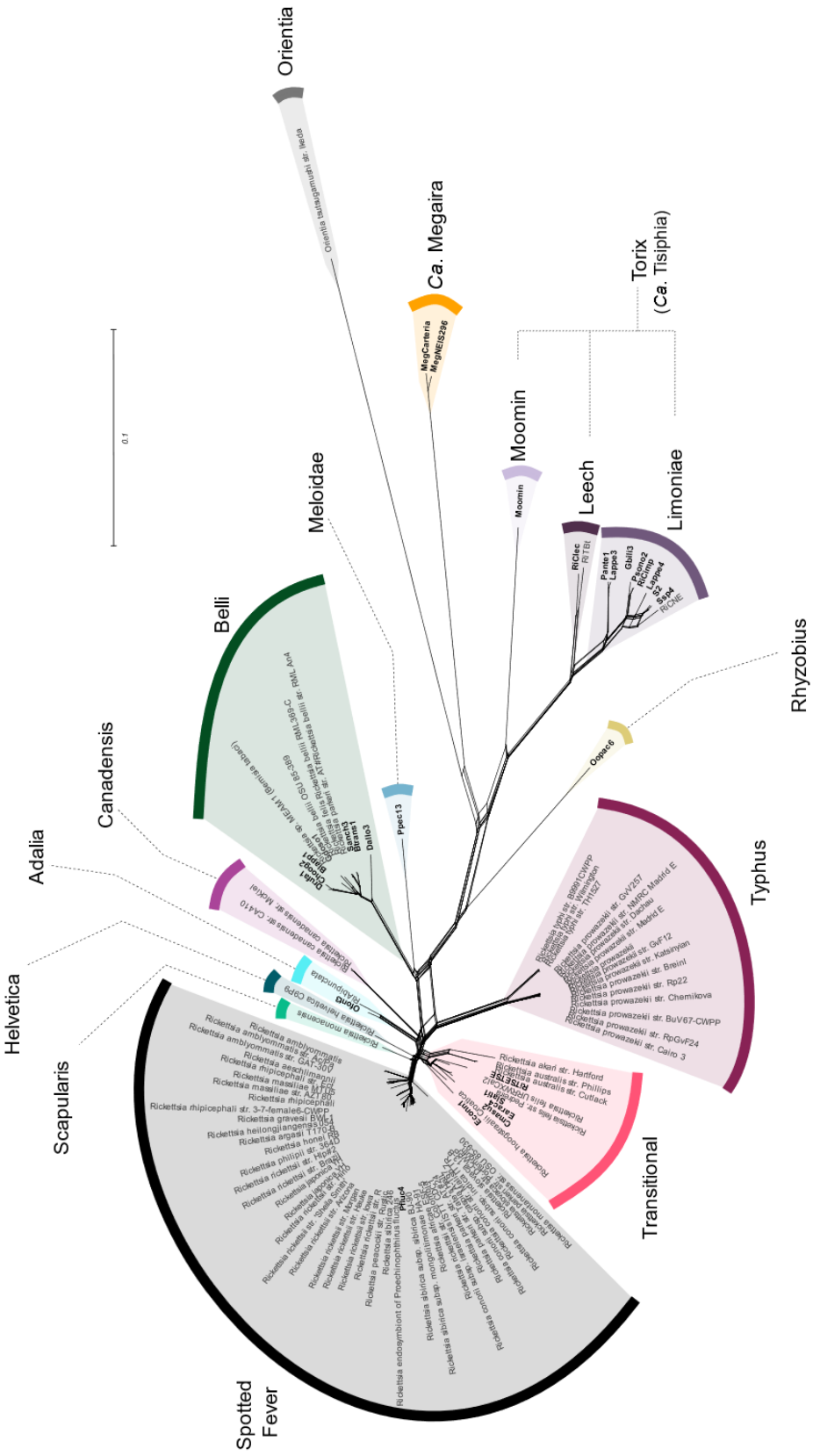
Appendix figure A.1. Whole genome alignment between the complete Torix limoniae (RiCimp) and Torix leech (RiClec) genomes reveals lack of synteny. Magenta represents forward matches and blue reverse matches.



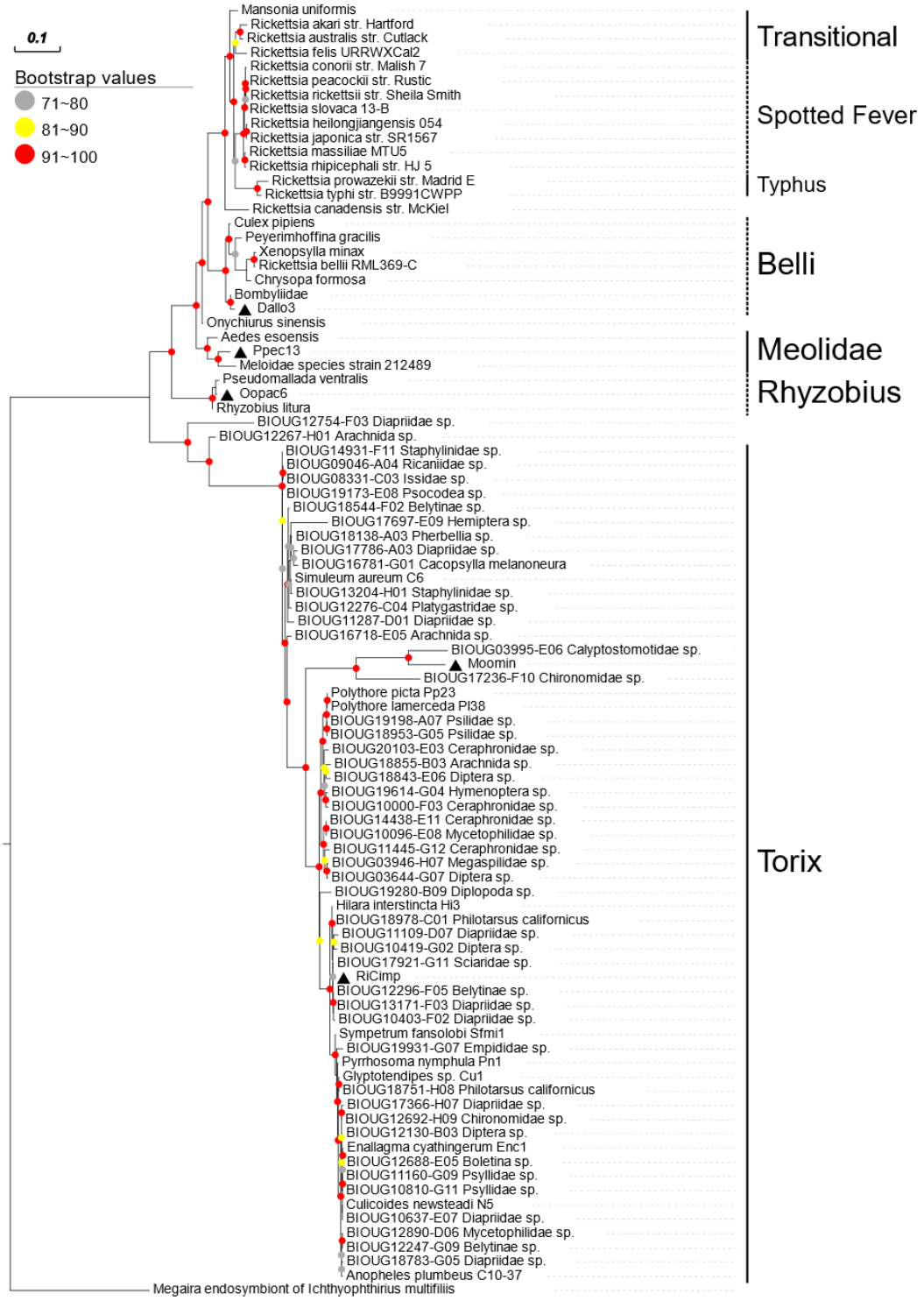
Appendix figure A.2. Maximum likelihood (ML) phylogeny of *Rickettsia* and '*Ca*. Megaira' constructed from 43 ribosomal protein gene clusters extracted from the pangenome. New genomes are written in bold and bootstrap values based on 1000 replicates are indicated with coloured diamonds. Asterisks pangenome. New genomes are written in bold and bootstrap values based on 1000 replicates are indicated with coloured diamonds. Asterisks indicate collapsed monophyletic branches and "//" represent breaks in the branch. New complete genomes are: RiCimp, RiClec and MegNEIS296. See Appendix B.1 for genome metadata.



Appendix figure A.3. The circular chromosomes of A) a Torix group *Rickettsia* (RiCimp) and B) a '*Ca*. Megaira' sp. (MegNEIS296). From outside to in, the circles represent: forward CDSs (Red), Reverse CDSs (blue), tRNAs (yellow) rRNAs (black), and GC content (green and magenta). Highlighted are the predicted domains that form non-ribosomal peptide synthase (NRPS) and hybrid non-ribosomal peptide synthase/polyketide synthase (NRPS/PKS) gene modules. Modules define individual amino acids in the synthesised peptide and show the catalytic domains within modules. Beneath the predicted domains are diagrams showing the similarity of modules to their closest relatives in other species determined by blastp. See appendix B/2 for raw data.

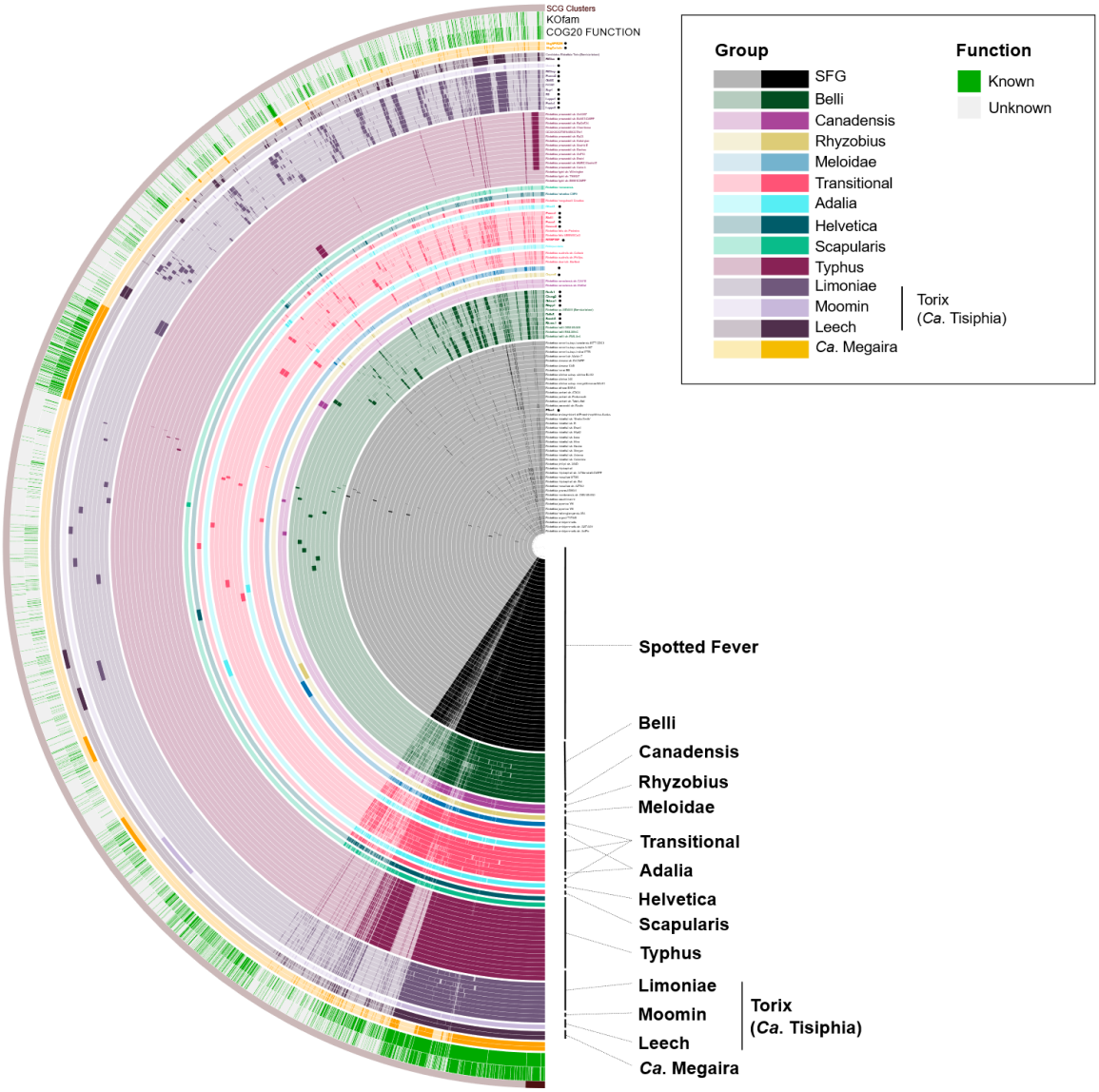


Appendix figure A.4. Nearest Neighbour Network, displaying the distances between the 74 core gene sets across all 104 Rickettsia, Ca. Megaira genomes, and the outgroup Orientia. New genomes are indicated with bold text. See Appendix B.1 for metadata.

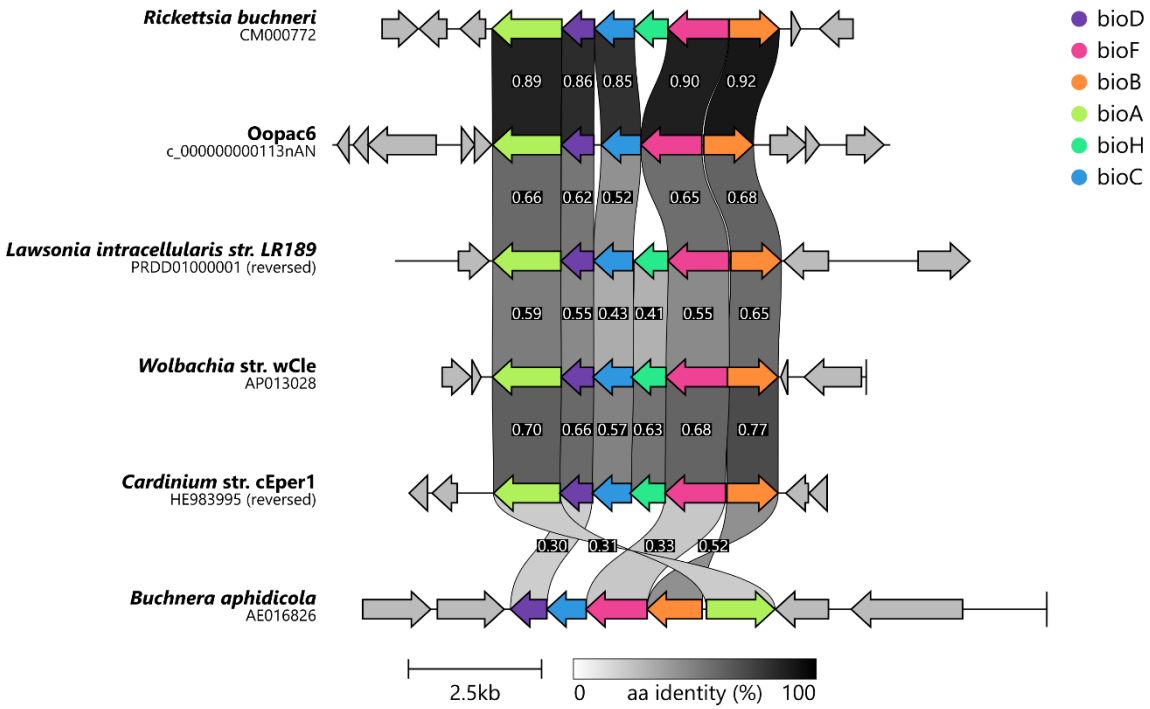


Supplementary Figure 5. Phylogram of a maximum likelihood (ML) tree of 90 *Rickettsia* multilocus profiles. The tree is based on 4 loci, 16S rRNA, 17Kda, gltA, and COI,under a partition model (2,781 bp total). https://doi.org/10.6084/m9.figshare.14865600

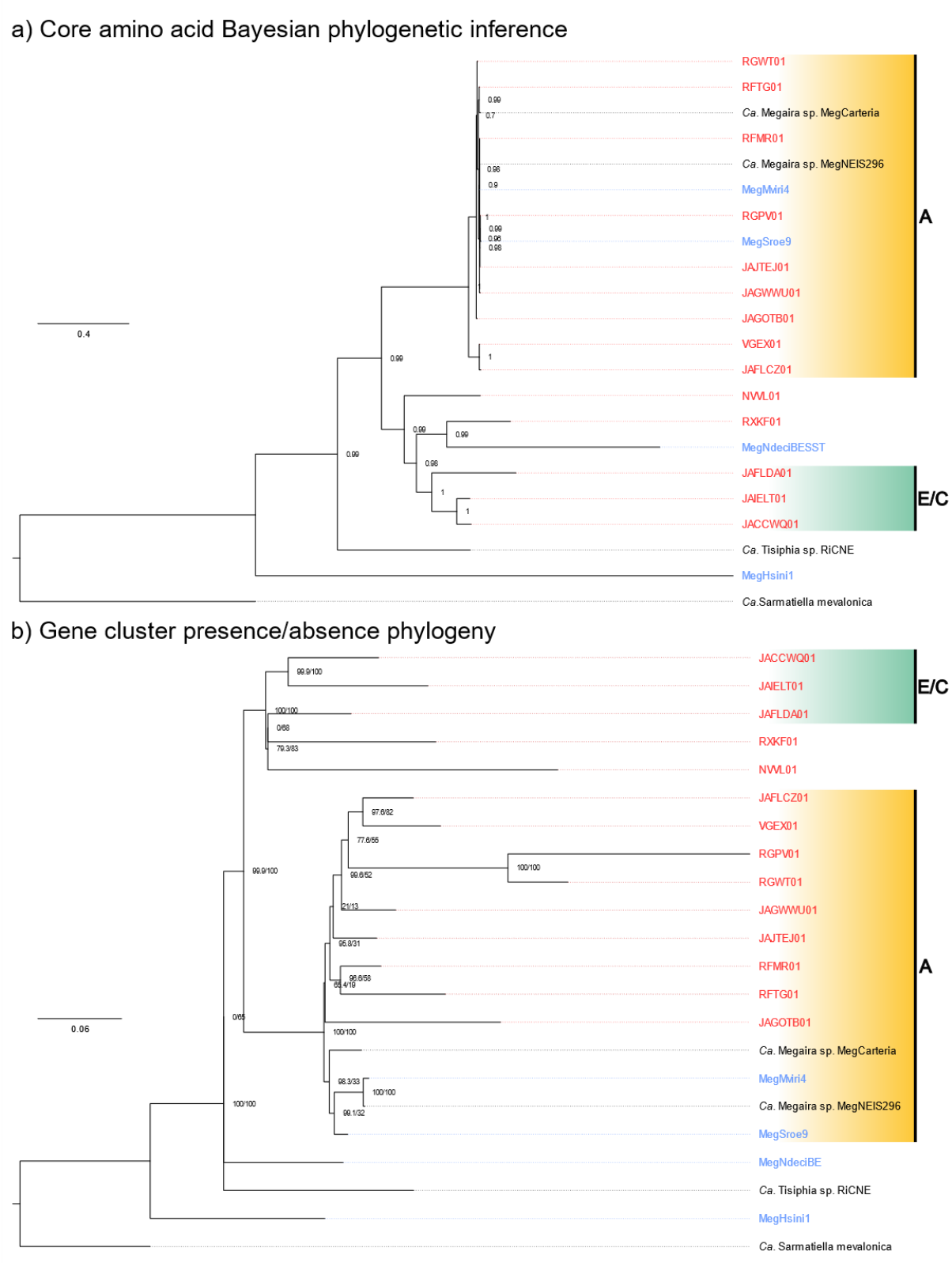
Appendix figure A.5. Phylogram of a maximum likelihood (ML) tree of 90 *Rickettsia* multilocus profiles. The tree is based on 4 loci, 16S rRNA, 17Kda, gltA, and COI, under a partition model (2,781 bp total). See Appendix B.1 for metadata.



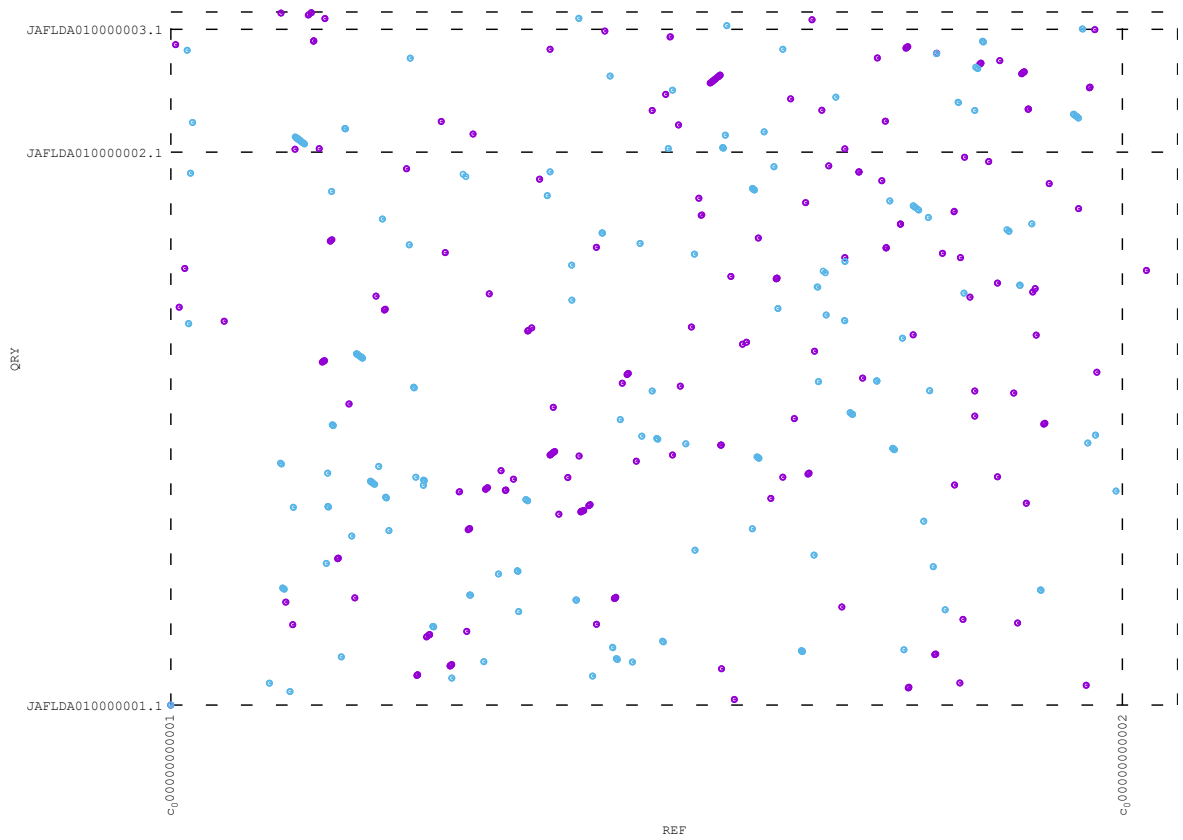
Appendix figure A.6 Pangenome of 103 genomes including *Rickettsia*, Torix, and Ca. Megaira. New genomes are indicated by ●. Each genome displays gene cluster presence/absence and is organised by gene cluster frequency. Group identity was assigned from phylogeny. SFG is Spotted Fever Group. See Appendix B.2 for raw data.



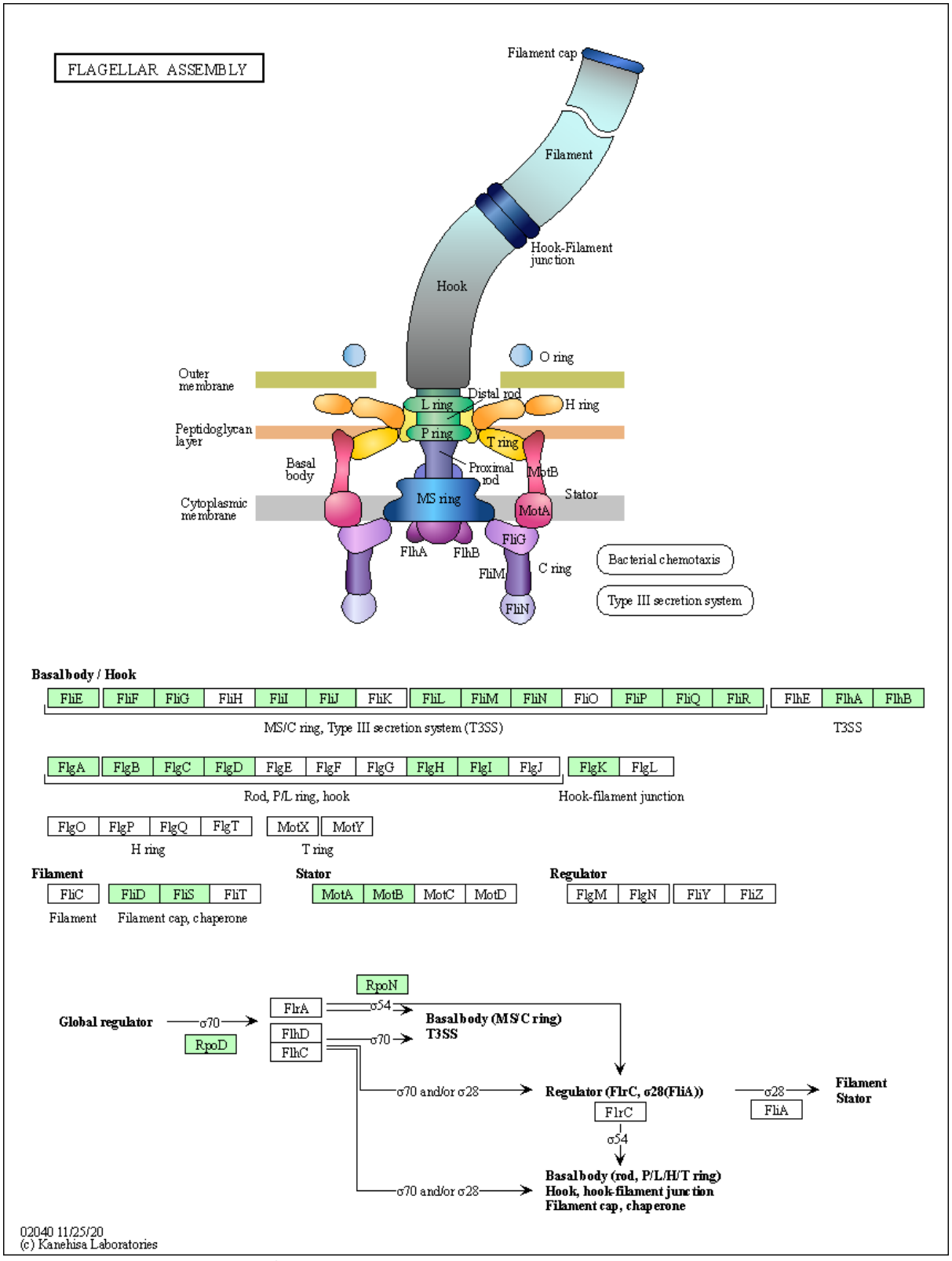
Appendix figure A.7. Biotin operon of the Oopac6 Rhyzobius *Rickettsia* and its surrounding genes compared with other known biotin pathways in other related symbionts. Similarity scores in the black boxes refer to the percentage identity between the genes of the operon above or below it, further illustrated by a greyscale bar. Operons are ordered by overall similarity, showing the closest relationships between all six. See Appendix B.2 for raw data.



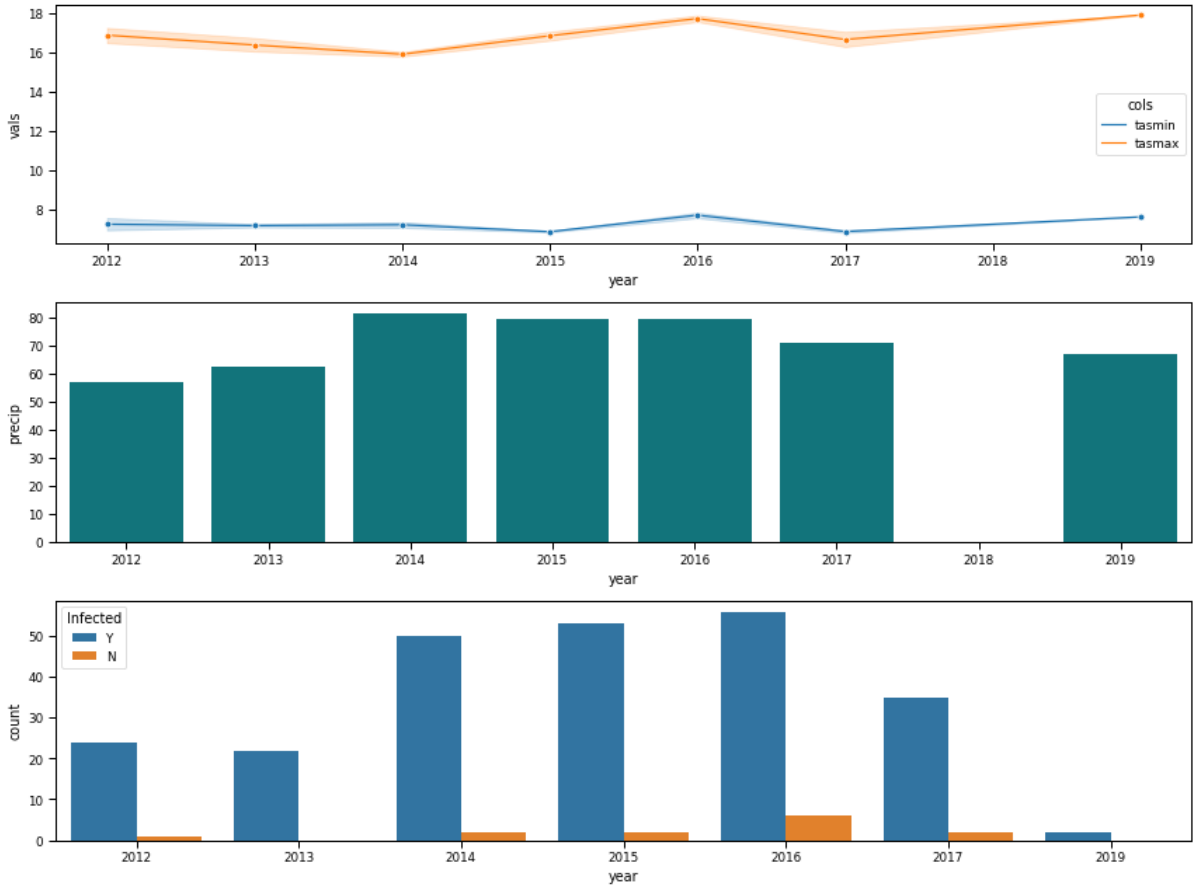
Appendix figure A.8. Supporting '*Ca*. Megaira' phylogenies for main Figure 5.1. a) core amino acid Bayesian phylogenetic tree with nodes displaying posterior probability values, and b) a maximum likelihood tree of gene cluster presence absence with 1000 SH-aLRT and ultrafast bootstraps (UFB). Support for each split is displayed as SH-aLRT/UFB values, with strong support being >=80/>=95. Samples from this study are blue and existing environmental metagenomes are red. See Appendic C.1 for genome metadata.



Appendix figure A.9 Whole genome alignment between JAFLDA01 and MegNEIS296 reveals lack of synteny. Purple are forward matches; blue are reverse matches.



Appendix figure A.10 Presence/Absence of components of the Flagella apparatus for '*Ca*. Megaira', JAFLDA01. Green highlighted boxes indicated genes found in KEGG annotation through Anvi'o. Further metadata for flagella components in this and other '*Ca*. Megaira' assemblies are in Appendix C.1.



Appendix figure A.11. Environmental data for *Anopheles plumbeus* collection sites across Germany extracted from the TerraClim database. (Top) average annual minimum and maximum temperature across all An. plumbeus collection sites in Germany. (Middle) average annual precipitation across all sites. (Bottom) counts of infected and uninfected individuals across all sites where dark blue = infected and orange = uninfected.

Appendix B. Additional data for chapter 2

Appendix B.1 Supplementary metadata tables

Appendix B1 contains all metadata for genomes assembled and genomes used, including accession numbers, CheckM scores and Gtdbtk taxonomy. You will also find supporting data for Chapter 2.

Yellow tabs contain:
> accessions and species information for all whole genomes used
> brief details on hosts and environment for new genomes described in this study
> metadata such as N50s and genome lengths for all new genomes
> completeness scores and assembly levels for all genomes
> information on where all the published genomes were used in this study
> taxonomy calculations from GTDBtk

Red tabs contain:
> Phi scores from reticulate analysis for all core genome clusters extracted from the pangenome aswell as their associated COG and KEGG functions
> Functional enrichment tables exploring the association of different metabolic functions with different groups of bacteria

To access Appendix B1 please follow this link: https://doi.org/10.5281/zenodo.7548404

Appendix B.2 Supplementary raw data

To access raw data used to produce the figures in Chapter 2 and Appendix figures A.1 to A.7 please follow this link: https://doi.org/10.5281/zenodo.7548404

Appendix C. Additional data for Chapter 3

Appendix C.1 Supplementary metadata tables

Appendix C1 contains all metadata for genomes assembled and genomes used, including accession numbers and CheckM scores. You will also find raw data used to produce the figures in Chapter 3.

Yellow tabs contain:
S1 - Meta data tables for draft genomes examined in this study
S2 - Accessions for additional genomes used

Red tabs contain:
S3 - AAI % similarity across Ca. Megaira used in figure 3a
S4 - ANIb % Similarity across Ca. Megaira used in figure 3b
S5 - KEGG ko hits
S6 - KEGG module completeness used in figure 5 and 6
S7 - 16S rRNA accessions used in phylogeny Figure 2
S8 - Gene cluster presence absence matrix used in figure 3.4 and Appendix figure 1
S9 - GTDBtk results for SRA and GenBank Environmental MAGs
S10 - top 10 blastp results for RiPP, NRPS and CDPS regions identified by antiSMASH

To access Appendix C1 please follow this link: https://doi.org/10.5281/zenodo.7548404

Appendix C.2 Supplementary raw data

Clinker diagram interactive html https://doi.org/10.6084/m9.figshare.20424894.v1

Appendix D. Additional data for Chapter 4

Appendix D.1 Supplementary metadata tables

Appendix D1 contains all metadata for genomes assembled and genomes used, including accession numbers, CheckM scores and Gtdbtk taxonomy. You will also find raw data used to produce the figures in Chapter 4.

Yellow tabs contain:
S1 - Metadata for genomes assembled in this study
S2 - Metadata for environmental MAGs recovered from NCBI non redundant sequence database
S3 - Metadata for additional Chlamydiota genomes used

Red tabs contain:
S4 - AAI percentage similarity scores used to produce genera similarity networks
S5 - ANIb percentage similarity scores used to produce species similarity networks
S6 - CRISPRcas finder results
S7 - KEGG pathway hits
S8 - KEGG pathway completeness
S9 - Gene cluster presence absence matrix used in figure 4.3a for Rhabdochlamydiaceae
S10 - Gene cluster presence absence matrix used in figure 4.3b for Simkaniaceae

To access Appendix D1 please follow this link: https://doi.org/10.5281/zenodo.7548404

Appendix D.2 phylogeny partition models

Best-fit model according to BIC:

LG+I+G4:GC_00000198,
Q.yeast+F+R9:GC_00000084,
Q.insect+I+G4:GC_00000212,
LG+R5:GC_00000105,
Q.insect+I+G4:GC_00000162,
LG+I+G4:GC_00000203,
Q.insect+R5:GC_00000179,
Q.yeast+R6:GC_00000158,
Q.yeast+R6:GC_00000243,
Q.yeast+R5:GC_00000110,
Q.insect+I+G4:GC_00000131,
Q.yeast+I+G4:GC_00000255,
Q.pfam+F+I+G4:GC_00000134,
mtInv+F+R5:GC_00000159,
LG+R5:GC_00000199,
Q.insect+I+G4:GC_00000197,
Q.yeast+I+G4:GC_00000187,
Q.yeast+R6:GC_00000144,
Q.insect+I+G4:GC_00000181,
Q.insect+R5:GC_00000128,
Q.insect+R6:GC_00000201,
Q.yeast+I+G4:GC_00000211,
Q.insect+R5:GC_00000125,
Q.yeast+R5:GC_00000126,
Q.insect+F+G4:GC_00000253,
Q.plant+I+G4:GC_00000178,
Q.yeast+I+G4:GC_00000244,
Q.insect+G4:GC_00000103,
LG+R5:GC_00000230,
Q.insect+I+G4:GC_00000246,
Q.yeast+R7:GC_00000252,
Q.insect+R6:GC_00000129

Appendix E. Additional data for Chapter 5

Appendix E.1 Supplementary metadata tables

Appendix E1 contains screening results, environmental data extracted from climate databases and additional genome information.

Yellow tabs contain:
S1 - 'Ca. Tisiphia' in Anopheles plumbeus across Germany. PCR screening and geographic data
S2 - additional genome accessions and metadata
Red tabs contain:
S3 - KEGG completeness
S4 - KEGG ko_hits presence

To access Appendix E1 please follow this link: https://doi.org/10.5281/zenodo.7548404