universität wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation /Title of the Doctoral Thesis

"Uncovering the hidden diversity and ancestral states of Chlamydiae genomes"

verfasst von / submitted by

Stephan Köstlbacher, BSc MSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Wien, 2022 / Vienna 2022

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 794 685 437
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Biology
Betreut von / Supervisor:	Univ.-Prof. Dr. Matthias Horn

Acknowledgements

Somehow I ended up in the deep field of chlamydiae and PVC superphylum research. While chlamydiae first piqued my interest due to their obligate intracellular lifestyle, an intimate symbiosis I have been fascinated with since before I started studying biology, I was captured by their peculiarities and intricate evolutionary heritage. This leads me to acknowledge the two people that put me onto the track I have followed for my master's thesis and graduate studies: **Han-Fei (Allen) Tsao** and Prof. **Matthias Horn**. I still remember vividly two situations that made me want to join Matthias' working group. First was a symbiosis lecture by Matthias introducing the evolution of life, contrasted by concepts of creationism and the flying spaghetti monster. Second was a presentation by Allen in a genomics introductory course, where Allen introduced the endosymbiotic bacterium *Procabacter* for the first time, while using a quote of uncle Ben (Spiderman): "With great power comes great responsibility".

I want to first of all thank **Matthias** for opening up this scientific PhD journey to me, while we were ice skating at Rathausplatz. You guided me through my PhD, while being open to at times seemingly random ideas and keeping me scientifically grounded. You challenged me and helped me ask better questions and helped me develop as a scientist. I am incredibly thankful for your support over the years.

Even though you were not my official supervisor, a shout out to **Astrid**. You simply rock, thanks for your patience, care, wisdom, and support. I will always remember especially fondly the bus ride back from the DOME trip and the subsequent pub visit.

I want to thank all collaborators over the years without whom this work would not have been possible, especially **Jennah** and **Thijs**. Thijs, thanks for hosting my research stay in Wageningen 2019, as well as the scientific meetings which were always insightful and taught me a lot. Also thanks for later on hiring me. Jennah, I learned a lot from you, thanks for the years of collaboration and friendship and the weekly meetings that contributed to me getting through the lockdowns of 2020 and 2021. I further want to thank my lovely new colleagues in Wageningen for being so welcoming and also

regularly reminding me to finish my thesis. Shout out to **Burak**, **Dani**, **Eric**, **Felix**, **Guillaume**, **Kassi**, **Jolanda**, **Patricia**, **Victor**.

For bioinformatics support and maintenance of the LISC infrastructure I want to thank **Florian** and **Thomas**. Additionally I want to thank Thomas for his bioinformatics courses, lectures, and general insights, which made an important impact on my understanding and my research.

Next I want to thank all my teammates in Matthias' group over the years. First, I want to thank the above-mentioned **Allen**, who directly supervised me during my Masters thesis. Where do you get the energy to be so kind, supportive, and friendly all the time? Thank you for your friendship and kisses to the girls. I want to thank **Daryl** and **Frederick** who showed me the ropes when I started working in genomics and phylogenetics, and became important collaborators in this work after they moved on to new labs. A place in my heart will always be occupied by **Florian**, **Ilias**, **Karin**, **Nadja**, **Patrick**, **Paul**, **Tamara**. **Florian** (Wascher), the few times I was in the lab were always hilarious thanks to you. Ilias, you were slightly mean but inspiring. **Karin**, you cleaned up my first centrifuge spill and were so kind about it, you will always be remembered and missed. **Nadja**, we go back to our bachelor studies, to keep it brief: we had a blast, even when we didn't, thank you so much. **Patrick**, I did not actually hate you when we first met. My friend, we'll meet again at the foosball table and I'll kick your ass next time! Until then, stay "noch gschissener". **Paul**, we went through thick and thin together and I will always remember Bubbles and Charlotte, cheers mate, it was a great PhD time together. **Tamara**, often quiet, but to the point and always helpful, thank you.

Angelika, you need your own section. It was wonderful to get to know you and work with you. I learned so much thanks to you and I admire your pure curiosity and strength. Thank you for your friendship and know that you have a special place in my heart!

There were so many DOMEies (google suggests "DUMMies") I had the pleasure of meeting over the years, big thanks to all of you. I want to thank especially **Alejandro, Andrew, Anouk, Avi, Bela**, **Buck**, **Claus**, **Conny**, **Craig**, **Chris**, **Isabelle**, **Fathima**,

Franziska, **Jay**, **Jessica**, **Johanna**, **Julius**, **Ken**, **Lena**, **Lena**, **Linda**, **Lisa**, **Michael**, **Michaela**, **Michi**, **Nora**, **Orest**, **Pala**, **Petra**, **Sascha**, **Steph**, **Thomas**. Thanks, y'all.

Bela Hausmann, my work (ex-?)husband and great friend. We had so much fun over the years, but also quite some scientific discussions on the couch in the social room. I loved working together with you, sharing coffee, tequila, beers and ideas. Also thanks for getting me into R. **Andrew**, I will keep it short, as you hate blabla. Great friend, love your input, let's have beers as often as possible. Bussi. **Petra**, I never stood a chance. Thanks for all the great times together and your support and advice, even if I didn't ask for it. You always believed in me and for that I am grateful. I look forward to having my ears ring again. **Chris**, even if I tried, I could never forget you, as long as rain falls down on Africa. You are great and time spent with you is never wasted. **Babsi**, we bring out the worst in each other, it is great. **Michael**, how lucky I was to start my PhD with you. I love you, man. I am grateful for your friendship, Sanja, and all the great times just drinking beer and talking about whatever, be it serious or a random Manga. You opened my mind to various topics. Talk soon. **Michi**, funny how we met more than ten years ago, didn't get along, but were somehow reunited through DOME. So glad that happened and thanks for your friendship, and being there for me in my happy and sad moments. You are a big inspiration to me.

I want to thank all my friends for their support but this thing is too long already. Special thanks to **Astrid**, **Georg**, **Lisa**, **Michael**, **Michi**, **Martin**, **Mathias**, **Jakob**, **Pablo**. All of you rock!

I am incredibly grateful to my family, who have always been there for me. I want to thank my Mum, **Monika**, who glued plants with me in "Unterstufe" into my "Herbarium" until late at night, because I somehow didn't manage in time. I want to thank my Dad, **Franz**, for introducing me to microscopy and the viennese forest. Besides raising me to be a somewhat reasonable person, you two cultivated my interest in science, even though I just recently have come to understand that. I want to thank my brother, **Felix**, one of the greatest people I know. From fixing my car to staying up late to help with moving. I will always remember the times we built LEGO, and read manga while listening to music. A final thanks to my grandparents.

I want to thank my extended family, i.e. families Zink, Schwarz and Winkler. **Doris**, you are the best mother in law I could have hoped for. Thanks to **Sabine** and **Ernst Johann** for lovely dinners and conversations. **Gerd** and **Marko**, my brothers by choice. I love you and thank you.

Lastly, I want to thank my beloved wife, **Isabelle**. You are my love and light. You made me a better person but also a better scientist. I am incredibly lucky and thankful that you came into my life. You always believe in me, motivate me, inspire me, and generally make life better. Looking forward to what is to come and facing it together with you.

Table of contents

CHAPTER I

Introduction

Outline of thesis

Introduction

Bacteria are among the smallest and oldest living entities of our planet. These fascinating unicellular organisms thrive under all kinds of conditions almost anywhere on our planet, from deep sea sediments, to the clouds, or even within us humans. Their 3 billion years of evolutionary history (Betts et al. 2018) evolved a yet to be fully conceived metabolic and physiological versatility. Single prokaryotic organisms have been shown to be able to grow on spilled oil or survive and thrive in extreme environments, be it temperature, toxicity, or even radioactivity. Over time, countless co-operations among different microorganisms have evolved. Without the billions of bacteria working together in the human gut, for example, humans would be unable to efficiently process food and lack essential vitamins.

However, nature brought forth even more intimate interactions, in which one organism thrives within the cells of another organism, called endosymbiosis (Greek: ἔνδον inner, internal; συμβίωσις, living together). This phenomenon led to the origin of eukaryotes, life forms with a nucleus, like us humans (Sagan 1967). The best supported scenario suggests that an ancestor of archaea, the other prokaryotic domain besides bacteria, engulfed an alphaproteobacterium up to two billion years ago, which in turn became the first mitochondrion in the ancestor of all eukaryotes (Martin, Garg, and Zimorski 2015; Spang et al. 2019; Betts et al. 2018). Only around a couple of hundred million years later, an ancestral cyanobacterial organism was acquired by a lineage of primordial eukaryotes (Betts et al. 2018), birthing photosynthetic eukaryotes, a lineage of which evolved into plants.

Since then, many bacterial lineages have adopted an intracellular lifestyle in a wide array of host organisms. From the host perspective endosymbionts can have positive, neutral, or negative effects, ranging from bacteria of the genus *Buchnera* that provide essential metabolites to their aphid hosts (Douglas 1998), to organisms that cause diseases in humans like pneumonia, urinary tract infection, or meningitis. This thesis focuses on Chlamydiae, a phylum of unique obligate endosymbiotic bacteria, likely spanning the whole spectrum of host influence, with global importance on health and beyond. Employing state of the art genomic and phylogenetic techniques I intended

to tackle some of the long-standing evolutionary and ecological questions concerning this phylum.

Enter the chlamydiae

In 1907 the german Ludwig Halberstädter and bohemian Stanislaus von Prowazek set out on a research expedition on syphilis to Java, Indonesia. However, whilst investigating conjunctival scrapings from an experimentally infected orangutan, they discovered unknown, "coated" intracellular parasites. They could neither classify these parasites with bacteria or protozoa and therefore suggested placing them between the aforementioned two as a new group called "Chlamydozoa" (Greel; χλαμύς covering, coat; ζῷον animal) (Halberstädter and von Prowazek 1907).

Only decades later, this pathogen that is today known as *Chlamydia trachomatis* was identified as an actual member of bacteria (Moulder 1964). Today *C. trachomatis* is recognized as the leading infectious cause of preventable blindness (Taylor et al. 2014) and the most frequent sexually transmitted disease of bacterial origin (Newman et al. 2015) with global implications on human health and 127 million new cases in 2016 alone (World Health Organization 2018). Other members of the genus *Chlamydia* are represented by the important human pathogen *Chlamydia pneumoniae* that causes respiratory disease and animal pathogens like *Chlamydia psittaci*, which can occasionally cause severe zoonotic infections (Sixt 2020).

For a long time Chlamydiaceae, the family containing all members of the genus *Chlamydia*, were thought to be the only representatives of the phylum Chlamydiae (Taylor-Brown et al. 2015). The picture started to change with the discovery of a contaminant in a human cell culture, that represented a putative novel genus in the Chlamydiaceae, the enigmatic microorganism 'Z', today known as *Simkania negevensis* Z (Kahane, Metzer, and Friedman 1995; Kahane et al. 1993). Few years later another bacterial endosymbiont was discovered in the protist *Acanthamoeba* sp. that, together with 'Z', formed a sister clade to Chlamydia spp. and was named *Parachlamydia acanthamoebae* BN9 (Amann et al. 1997). Like its relative, *S. negevensis* turned out to be fully capable of thriving in acanthamoebae (Kahane et al. 2001). *S. negevensis* and *P. acanthamoebae* were the first representatives of the families Simkaniaceae and Parachlamydiaceae, respectively (Figure 1). (Taylor-Brown et al. 2015). Chlamydiae not affiliated with the Chlamydiaceae are collectively referred to as "environmental

chlamydiae", as they are largely associated with protists living in the environment. Further novel chlamydiae were isolated from diverse sources, i.e. *Waddlia chondrophila* (Waddliaceae) from an aborted bovine fetus (Rurangirwa et al. 1999), *Neochlamydia hartmannellae* (Horn et al. 2000) from a water conduit system, *Protochlamydia amoebophila* UWE25 (both Parachlamydiaceae) (Collingro et al. 2005) from soil, and *Criblamydia sequanensis* (Criblamydiaceae) using a protist co-culture approach from freshwater (Thomas, Casson, and Greub 2006). While some of these environmental chlamydiae have been associated with human and animal disease (Taylor-Brown et al. 2015), their primary hosts seem to be ubiquitous unicellular eukaryotes, i.e. protists (Collingro, Köstlbacher, and Horn 2020; Horn 2008). However not all environmental chlamydiae are able to thrive in known protists. The cultured representative of the Rhabdochlamydiaceae, *Rhabdochlamydia porcellionis* (Kostanjšek et al. 2004), was isolated from the hepatopancreas of the crustacean host *Porcellio scaber* and could only be cultured in Sf9 insect cells so far (Sixt et al. 2013).

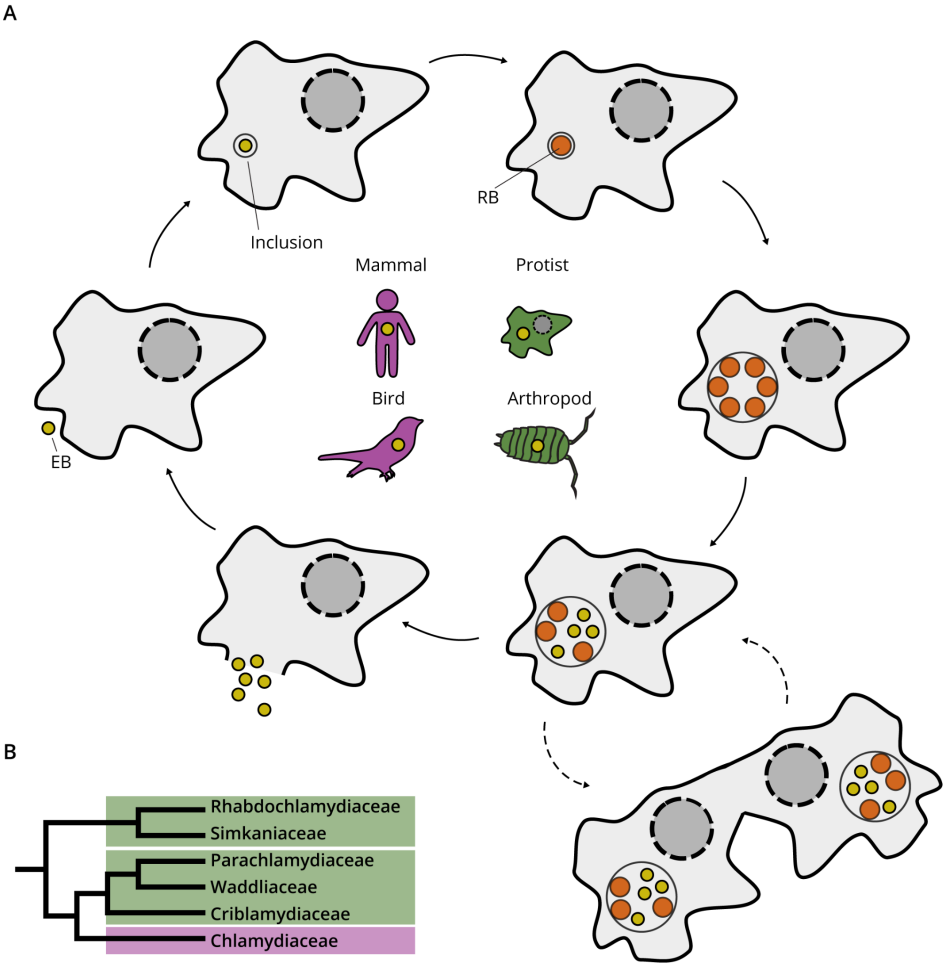


Figure 1. Developmental cycle, hosts, and phylogeny of cultured chlamydiae. *(A) The developmental cycle starts when the elementary body (EB) enters the host cell and differentiates into the reticulate body (RB). The RB either thrives directly in the cytoplasm or within a host-derived vacuole termed inclusion. Upon multiple rounds of replication, RBs differentiate into EBs and leave the host via cell lysis or exocytosis. In the absence of cell lysis, some chlamydiae can be vertically transmitted, meaning they are present in the daughter cells upon host cell division. Cartoons in the center show representative hosts of cultured chlamydiae colored by their affiliation with Chlamydiaceae in purple or environmental chlamydiae in green. (B) Schematic phylogeny of cultured chlamydiae families based on recent phylogenomic studies (Köstlbacher et al. 2020; Dharamshi et al. 2020). Clade colors of chlamydiae family affiliation as indicated in (A).*

So far all cultured chlamydiae have an obligate intracellular lifestyle. Despite the variety of hosts, a common denominator is their biphasic life-cycle consisting of an extracellular, infective stage, the elementary body (EB) and an intracellular replicative stage, the reticulate body (RB, Figure 1) (Elwell, Mirrashidi, and Engel 2016). The EB is taken up by the host cell and resides within a host-derived vacuole, called the inclusion. Upon differentiation into the RB, replication is initiated and chlamydiae divide by polarized budding (Ouellette, Lee, and Cox 2020). RBs then differentiate into EBs and either leave the host cell by lysis, or exocytosis. Of note, some chlamydiae do not (always) reside in inclusions (Horn et al. 2000; Bou Khalil et al. 2016; Benamar et al. 2017; Bou Khalil et al. 2017; Nylund et al. 2018) or might not leave the host cell at all in the case of *Neochlamydia* sp. S13 *(Okude et al. 2020)*. The chlamydial biphasic life-cycle seems to be coupled to a biphasic metabolism, as exemplified by proteome or transcriptome studies in *C. trachomatis* (Grieshaber et al. 2018; Saka et al. 2011; Omsland et al. 2012) and *P. amoebophila* (König et al. 2017). Energy metabolism is roughly separated into host ATP-scavenging termed “energy parasitism” in RBs and an endogenous glucose-based ATP production that gets upregulated in the transitional stages from RBs to EBs.

The genome sequencing of *C. trachomatis* unveiled the genomic background of this highly effective pathogen. This 1,042,519 base pair (bp) genome was lacking many biosynthetic capabilities (Stephens et al. 1998). Instead it encoded the necessary enzymes to utilize and interconvert host derived metabolites and numerous virulence-associated proteins, indicating host dependence for its proliferation. Genome sequencing of the second important model organism *C. pneumoniae* revealed a slightly

bigger genome (1,230,230 bp), but high conservation of gene content and organization between the two *Chlamydia* species (Kalman et al. 1999). This narrow understanding of chlamydial genomes got expanded when cultured environmental chlamydiae representatives were sequenced, revealing generally larger genomes of up to 3,072,383 bp in size (*P. acanthamoebae* UV-7) (Horn et al. 2004; Greub et al. 2009; Bertelli et al. 2010; Collingro et al. 2011). Comparative analysis of available chlamydial genomes revealed a generally larger biosynthetic potential in environmental chlamydiae, indicating adaptation to a less homeostatic environment than their pathogenic sisters (Horn et al. 2004; Collingro et al. 2011). It did however support the biphasic endosymbiotic lifestyle of the phylum, exemplified by the conserved and strictly vertically inherited type III secretion system (T3SS) (Collingro et al. 2011) and highly conserved regulatory components of genes involved in the exploitation of the intracellular niche (Domman and Horn 2015).

Chlamydiae are one of the most successful groups of bacteria specialized in inhabiting the interior of eukaryotic cells. As different chlamydiae can infect a wide variety of eukaryotes, ranging from unicellular protists to our very own bodies, they make an intriguing object to study the evolution of endosymbiosis and the different flavours it comes in. This thesis aims to illuminate the evolutionary trajectories shaping chlamydiae lineages. Chapter II holds a brief outline of the manuscripts produced in the course of this thesis. In Chapter III I explore how chlamydial plasmids have been ancient partners of their chlamydial hosts. In Chapter IV I explore untapped chlamydiae genome diversity and shed light on novel chlamydial lifestyles. In Chapter V I set out to reconstruct chlamydiae evolutionary history back to before eukaryotes evolved.

Chlamydiae, plasmids and horizontal gene transfer

The totality of chromosomes and other genetic elements of an organism are referred to as its genome. In bacteria, genome sizes can vary substantially and typically range in free-living bacteria between 3 and 6 megabase pairs (Mbp) (Qin et al. 2019). Chlamydiaceae genomes with approximately one Mbp in size fall distinctively below this range and even the environmental chlamydiae with up to three Mbp barely scratch the realm of most free-living bacteria. The small genomes of chlamydiae can be attributed to their intracellular lifestyle. Small population sizes and strong genetic drift result in a bias for gene loss and lead to reduced genomes (Moran 2002). In all domains of life and

especially in prokaryotes there is, however, a powerful mechanism that can balance gene loss. The acquisition of new genes from other, potentially distantly related organisms, is called horizontal gene transfer (HGT, also known as lateral gene transfer) and brings innovation to recipient genomes, therefore expanding their genetic repertoire (Ochman, Lawrence, and Groisman 2000). However, due to the isolation of chlamydiae from other prokaryotes due to their obligate intracellular lifestyle, i.e. likely HGT donors, leads to a reduced influx of new genes.

A peculiar finding during the investigation of chlamydiae genomes was the discovery of plasmids in many chlamydiae that, despite a great disparity in size (7.5 - 132 kb), shared chlamydiae specific plasmid genes, some of which are even present on all chlamydial chromosomes (Collingro et al. 2011). Plasmids are extrachromosomal genetic elements that can vary in size (see chlamydiae) and contain genes involved in functions ranging from metabolism and defense to virulence (Dietel, Kaltenpoth, and Kost 2018). An inherent nature of many plasmids is their transferability, meaning they can pose as vectors for (large scale) HGT (Smillie et al. 2010). In Chlamydiaceae plasmids have mostly been associated with virulence. For environmental chlamydiae plasmids are generally larger (> 20 kbp), little is known concerning their role in chlamydial life.

In the third chapter of this thesis we systematically analyze the evolutionary aspects of chlamydial plasmids in relation to their host bacterial chromosomes (Köstlbacher et al. 2020). We first characterized the general characteristics of plasmid genetic makeup and retrieved first evidence for a coevolution of plasmids with their chlamydial hosts. Using statistical approaches we defined gene modules on plasmids that encompass highly conserved to highly specific genes. We establish monophyly of key conserved plasmid genes by phylogenetic analysis, indicating a common ancestry of chlamydiae plasmids. With the reconciliation of gene trees and species trees we demonstrate a predominantly vertical mode of inheritance in conserved plasmid genes, while other plasmid genes seem to be enriched in HGT. By integrating all phylogenetic data we synthesize a common origin of chlamydial plasmids followed by coevolution with chlamydial hosts that led to increased HGT rates between chlamydiae, potentially historically balancing the deletion bias in these endosymbionts. This study adds the presumably oldest documented system of host-plasmid coexistence and evolution and such evolutionary dynamics might be important to other endosymbiotic or highly specialized systems.

Chlamydiae and hidden lifestyles in the environment

In the general public, chlamydiae are almost exclusively perceived as human pathogens, best known for the sexually transmitted disease caused by *C. trachomatis*. While the role of human and animal pathogens should not be downplayed (see above), the majority of chlamydiae do not directly affect, i.e. infect us. From the perspective of putative hosts, protists - common hosts of many environmental chlamydiae - globally outweigh humans in biomass ~67 times (Bar-On, Phillips, and Milo 2018). The 16S rRNA gene sequence is a molecular marker commonly used to study prokaryotic diversity and distribution in the environment. Mining efforts of publicly available 16S rRNA amplicon data from all sorts of environments and all over the globe databases yielded colossal estimates of the existence of hundreds of chlamydial families occurring in diverse environments (Lagkouvardos et al. 2014; Collingro, Köstlbacher, and Horn 2020), standing in stark contrast to the cultured representatives of only six families. However even for free-living bacteria it is known that there are vastly more bacterial species in the environment than can be cultured, a phenomenon referred to as “the great plate count anomaly” (Staley and Konopka 1985). For a long time, our knowledge about chlamydiae was based on their culturability and therefore limited to only a few representatives. With the rise of metagenomics, a technique to sequence genomes from environmental samples in an untargeted and culture-independent fashion, and related techniques, that picture slowly started to change, yielding genomes of uncultured chlamydial clades from previously untapped environments and hosts (Baker et al. 2015; Collingro et al. 2017; Dharamshi et al. 2020; Pillonel, Bertelli, and Greub 2018; Taylor-Brown et al. 2016, 2017, 2018). For the first time, this enables researchers to make inferences about the genetic potential and biology of these uncultured lineages.

In the fourth chapter of this thesis we harvest the power of metagenomics to add the metagenome-assembled genomes (MAGs) of an additional 82 uncultured chlamydiae. Phylogenomic analysis shows that our dataset almost doubles the chlamydial phylogenetic diversity. We underline the conservation of the genetic potential for the chlamydial endosymbiotic, biphasic life-cycle, and uncover novel chlamydial lifestyles in the environment. This study provides the most comprehensive picture of the chlamydiae pangenome and its phylum-wide dynamics put into an environmental context yet.

Chlamydiae and where they come from

Throughout the previous sections chlamydiae were referred to as a phylum, without explaining the meaning of this term. A phylum represents a taxonomic rank that, like any taxonomic rank, summarizes a group of organisms by their shared properties. In modern prokaryotic taxonomy the shared property that is commonly applied for classification is evolutionary relationship (Mayr 1981). In the case of chlamydiae this means that all members of this phylum descended from one common ancestor. While the definitions for most taxonomic ranks are somewhat arbitrary, the phylum rank indicates a very ancient common heritage of a group. Estimates place the last common chlamydiae ancestor (LCCA) into a time period of 1-2 billion years ago (Gya) and evidence suggests it already possessed the genetic toolkit necessary for the conserved chlamydial endosymbiotic lifestyle (Horn et al. 2004; Kamneva et al. 2012; Subtil, Collingro, and Horn 2014; Betts et al. 2018). This time frame overlaps with some of the most important events in earth's biological history, i.e. eukaryogenesis and the primary plastid acquisition in the ancestors of extant plants. And indeed, chlamydiae have been suggested to be involved in both events, e.g. (Cenci et al. 2017; Stairs et al. 2020).

To formulate improved scenarios for these ancient events we need a better understanding of the LCCA. To make inferences about LCCA, however, we have to know the genomic background it evolved from, i.e., we need information about the closest relatives of chlamydiae. Together with the bacterial phyla Planctomycetes and Verrucomicrobia, chlamydiae form the PVC superphylum (Wagner and Horn 2006), the name deriving from the acronym of its founding phyla (Figure 3). Since the initial definition of the PVC superphylum it has gained additional members in the Lentisphaerae, Kirimatiellaeota, and potentially additional candidate phyla like Omnitrophica (Rivas-Marín and Devos 2018; Astrid Collingro, Köstlbacher, and Horn 2020). Unlike chlamydiae, the other members of the superphylum are largely free-living bacteria, globally involved in biogeochemical cycles and of importance for biotechnology and medicine (Devos and Ward 2014; Wagner and Horn 2006; Rivas-Marín and Devos 2018)

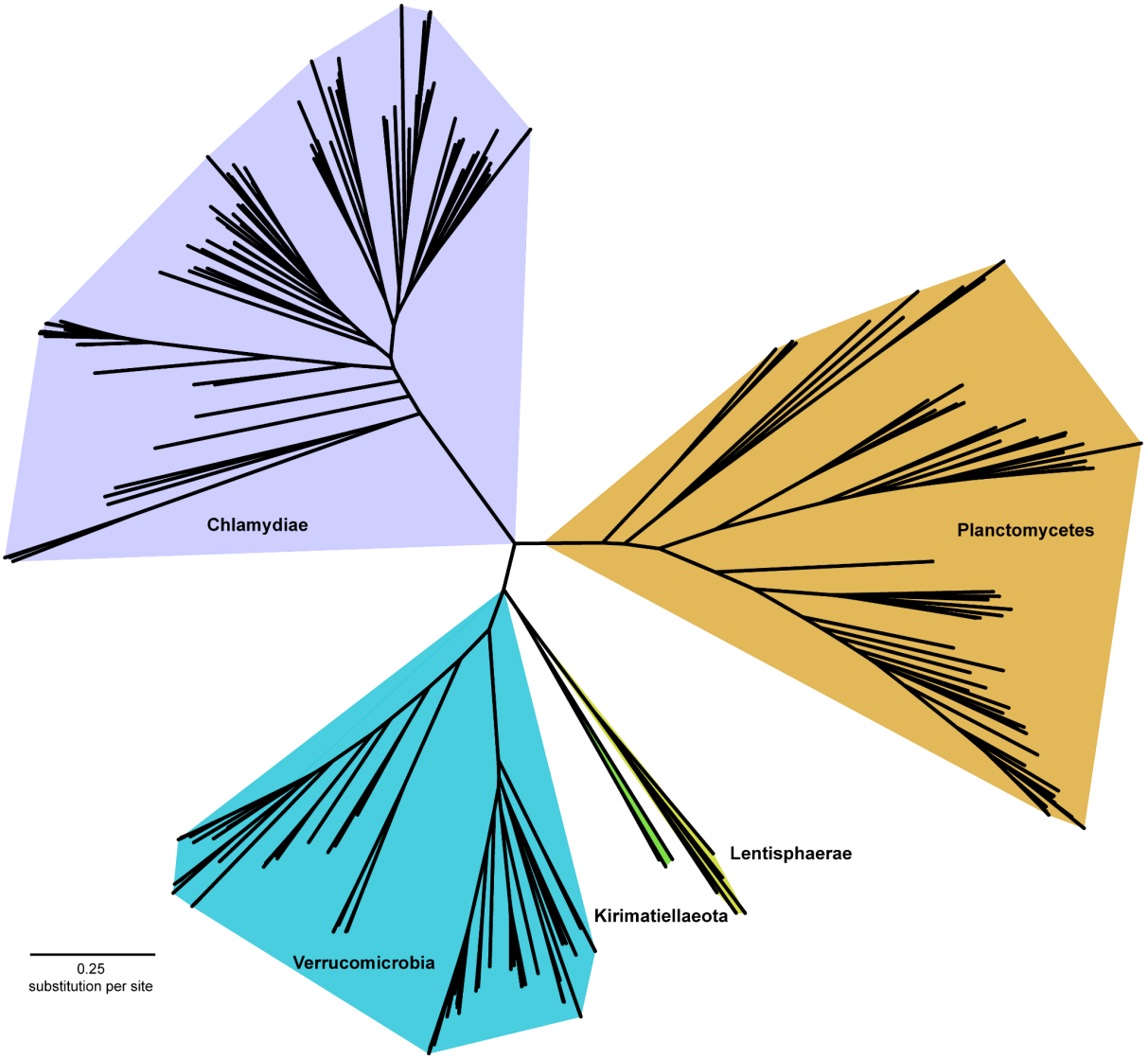


Figure 3: Phylogenomic species tree of PVC bacteria. *Adapted from Chapter IV (Dharamshi, Köstlbacher et al., under review). Highlighted clades indicate phylum level clades. Species tree is based on 43 conserved bacterial markers (Parks et al. 2015). 0.25 substitutions per site in the alignment.*

In the fifth chapter of this thesis we collect the to date most comprehensive genome collection of PVC bacteria (Figure 3). Using state of the art phylogenomic techniques, including the application of complex phylogenetic models and modeling of compositional bias, we reconstruct an updated picture of PVC species evolution. An ancestral state reconstruction of the last free living common ancestor of chlamydiae and their sister phylum Verrucomicrobia suggests a motile, facultative anaerobic ancestor with fermentative metabolism coupled to molecular hydrogen production with some traits pointing at a pre-adaptation to symbiotic lifestyle. Down the road LCCA

underwent a reduction of the electron transport chain as well as biosynthetic capabilities coupled to the gain of a T3SS as could be explained by the switch to obligate endosymbiosis. We further show that the phylum Chlamydiae since then underwent complex genome dynamics, including substantial contribution of HGT from all domains of life.

References

Amann, R., N. Springer, W. Schönhuber, W. Ludwig, E. N. Schmid, K. D. Müller, and R. Michel. 1997. “Obligate Intracellular Bacterial Parasites of Acanthamoebae Related to Chlamydia Spp.” *Applied and Environmental Microbiology* 63 (1): 115–21.

Baker, Brett J., Cassandre Sara Lazar, Andreas P. Teske, and Gregory J. Dick. 2015. “Genomic Resolution of Linkages in Carbon, Nitrogen, and Sulfur Cycling among Widespread Estuary Sediment Bacteria.” *Microbiome* 3 (April): 14.

Bar-On, Yinon M., Rob Phillips, and Ron Milo. 2018. “The Biomass Distribution on Earth.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (25): 6506–11.

Benamar, Samia, Jacques Y. Bou Khalil, Caroline Blanc-Tailleur, Melhem Bilen, Lina Barrassi, and Bernard La Scola. 2017. “Developmental Cycle and Genome Analysis of Sp. Nov. a New Species in the Family.” *Frontiers in Cellular and Infection Microbiology* 7 (August): 385.

Betts, Holly C., Mark N. Puttick, James W. Clark, Tom A. Williams, Philip C. J. Donoghue, and Davide Pisani. 2018. “Integrated Genomic and Fossil Evidence Illuminates Life’s Early Evolution and Eukaryote Origin.” *Nature Ecology & Evolution*. https://doi.org/10.1038/s41559-018-0644-x.

Bou Khalil, Jacques Y., Samia Benamar, Jean-Pierre Baudoin, Olivier Croce, Caroline Blanc-Tailleur, Isabelle Pagnier, Didier Raoult, and Bernard La Scola. 2016. “Developmental Cycle and Genome Analysis of ‘Rubidus Massiliensis,’ a New Vermamoeba Vermiformis Pathogen.” *Frontiers in Cellular and Infection Microbiology* 6 (March): 31.

Bou Khalil, Jacques Y., Samia Benamar, Fabrizio Di Pinto, Caroline Blanc-Tailleur, Didier Raoult, and Bernard La Scola. 2017. “Protochlamydia Phocaeensis Sp. Nov., a New Chlamydiales Species with Host Dependent Replication Cycle.” *Microbes and Infection / Institut Pasteur* 19 (6): 343–50.

Cenci, Ugo, Debashish Bhattacharya, Andreas P. M. Weber, Christophe Colleoni, Agathe Subtil, and Steven G. Ball. 2017. “Biotic Host–Pathogen Interactions As Major Drivers of Plastid Endosymbiosis.” *Trends in Plant Science*. https://doi.org/10.1016/j.tplants.2016.12.007.

Collingro, Astrid, Stephan Köstlbacher, and Matthias Horn. 2020. “Chlamydiae in the Environment.” *Trends in Microbiology*, June. https://doi.org/10.1016/j.tim.2020.05.020.

Collingro, Astrid, Stephan Köstlbacher, Marc Mussmann, Ramunas Stepanauskas, Steven J. Hallam, and Matthias Horn. 2017. “Unexpected Genomic Features in Widespread Intracellular Bacteria: Evidence for Motility of Marine Chlamydiae.” *The ISME Journal* 11 (10): 2334–44.

Collingro, Astrid, Patrick Tischler, Thomas Weinmaier, Thomas Penz, Eva Heinz, Robert C. Brunham, Timothy D. Read, et al. 2011. “Unity in Variety--the Pan-Genome of the Chlamydiae.” *Molecular Biology and Evolution* 28 (12): 3253–70.

Collingro, Astrid, Elena R. Toenshoff, Michael W. Taylor, Thomas R. Fritsche, Michael Wagner, and Matthias Horn. 2005. “‘Candidatus Protochlamydia Amoebophila’, an Endosymbiont of Acanthamoeba Spp.” *International Journal of Systematic and Evolutionary Microbiology*. https://doi.org/10.1099/ijs.0.63572-0.

Devos, Damien P., and Naomi L. Ward. 2014. “Mind the PVCs.” *Environmental*

Microbiology 16 (5): 1217–21.
Dharamshi, Jennah E., Daniel Tamarit, Laura Eme, Courtney W. Stairs, Joran Martijn, Felix Homa, Steffen L. Jørgensen, Anja Spang, and Thijs J. G. Ettema. 2020. "Marine Sediments Illuminate Chlamydiae Diversity and Evolution." *Current Biology: CB* 30 (6): 1032–48.e7.
Dietel, Anne-Kathrin, Martin Kaltenpoth, and Christian Kost. 2018. "Convergent Evolution in Intracellular Elements: Plasmids as Model Endosymbionts." *Trends in Microbiology* 26 (9): 755–68.
Domman, D., and M. Horn. 2015. "Following the Footsteps of Chlamydial Gene Regulation." *Molecular Biology and Evolution* 32 (12): 3035–46.
Douglas, A. E. 1998. "Nutritional Interactions in Insect-Microbial Symbioses: Aphids and Their Symbiotic Bacteria Buchnera." *Annual Review of Entomology* 43: 17–37.
Elwell, Cherilyn, Kathleen Mirrashidi, and Joanne Engel. 2016. "Chlamydia Cell Biology and Pathogenesis." *Nature Reviews. Microbiology* 14 (6): 385–400.
Grieshaber, Scott, Nicole Grieshaber, Hong Yang, Briana Baxter, Ted Hackstadt, and Anders Omsland. 2018. "Impact of Active Metabolism on Chlamydia Trachomatis Elementary Body Transcript Profile and Infectivity." *Journal of Bacteriology* 200 (14). https://doi.org/10.1128/JB.00065-18.
Halberstädter, L., and S. V. von Prowazek. 1907. "Über Zelleinschlüsse Parasitärer Natur Beim Trachom." *Arbeiten Aus Dem Kaiserlichen Gesundheitsamte* 26: 44–47.
Horn, Matthias. 2008. "Chlamydiae as Symbionts in Eukaryotes." *Annual Review of Microbiology* 62: 113–31.
Horn, Matthias, Astrid Collingro, Stephan Schmitz-Esser, Cora L. Beier, Ulrike Purkhold, Berthold Fartmann, Petra Brandt, et al. 2004. "Illuminating the Evolutionary History of Chlamydiae." *Science* 304 (5671): 728–30.
Horn, Matthias, Michael Wagner, Karl-Dieter Müller, Ernst N. Schmid, Thomas R. Fritsche, Karl-Heinz Schleifer, and Rolf Michel. 2000. "Neochlamydia Hartmannellae Gen. Nov., Sp. Nov. (Parachlamydiaceae), an Endoparasite of the Amoeba Hartmannella Vermiformis." *Microbiology* 146 (Pt 5) (May): 1231–39.
Kahane, S., B. Dvoskin, M. Mathias, and M. G. Friedman. 2001. "Infection of Acanthamoeba Polyphaga with Simkania Negevensis and S. Negevensis Survival within Amoebal Cysts." *Applied and Environmental Microbiology* 67 (10): 4789–95.
Kahane, S., R. Gonen, C. Sayada, J. Elion, and M. G. Friedman. 1993. "Description and Partial Characterization of a New Chlamydia-like Microorganism." *FEMS Microbiology Letters* 109 (2-3): 329–33.
Kahane, S., E. Metzer, and M. G. Friedman. 1995. "Evidence That the Novel Microorganism 'Z' May Belong to a New Genus in the Family Chlamydiaceae." *FEMS Microbiology Letters* 126 (2): 203–7.
König, Lena, Alexander Siegl, Thomas Penz, Susanne Haider, Cecilia Wentrup, Julia Polzin, Evelyne Mann, Stephan Schmitz-Esser, Daryl Domman, and Matthias Horn. 2017. "Biphasic Metabolism and Host Interaction of a Chlamydial Symbiont." *mSystems* 2 (3). https://doi.org/10.1128/mSystems.00202-16.
Kostanjšek, Rok, Jasna Štrus, Damjana Drobne, and Gorazd Avguštin. 2004. "'Candidatus Rhabdochlamydia Porcellionis', an Intracellular Bacterium from the Hepatopancreas of the Terrestrial Isopod Porcellio Scaber (Crustacea: Isopoda)." *International Journal of Systematic and Evolutionary Microbiology*. https://doi.org/10.1099/ijs.0.02802-0.
Köstlbacher, Stephan, Astrid Collingro, Tamara Halter, Daryl Domman, and Matthias Horn. 2020. "Coevolving Plasmids Drive Gene Flow and Genome Plasticity in Host-Associated Intracellular Bacteria." *Current Biology: CB*, October. https://doi.org/10.1016/j.cub.2020.10.030.

Lagkouvardos, Ilias, Thomas Weinmaier, Federico M. Lauro, Ricardo Cavicchioli, Thomas Rattei, and Matthias Horn. 2014. “Integrating Metagenomic and Amplicon Databases to Resolve the Phylogenetic and Ecological Diversity of the Chlamydiae.” *The ISME Journal* 8 (1): 115–25.

Martin, William F., Sriram Garg, and Verena Zimorski. 2015. “Endosymbiotic Theories for Eukaryote Origin.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 370 (1678): 20140330.

Mayr, E. 1981. “Biological Classification: Toward a Synthesis of Opposing Methodologies.” *Science* 214 (4520): 510–16.

Moran, Nancy A. 2002. “Microbial Minimalism: Genome Reduction in Bacterial Pathogens.” *Cell* 108 (5): 583–86.

Moulder, James W. 1964. *The Psittacosis Group as Bacteria*. New York: Wiley.

Newman, Lori, Jane Rowley, Stephen Vander Hoorn, Nalinka Saman Wijesooriya, Magnus Unemo, Nicola Low, Gretchen Stevens, Sami Gottlieb, James Kiarie, and Marleen Temmerman. 2015. “Global Estimates of the Prevalence and Incidence of Four Curable Sexually Transmitted Infections in 2012 Based on Systematic Review and Global Reporting.” *PloS One* 10 (12): e0143304.

Nylund, Are, Dario Pistone, Christiane Trösse, Steffen Blindheim, Linda Andersen, and Heidrun Plarre. 2018. “Genotyping of Candidatus Syngnamydia Salmonis (chlamydiales; Simkaniaceae) Co-Cultured in Paramoeba Perurans (amoebozoa; Paramoebidae).” *Archives of Microbiology* 200 (6): 859–67.

Ochman, H., J. G. Lawrence, and E. A. Groisman. 2000. “Lateral Gene Transfer and the Nature of Bacterial Innovation.” *Nature* 405 (6784): 299–304.

Okude, Miho, Junji Matsuo, Tomohiro Yamazaki, Kentaro Saito, Yoshokazu Furuta, Shinji Nakamura, Jeewan Thapa, Torahiko Okubo, Hideaki Higashi, and Hiroyuki Yamaguchi. 2020. “Distribution of Amoebal Endosymbiotic Environmental Chlamydia Neochlamydia S13 via Amoebal Cytokinesis.” *Microbiology and Immunology*, December. https://doi.org/10.1111/1348-0421.12871.

Omsland, Anders, Janet Sager, Vinod Nair, Daniel E. Sturdevant, and Ted Hackstadt. 2012. “Developmental Stage-Specific Metabolic and Transcriptional Activity of Chlamydia Trachomatis in an Axenic Medium.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (48): 19781–85.

Ouellette, Scot P., Junghoon Lee, and John V. Cox. 2020. “Division without Binary Fission: Cell Division in the FtsZ-Less.” *Journal of Bacteriology* 202 (17). https://doi.org/10.1128/JB.00252-20.

Parks, Donovan H., Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W. Tyson. 2015. “CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes.” *Genome Research* 25 (7): 1043–55.

Pillonel, Trestan, Claire Bertelli, and Gilbert Greub. 2018. “Environmental Metagenomic Assemblies Reveal Seven New Highly Divergent Chlamydial Lineages and Hallmarks of a Conserved Intracellular Lifestyle.” *Frontiers in Microbiology* 9 (February): 79.

Qin, Qi-Long, Yi Li, Lin-Lin Sun, Zhi-Bin Wang, Shi Wang, Xiu-Lan Chen, Aharon Oren, and Yu-Zhong Zhang. 2019. “Trophic Specialization Results in Genomic Reduction in Free-Living Marine Bacteria.” *mBio* 10 (1). https://doi.org/10.1128/mBio.02545-18.

Rivas-Marín, Elena, and Damien P. Devos. 2018. “The Paradigms They Are a-Changin’: Past, Present and Future of PVC Bacteria Research.” *Antonie van Leeuwenhoek* 111 (6): 785–99.

Rurangirwa, F. R., P. M. Dilbeck, T. B. Crawford, T. C. McGuire, and T. F. McElwain. 1999. “Analysis of the 16S rRNA Gene of Micro-Organism WSU 86-1044 from an Aborted

Bovine Foetus Reveals That It Is a Member of the Order Chlamydiales: Proposal of Waddliaceae Fam. Nov., Waddlia Chondrophila Gen. Nov., Sp. Nov." *International Journal of Systematic Bacteriology* 49 Pt 2 (April): 577–81.

Sagan, L. 1967. "On the Origin of Mitosing Cells." *Journal of Theoretical Biology* 14 (3): 255–74.

Saka, Hector A., J. Will Thompson, Yi-Shan Chen, Yadunanda Kumar, Laura G. Dubois, M. Arthur Moseley, and Raphael H. Valdivia. 2011. "Quantitative Proteomics Reveals Metabolic and Pathogenic Properties of Chlamydia Trachomatis Developmental Forms." *Molecular Microbiology* 82 (5): 1185–1203.

Sixt, Barbara S. 2020. "Host Cell Death during Infection with Chlamydia: A Double-Edged Sword." *FEMS Microbiology Reviews*, September. https://doi.org/10.1093/femsre/fuaa043.

Sixt, Barbara S., Rok Kostanjšek, Azra Mustedanagic, Elena R. Toenshoff, and Matthias Horn. 2013. "Developmental Cycle and Host Interaction of Rhabdochlamydia Porcellionis, an Intracellular Parasite of Terrestrial Isopods." *Environmental Microbiology* 15 (11): 2980–93.

Smillie, Chris, M. Pilar Garcillán-Barcia, M. Victoria Francia, Eduardo P. C. Rocha, and Fernando de la Cruz. 2010. "Mobility of Plasmids." *Microbiology and Molecular Biology Reviews*. https://doi.org/10.1128/mmbr.00020-10.

Spang, Anja, Courtney W. Stairs, Nina Dombrowski, Laura Eme, Jonathan Lombard, Eva F. Caceres, Chris Greening, Brett J. Baker, and Thijs J. G. Ettema. 2019. "Proposal of the Reverse Flow Model for the Origin of the Eukaryotic Cell Based on Comparative Analyses of Asgard Archaeal Metabolism." *Nature Microbiology* 4 (7): 1138–48.

Stairs, Courtney W., Jennah E. Dharamshi, Daniel Tamarit, Laura Eme, Steffen L. Jørgensen, Anja Spang, and Thijs J. G. Ettema. 2020. "Chlamydial Contribution to Anaerobic Metabolism during Eukaryotic Evolution." *Science Advances* 6 (35): eabb7258.

Staley, J. T., and A. Konopka. 1985. "Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats." *Annual Review of Microbiology*. https://doi.org/10.1146/annurev.mi.39.100185.001541.

Taylor-Brown, Alyce, Nathan L. Bachmann, Nicole Borel, and Adam Polkinghorne. 2016. "Culture-Independent Genomic Characterisation of Candidatus Chlamydia Sanzinia, a Novel Uncultivated Bacterium Infecting Snakes." *BMC Genomics* 17 (September): 710.

Taylor-Brown, Alyce, Trestan Pillonel, Andrew Bridle, Weihong Qi, Nathan L. Bachmann, Terrence L. Miller, Gilbert Greub, et al. 2017. "Culture-Independent Genomics of a Novel Chlamydial Pathogen of Fish Provides New Insight into Host-Specific Adaptations Utilized by These Intracellular Bacteria." *Environmental Microbiology* 19 (5): 1899–1913.

Taylor-Brown, Alyce, Trestan Pillonel, Gilbert Greub, Lloyd Vaughan, Barbara Nowak, and Adam Polkinghorne. 2018. "Metagenomic Analysis of Fish-Associated Ca. Parilichlamydiaceae Reveals Striking Metabolic Similarities to the Terrestrial Chlamydiaceae." *Genome Biology and Evolution* 10 (10): 2587–95.

Taylor-Brown, Alyce, Lloyd Vaughan, Gilbert Greub, Peter Timms, and Adam Polkinghorne. 2015. "Twenty Years of Research into Chlamydia-like Organisms: A Revolution in Our Understanding of the Biology and Pathogenicity of Members of the Phylum Chlamydiae." *Pathogens and Disease* 73 (1): 1–15.

Taylor, Hugh R., Matthew J. Burton, Danny Haddad, Sheila West, and Heathcote Wright. 2014. "Trachoma." *The Lancet* 384 (9960): 2142–52.

Thomas, Vincent, Nicola Casson, and Gilbert Greub. 2006. "Criblamydia Sequanensis, a

New Intracellular Chlamydiales Isolated from Seine River Water Using Amoebal Co-Culture." *Environmental Microbiology* 8 (12): 2125–35.
Wagner, Michael, and Matthias Horn. 2006. "The Planctomycetes, Verrucomicrobia, Chlamydiae and Sister Phyla Comprise a Superphylum with Biotechnological and Medical Relevance." *Current Opinion in Biotechnology* 17 (3): 241–49.
World Health Organization. 2018. "Report on Global Sexually Transmitted Infection Surveillance 2018." 2018. https://apps.who.int/iris/bitstream/handle/10665/277258/9789241565691-eng.pdf?sequence=5&isAllowed=y.

CHAPTER II

Overview of

publications and manuscripts

Chapter III

Manuscript title:

Coevolving Plasmids Drive Gene Flow and Genome Plasticity in Host-Associated Intracellular Bacteria

Author names:

Stephan Köstlbacher, Astrid Collingro, Tamara Halter, Daryl Domman, Matthias Horn

Reference:

Current Biology, 2020, https://doi.org/10.1016/j.cub.2020.10.030.

Author contributions:

S.K., A.C., D.D., and M.H. conceptualized the study. S.K. and A.C. performed comparative genomic analysis. S.K. performed phylogenetic analyses and gene tree-species tree reconciliation analyses. S.K., A.C., T.H., and M.H. interpreted the results. All authors wrote and edited the manuscript.

Chapter IV

Manuscript title:

Pangenomics reveals alternative environmental lifestyles among chlamydiae

Author names:

Stephan Köstlbacher, Astrid Collingro, Tamara Halter, Frederik Schulz, Sean P. Jungbluth, and Matthias Horn

Reference:

Nature Communications, 2021, https://doi.org/10.1038/s41467-021-24294-3.

Author contributions:

S.K., M.H., S.P.J., and F.S. conceptualized the study. S.K. and A.C. performed comparative genomic analysis. S.K., A.C., T.H., and M.H. interpreted the results. S.K. and M.H. wrote and all authors edited the manuscript.

Chapter V

Manuscript title:

Gain of symbiotic traits underpins evolutionary transitions across the phylum Chlamydiae

Author names:

Jennah E. Dharamshi‡, Stephan Köstlbacher‡, Max-Emil Schön, Astrid Collingro, Thijs J. G. Ettema‡ , Matthias Horn‡

‡ Equal contribution

‡ Equal contribution

Reference:

Submitted to Nature Microbiology on November 18, 2021.

Author contributions:

S.K., J.D., T.J.G.E., and M.H. conceptualized the study. S.K. and J.D. performed all analysis with support by M.E.S. and A.C.. S.K., J.D., T.J.G.E., and M.H. interpreted the results. S.K., J.D., T.J.G.E., and M.H. wrote and all authors edited the manuscript. The contributions of J.E.D. and S.K., and T.J.G.E. and M.H. should be regarded as equal, respectively.

CHAPTER III

Coevolving Plasmids Drive Gene Flow and Genome Plasticity in Host-Associated Intracellular Bacteria

Authors:

Stephan Köstlbacher, Astrid Collingro, Tamara Halter, Daryl Domman, Matthias Horn

Published in:

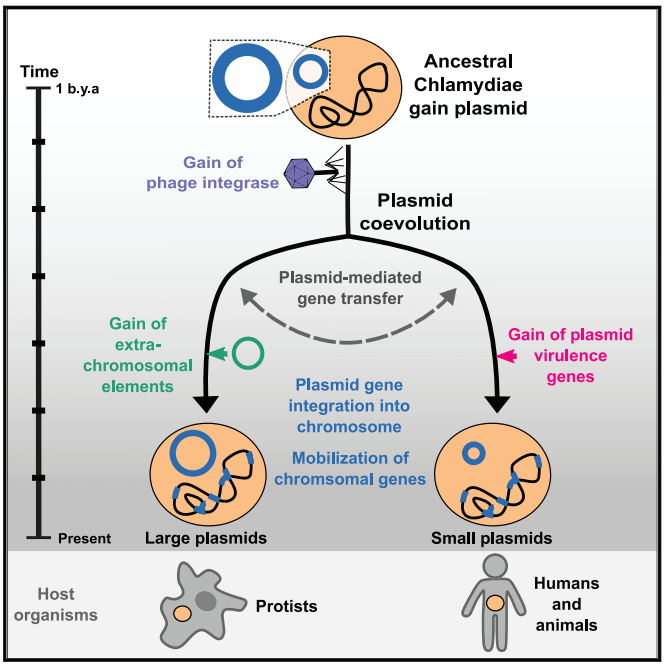
Current Biology (2020)

Article

Current Biology

Coevolving Plasmids Drive Gene Flow and Genome Plasticity in Host-Associated Intracellular Bacteria

Graphical Abstract



Highlights

- Chlamydial plasmids coevolved with their bacterial hosts over a billion years
- Recombination with extrachromosomal elements and viruses shaped plasmid gene content
- Plasmid-mediated chromosomal gene mobilization and transfer drove genome evolution
- Plasmids contributed to adaptation of chlamydiae to diverse eukaryotic hosts

Authors

Stephan Köstlbacher, Astrid Collingro, Tamara Halter, Daryl Domman, Matthias Horn

Correspondence

matthias.horn@univie.ac.at

In Brief

Köstlbacher et al. illustrate how plasmids of intracellular bacteria in the phylum Chlamydiae have coevolved with their hosts over a billion years. By mobilizing chromosomal genes, plasmids contributed to host adaptation and might have mitigated the degenerative effects of Muller's ratchet in this group of intracellular pathogens and symbionts.

Köstlbacher et al., 2021, Current Biology *31*, 1–12
January 25, 2021
https://doi.org/10.1016/j.cub.2020.10.030

CellPress

Please cite this article in press as: Köstlbacher et al., Coevolving Plasmids Drive Gene Flow and Genome Plasticity in Host-Associated Intracellular Bacteria, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.10.030

Current Biology

CellPress
OPEN ACCESS

Article

Coevolving Plasmids Drive Gene Flow and Genome Plasticity in Host-Associated Intracellular Bacteria

Stephan Köstlbacher,[1] Astrid Collingro,[1] Tamara Halter,[1] Daryl Domman,[2,3] and Matthias Horn[1,4,*]
[1]University of Vienna, Centre for Microbiology and Environmental Systems Science, Division of Microbial Ecology, Althanstrasse 14, Vienna 1090, Austria
[2]Wellcome Sanger Institute, Parasites and Microbes Programme, Hinxton, Cambridge CB10 1SA, UK
[3]Center for Global Health, Department of Internal Medicine, University of New Mexico Health Sciences Center, Albuquerque, NM 87131, USA
[4]Lead Contact
*Correspondence: matthias.horn@univie.ac.at
https://doi.org/10.1016/j.cub.2020.10.030

SUMMARY

Plasmids are important in microbial evolution and adaptation to new environments. Yet, carrying a plasmid can be costly, and long-term association of plasmids with their hosts is poorly understood. Here, we provide evidence that the Chlamydiae, a phylum of strictly host-associated intracellular bacteria, have coevolved with their plasmids since their last common ancestor. Current chlamydial plasmids are amalgamations of at least one ancestral plasmid and a bacteriophage. We show that the majority of plasmid genes are also found on chromosomes of extant chlamydiae. The most conserved plasmid gene families are predominantly vertically inherited, while accessory plasmid gene families show significantly increased mobility. We reconstructed the evolutionary history of plasmid gene content of an entire bacterial phylum over a period of around one billion years. Frequent horizontal gene transfer and chromosomal integration events illustrate the pronounced impact of coevolution with these extrachromosomal elements on bacterial genome dynamics in host-dependent microbes.

INTRODUCTION

Plasmids are extrachromosomal genetic elements encoding a wide range of genes that allow organisms from all domains of life to adapt to different stresses or niches.[1] Ranging in size from below 1 kb to more than 2.5 Mb, the effect of plasmids on their hosts is often poorly understood, as most plasmids have not been fully characterized.[2] Among bacteria, plasmids spread genetic information within and between populations, strains, species, and even more distantly related microbes.[3] This mechanism of horizontal gene transfer (HGT) is not only an important driver of the evolution of natural microbial populations, but plasmids are also essential tools in diverse applications in genetics and biotechnology, and they have important implications in public health. Major human pathogens, such as enterohemorrhagic *E. coli* (EHEC), emerge through plasmid acquisition.[4] Importantly, plasmid-mediated transfer of antibiotic resistance is a key factor in the spread of antibiotic resistance and the increase in multi-resistant bacterial pathogens.[5]

Acquisition of a plasmid implies gain of genetic potential, yet there are usually negative side effects. A number of plasmids encode toxin-antitoxin (TA) modules—genetic elements that encode a protein capable of inhibiting cell growth and an antitoxin that counteracts the toxin.[6] Loss of such a plasmid, therefore, can be detrimental to the host. Even in the absence of TA systems, production of plasmid proteins (as well as maintenance and repair of plasmid DNA requires host resources) occupies cellular machinery such as ribosomes and disrupts the cellular environment.[7–9] Newly acquired plasmids are thus lost quickly without selection for plasmid-encoded genes.[10] In addition, lateral transfer of plasmids and compensatory mutations that reduce the costs for plasmid maintenance are important factors in plasmid persistence.[10–12] During longer phases of host-plasmid coexistence, plasmids can coevolve with their hosts,[13–17] and plasmid-mediated HGT has been proposed to represent a coevolutionary process.[18] Plasmids can be altered through coresiding mobile genetic elements like integrative conjugative elements (ICEs), transposons, phages, or even other plasmids.[19,20] Longer histories of host-plasmid coexistence are often found in strictly intracellular bacteria. The potentially longest described case is found in *Buchnera* species, primary endosymbionts of aphids, which seem to be coevolving with their plasmids for up to 70 My.[21] Around 25 My years of association with their 8 kb plasmids is found in *Riesa* species, endosymbionts of blood-sucking lice parasitizing primates.[22]

To investigate the association of bacteria with plasmids over an extended evolutionary time period, we chose the Chlamydiae, a phylum of obligate intracellular pathogens and symbionts that have engaged in a host-associated lifestyle around a billion years ago.[23–25] A strictly host dependent lifestyle has severe evolutionary consequences for bacterial genomes. Due to small population sizes, genetic drift, and limited access to large gene

CURBIO 16985

Please cite this article in press as: Köstlbacher et al., Coevolving Plasmids Drive Gene Flow and Genome Plasticity in Host-Associated Intracellular Bacteria, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.10.030

pools, endosymbiont genomes accumulate deleterious mutations eventually leading to genome size reduction.[26–28] These constraints make obligate intracellular bacteria an interesting subject to study genome and plasmid evolution.[29] Most human and animal pathogens classified in the family Chlamydiaceae carry a conserved 7.5 kb plasmid with eight plasmid encoded proteins, referred to as plasmid glycoproteins Pgp1–8.[30–33] These low copy number plasmids[34] represent an important virulence factor in the natural host.[35–38] Accumulating evidence indicates coevolution of chlamydial plasmids and chromosomes within the family Chlamydiaceae.[39–42] HGT among *Chlamydia trachomatis* strains is common, and both intra- and inter-species HGT has been demonstrated experimentally,[43–45] yet the role of plasmids therein is unclear. Intriguingly, all other chlamydial families with cultured representatives have members with plasmids up to 145.3 kb in size.[46–52] Despite the heterogeneity in plasmid size and gene content, based on the presence of conserved plasmid genes it has been proposed that all chlamydial plasmids originated from a single plasmid in the last common ancestor (LCA) of the phylum Chlamydiae.[50]

In this study, we aimed to recapitulate more than a billion years of plasmid gene content evolution in the bacterial phylum Chlamydiae. We demonstrate that a core set of plasmid genes is conserved, despite the plasticity of plasmid size across the phylum. We investigated the shared ancestry and putative origin of key core plasmid genes by integrating virus and plasmid sequence databases in our evolutionary analysis. We present evidence for an ancient acquisition of the chlamydial plasmid and find that the evolutionary trajectory of plasmid genes is characterized by frequent chromosomal integration and HGT. We propose that vertically inherited plasmids have been important partners in genome evolution in these strictly intracellular bacteria, facilitating genome evolution in the face of small population sizes and genetic drift.

RESULTS AND DISCUSSION

Diversity and Conservation of Chlamydial Plasmids

The monophyly of the phylum chlamydiae and its major families is well supported by phylogenomic analysis in previous studies[24,50,53] and confirmed with our comprehensive dataset comprising high-quality genomes of plasmid-containing and plasmid-less chlamydiae (Figure S1; Data S1). First, we compared the chlamydial plasmids in our dataset to known plasmids from other bacterial phyla and found that their size of 7.5–145 kb falls into the range of described bacterial plasmids (Figure S2A). The GC content is with 28% to 44% slightly lower than in most other phyla (Figure S2B), and on average 4.8% lower than the GC content of the host chromosomes (Pearson's correlation coefficient $r = 0.603$, $p = 0.005$; Figure S2C), a feature also seen in other host-associated bacteria.[54,55] Importantly, statistical analysis shows that the trinucleotide composition of most chlamydial plasmids matches that of the respective chromosomes, indicating plasmid acquisition of the host genomic signature (Data S2A).[56]

We next performed *de novo* clustering of the 124,183 proteins encoded on chlamydial plasmids and chromosomes in our dataset into 22,565 gene families. The plasmid proteome comprising in total 733 proteins is represented in 302 chlamydial plasmid gene families, whose members are encoded on at least two plasmids, or on one plasmid and one chromosome (Data S2B). Surprisingly, this amounts to more than 30% of the gene content among all chlamydial plasmids (Figure 1). The plasmids of the Chlamydiaceae and of the fish pathogen *Clavichlamydia salmonicola* are all smaller than 9 kb in size but are comprised of 100% conserved chlamydial plasmid genes as observed previously.[30] The large plasmids (>20 kb) include between 42% (*Protochlamydia naegleriophila*) and 89% (*Criblamydia sequanensis*) plasmid genes. Despite the variability in size, chlamydial plasmids are thus remarkably well conserved with respect to their gene content across all of the seven chlamydial families analyzed.

Taken together, acquisition of the chromosome trinucleotide signature and the high proportion of genes shared among chlamydial plasmids and between plasmids and chromosomes provide first evidence for an extended period of coexistence and a shared evolutionary history of chlamydial plasmids with their bacterial hosts.

A Mosaic Plasmid Building Set

To understand better the evolutionary building blocks that formed the extant chlamydial plasmids, we focused on the most highly conserved plasmid gene families and asked whether it was possible to recover a common plasmid gene set. Consistent with previous observations, chlamydial plasmids lack a pronounced backbone, i.e., a larger set of genes present in all chlamydial plasmids.[47] Nonetheless, there are common gene families between subsets of plasmids (Figures 2B and S2D). To investigate the relations between these gene families, we performed partial correlation network analysis. Briefly, we measured the degree of association between gene families based on their occurrence patterns on chlamydial plasmids. Of 151 gene families occurring on at least two plasmids, 92 were included in the network because they showed a statistically significant correlation (false discovery rate [FDR] corrected $p \leq 0.05$) based on their presence/absence on diverse chlamydial plasmids (Figure 2A). Using an algorithm for the identification of densely connected regions in the correlation network, these conserved plasmid gene families clustered into three statistically significant subgraphs (with $p \leq 0.05$). Based on their abundance and predicted functions, we refer to these subgraphs as (1) core group, (2) type IV secretion (T4SS) group, and (3) phage group, respectively (Figure 2A).

The core group represents the largest and most conserved set of plasmid gene families, comprising 46 (15.2%) of all conserved plasmid gene families (Figure 2B; Data S2C). Many of these have characteristic plasmid functions, and five of seven gene families that make up the Chlamydiaceae plasmid (Figure 2B) belong to this group. This includes the helicase Pgp1 essential for plasmid maintenance in the Chlamydiaceae,[57] the predicted plasmid partitioning protein ParA/Pgp5, the integrases Pgp7 and Pgp8, as well as Pgp2 and Pgp6, two proteins of unknown function, which are essential for plasmid maintenance.[57,58] Some of these genes are known to modulate gene expression,[58,59] and in *C. trachomatis* two highly expressed antisense sRNAs are encoded in *pgp5* and *pgp7/8*.[60,61] Other gene families in the core group function in stress response or are involved in plasmid persistence, such as an efflux transporter and a TA system (Figure 2B).

Please cite this article in press as: Köstlbacher et al., Coevolving Plasmids Drive Gene Flow and Genome Plasticity in Host-Associated Intracellular Bacteria, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.10.030

Current Biology
Article

CellPress
OPEN ACCESS

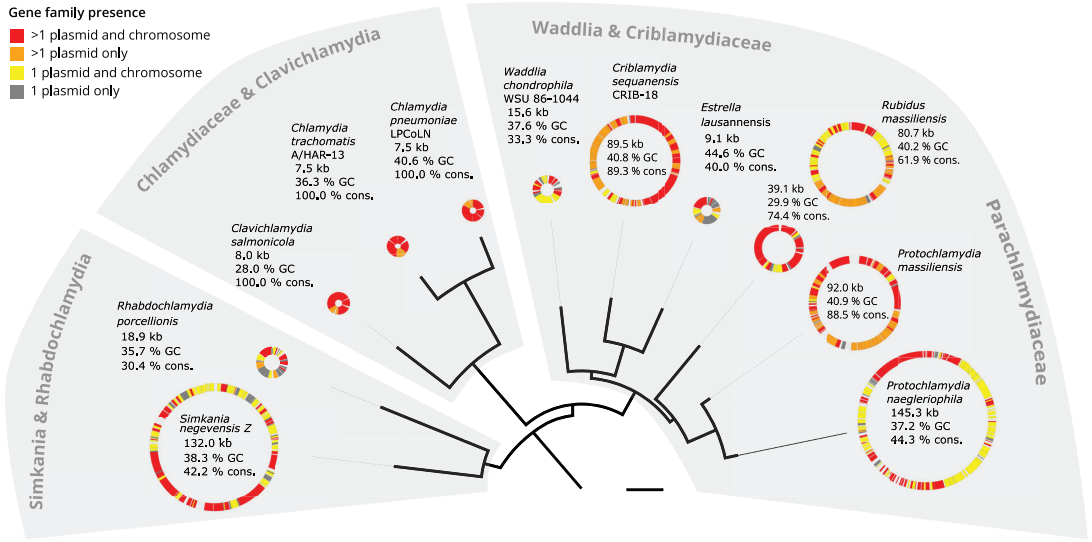


Figure 1. Highly Conserved Gene Content of Chlamydial Plasmids

Chlamydial species tree relating chlamydial plasmids and conservation of plasmid genes. Circles depict chlamydial plasmids and include plasmid size in kilobases, GC content in percent, and the proportion of conserved plasmid-encoded genes. Genes present on other chlamydial plasmids and chromosomes are shown in red, genes present on other chlamydial plasmids only in orange, genes present on one plasmid only and on other chlamydial chromosomes in yellow, and genes present only on a single plasmid within the chlamydiae are gray. Bar indicates 0.2 substitutions per site. See Figure S1 for the full species tree. See also Figure S2 and Data S1 and S2.

The T4SS group comprises a set of gene families associated with type IV secretion. The role of the chlamydial T4SS is still unclear, but it is monophyletic based on phylogenetic analysis of the outer membrane protein TraN[50] (Figure S3A) and occurs on the plasmids of *S. negevensis*, *P. naegleriophila*, and *R. massiliensis*. The T4SS is integrated into the genome of some members of the Parachlamydiaceae and Simkaniaceae (Figure S3B) and was suggested to originate from an Alphaproteobacteria donor.[50]

Finally, the phage group contains gene families almost exclusively present on the *P. massiliensis* and *C. sequanensis* plasmids, which encode among others a phage terminase (OG0004061), tail tip protein L (OG0004637), and RNA polymerase-associated protein Gp33 (OG0000297), indicating a putative phage origin for these gene families (Data S2C).

Overall, we identified a mosaic plasmid gene set consisting of a large core and two gene sets likely originating from other plasmids and prophages. One conceivable scenario would be that the core gene set is a remnant of an ancestral plasmid acquired by an early chlamydiae ancestor.

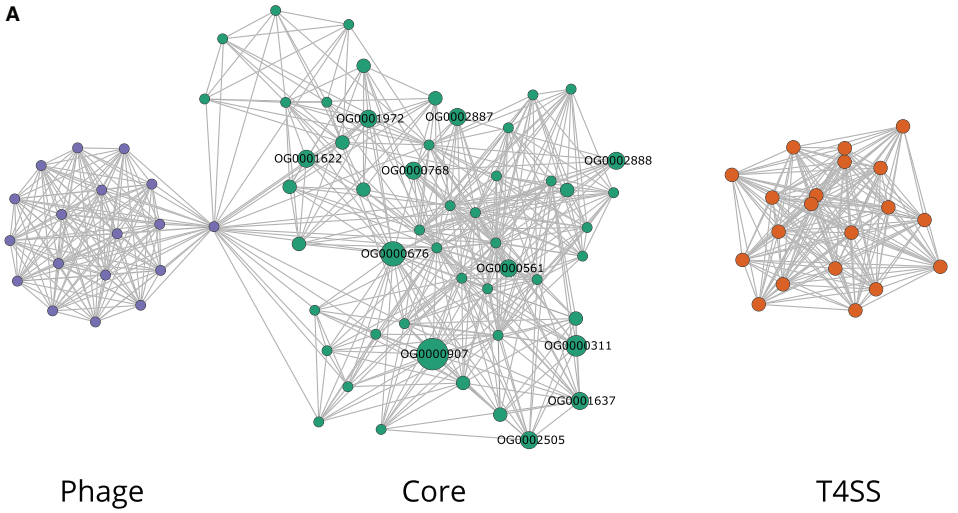
Extrachromosomal Origin of Conserved Plasmid Gene Families

We thus next asked whether gene families in the plasmid core gene set indicate a common origin of chlamydial plasmids. To address this, we analyzed the phylogeny of the most well represented gene families, *parA*/*pgp5* and *pgp7*/*8*, both of which have predicted functions typically associated with extrachromosomal elements (Figure 2B).

Homologs of *parA*/*pgp5* are found on all chlamydial plasmids and all chromosomes (Data S3). This gene family encodes ATPases with cytoskeletal properties.[62] ParA (or homologs like RepA, SopA) interacts with the DNA-binding protein ParB (RepB, SopB) and is integral for the partitioning of many low copy plasmids and phages.[62–64] The system is also often encoded chromosomally in bacteria and can contribute to chromosome partitioning.[64–66] Of note, chlamydial plasmids lack *parB* homologs, although *parB* is present on most chlamydial chromosomes.

The *parA*/*pgp5* gene family containing chlamydial plasmid and chromosomal copies is large (n = 71) and comprises five eggNOG Clusters of Orthologous Groups (COG) (Data S3). Yet, the original Chlamydiaceae *pgp5* and the highly conserved chromosomal copy of *parA* found on 38 chlamydial chromosomes all belong to a single eggNOG COG (ENOG4105C2U). Phylogenetic analysis shows that all chlamydial members of this COG are monophyletic, with plasmid Pgp5 and chromosomal ParA proteins representing sister groups (Figure 3A). This suggests that *parA*/*pgp5* was present already in the last common chlamydial ancestor, underwent gene duplication, and was subsequently maintained on some plasmids and on all closed chlamydial genomes. The closest relatives of chlamydial *parA*/*pgp5* are *parA* homologs found on plasmids of cyanobacteria and actinobacteria. This indicates that the ancestral chlamydial *parA*/*pgp5* originated from a plasmid and was subsequently integrated in chlamydial chromosomes. The presence of additional yet more distantly related plasmid-encoded *parA*/*pgp5* genes in some chlamydiae (in eggNOG ENOG4107QJE) suggests that the

Please cite this article in press as: Köstlbacher et al., Coevolving Plasmids Drive Gene Flow and Genome Plasticity in Host-Associated Intracellular Bacteria, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.10.030



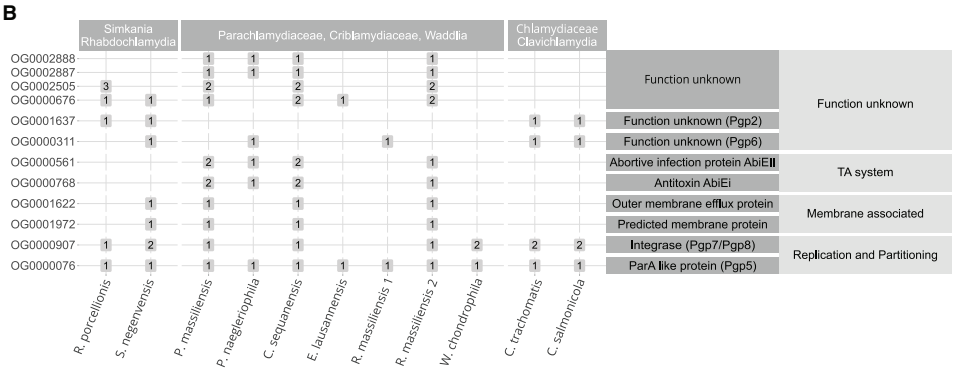


Figure 2. The Mosaic Gene Set of Chlamydial Plasmids

(A) Partial correlation network of plasmid gene families present on more than two chlamydial plasmids (n = 151, Data S4). The network represents the degree of association between gene families based on their occurrence patterns on chlamydial plasmids. Nodes represent gene families and edges represent the correlation coefficient. Only statistically significant correlations with an FDR corrected $p \leq 0.05$ are shown. Three highly connected groups of gene families can be identified, a core group (green), a type IV secretion system (T4SS) group (yellow), and a phage group (violet). Gray nodes are outliers or overlap between two clusters. Labels indicate highly conserved plasmid gene families (present on ≥ 4 plasmids).

(B) Distribution of highly conserved plasmid gene families and their predicted function. Numbers in boxes represent the gene family copy number on plasmids. See also Figures S2 and S3 and Data S2.

ancestral chlamydial *parA/pgp5* has been replaced by a homolog from an unrelated plasmid in at least one lineage, the Parachlamydiaceae (Figure 3B; Data S3). This scenario is consistent with the presence of two plasmids with *parA/pgp5* orthologs of different origin in *R. massiliensis* and earlier analysis.[46]

The second most conserved gene family on chlamydial plasmids is a putative integrase referred to here as Pgp7/8 (OG0000907, Figure 3B) due to the presence of two distinct copies on extant Chlamydiaceae plasmids. *pgp7/8* is exclusively found on chlamydial plasmids and chromosomes and is notably absent from all other known prokaryotic genomes (EggNOG ENOG4106VZX). This led us to investigate a putative viral origin by performing homology searches of Pgp7/8 proteins against the Virus Orthologous Groups database (VOGDB, http://vogdb.org/, Data S4). Hidden Markov-model-based search places Pgp7/8 into a large viral orthologous group (VOG, VOG000016) with 652 members. Phylogenetic analysis of this dataset merged with all chlamydial integrases demonstrated that chlamydial Pgp7/8 is a monophyletic clade deeply branching among viral homologs (Figure 3C). The closest relatives include the putative

Please cite this article in press as: Köstlbacher et al., Coevolving Plasmids Drive Gene Flow and Genome Plasticity in Host-Associated Intracellular Bacteria, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.10.030

Current Biology
Article

CellPress
OPEN ACCESS

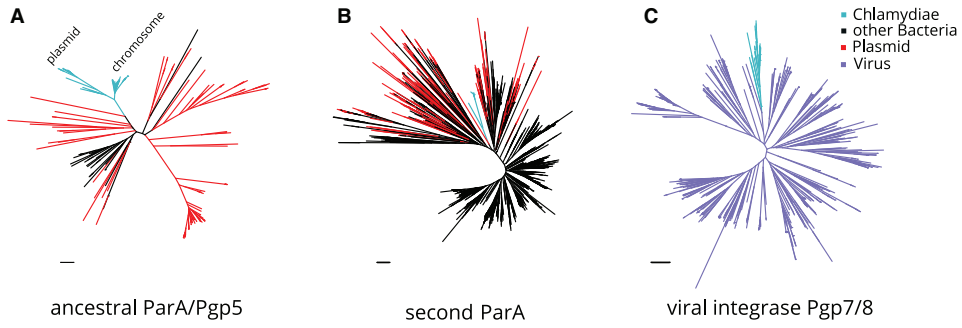


Figure 3. A Plasmid-Derived ParA/Pgp5 in the Chlamydial Ancestor and Viral Origin of Integrase Pgp7/8

(A) Phylogenetic analysis of chlamydial parA/pgp5 gene copies in EggNOG ENOG4105C2U and its plasmid representatives. Chlamydial plasmid and chromosomal clades are indicated and represent monophyletic sister groups.

(B) Phylogenetic analysis of the second chlamydial parA family in ENOG4107QJE and its plasmid representatives.

(C) Phylogenetic analysis of chlamydial pgp7/8 and its closest relatives of viral origin, the VOGDB VOG000016. Light blue indicates chlamydial branches, black other bacterial branches, red plasmid genes from the dereplicated RefSeq plasmid dataset, and purple viral genes. Maximum likelihood phylogenetic trees with best fit models (LG+C40+F, LG+C10+G+F, and LG+C60+G+F, respectively) with 1,000 ultrafast bootstraps are shown. Bootstrap support for monophyly of chlamydial clades in all trees is ≥95% and the SH-like approximate likelihood ratio is ≥80%.

Scale bars indicate one substitution per position. See also Data S3 and S4.

integrases of *Mycoplasma* phage MAV1 (NP_047270.1) and a clade of Siphoviridae that infect diverse bacteria and archaea. This suggests that *pgp7/8* was acquired once early in chlamydial evolution. Phages are known to have had a long-standing relationship with plasmids and can contribute to plasmid gene influx.[67]

Altogether, our phylogenetic analysis of the two most well-represented gene families on chlamydial plasmids suggests the presence of key plasmid genes in the last common chlamydial ancestor. The monophyly of the chlamydial partitioning protein ParA/Pgp5 indicates that this gene evolved independently on plasmids and chromosomes after an ancestral duplication event. The closest relatives are encoded on extrachromosomal genetic elements, pointing to an extrachromosomal origin of these genes.

High Frequency of Gene Flow between Plasmids and Chromosomes

A noticeable finding of our gene content analysis was that the majority of chlamydial plasmid gene families is also represented on chlamydial chromosomes (n = 255, 84.4%; Table S1). Inversely, the chromosomes of all known chlamydiae encode on average 6.4% plasmid gene families (31–204 genes, standard deviation [SD] ±1.34%; Figures 4, S4A, and S4B). This may be explained in two ways: either by integration of chromosomal genes into the plasmid or by integration of plasmid genes into the chromosome. The integration of plasmid genes into chlamydial chromosomes has been documented for a foreign *tetC* gene in the pathogen *Chlamydia suis* Tcr[68] and for the T4SS in the plasmidless amoeba symbionts *Protochlamydia amoebophila* and *Parachlamydia* acanthamoebae.[50] A high frequency of gene transfer between plasmids and chromosomes has also been observed in other bacteria[69] and has been experimentally shown in artificial soil bacterial communities.[70] This process, also referred to as gene externalization, represents an important driver of bacterial genome evolution.[71] In addition, a number of plasmid genes are apparently being maintained both on chlamydial plasmids and chromosomes in the same organism (Figure 4). Such redundancy is thought to facilitate innovation through neo-functionalization.[72] On the other hand, in small populations, as in the case of obligate endosymbionts, genetic redundancy can counteract Muller's ratchet—the fixation of slightly deleterious mutations combined with the random loss of the fittest genotypes that may lead to extinction.[73–75]

How did the high frequency of gene flow between chlamydial plasmids and chromosomes affect the functional role of both? To this end, we compared all gene families with at least one plasmid encoded copy with respect to their predicted function in cellular pathways according to eggNOG functional categories. This analysis showed that the functional profile of the plasmids is diverse but markedly differs from that of the chromosomes (Figures S4C and S4D). Chlamydial plasmid gene families for which a function could be predicted are involved in diverse cellular processes including secretion, transport, energy production/conversion, and transcription. Notably, plasmids are lacking genes functioning in translation, ribosomal structure and biogenesis, and cell motility (Figures S4C and S4D; Data S2B). The largest fraction of plasmid genes was assigned to the category "replication, recombination, and repair," which was significantly enriched in comparison to chromosomal genes (22% versus 8%; $p = 6.38 \times 10^{-16}$, one-tailed Fisher's exact test; Figure S4C). The majority of these genes represent transposases, which are considered important factors in genome evolution and may represent high turnover genes on extrachromosomal elements.[71]

Taken together, our analysis documents a high frequency of gene transfer events between chlamydial plasmids and chromosomes, possibly facilitated by transposases, which are abundantly present on most chlamydial plasmids. Despite this, chlamydial plasmids have maintained a characteristic functional profile different from chlamydial chromosomes. The high level of gene flow dynamics and the presence of characteristic

Please cite this article in press as: Köstlbacher et al., Coevolving Plasmids Drive Gene Flow and Genome Plasticity in Host-Associated Intracellular Bacteria, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.10.030

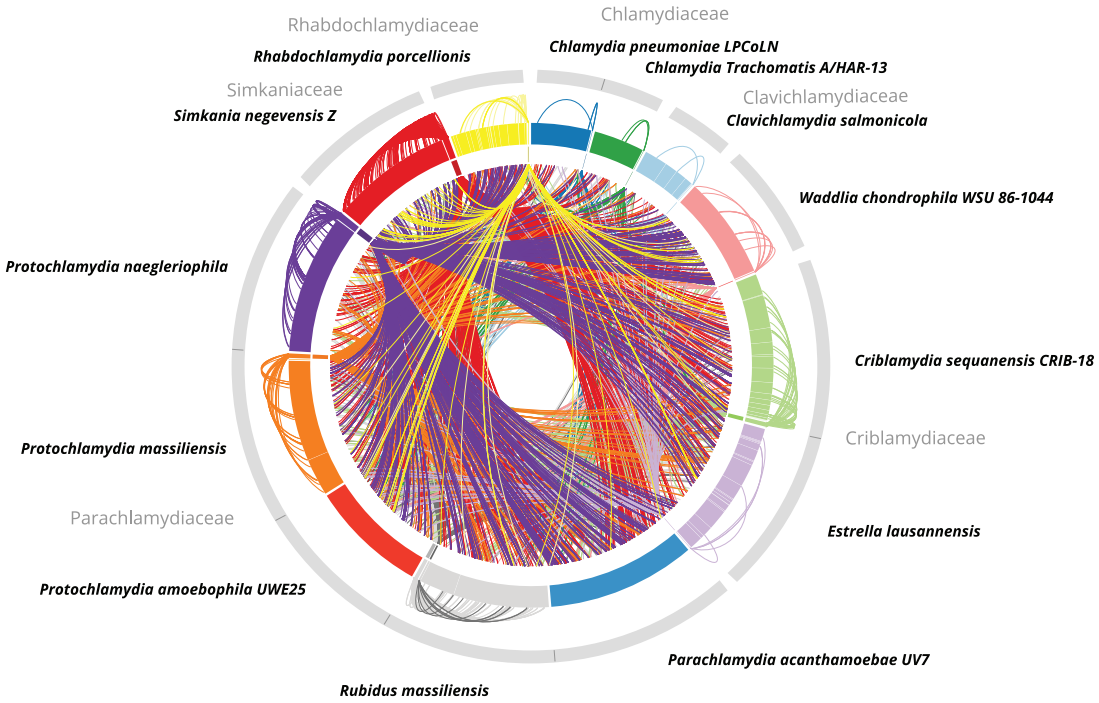


Figure 4. High Mobility of Genes between Plasmids and Host Chromosomes

The outer ring shows representations of chlamydial genome sequences including 13 chromosomes and 12 plasmids. The inner ring illustrates plasmids only. Outer links connect plasmid genes with their chromosomal homologs in the respective host chromosome. Inner links connect plasmid genes to chromosomal homologs in other chlamydial species. All chlamydial chromosomes, including those of plasmidless representatives such as *P. acanthamoebae* and *P. amoebophila*, encode a high percentage of conserved plasmid gene families (6.4% on average). See also Figure S4 and Table S1.

plasmid genes on nearly all chlamydial chromosomes further support a long-standing relationship between chlamydiae and their plasmids.

Increased Mobility and HGT among Plasmid Gene Families

We next investigated the impact of gene transfer on the chlamydial plasmid during its prolonged association with its bacterial hosts. To this end, we calculated maximum likelihood phylogenetic trees for all chlamydial gene families and applied a gene tree-species tree reconciliation approach as implemented in ecceTERA.[76] Briefly, to reduce gene tree uncertainty, ecceTERA reconciles samples of gene family trees with the species tree (Figure S1) and creates species tree aware gene trees.[77] Based on these more accurate gene trees, gene duplication, transfer, and loss events are estimated using all parsimonious reconciliations (see STAR Methods).

We first compared two sets of gene families, those that are predominantly encoded on plasmids and those predominantly encoded on chromosomes. We determined the number of gene transfers per node in a gene tree for each gene family, referred to as the number of normalized transfers per gene family. We observed a significantly increased transfer rate for plasmid-encoded gene families in comparison to chromosomal gene families (median of 0.125 versus 0.066 normalized transfers per gene family; $p = 2.9 \times 10^{-8}$, unpaired Wilcoxon signed-rank test; Figure 5A). The apparent higher mobility of plasmid-encoded genes indicates a dynamic evolutionary history and suggests that chlamydial plasmids were important mediators of HGT during the evolution of chlamydial genomes. This analysis also revealed that chlamydial genomes were differently affected by inter-species gene transfer with respect to plasmid gene families (Figure 5B). The most striking set of transfers was observed between *Parachlamydia massiliensis* and *Criblamydia sequanensis*, with 29 transfer events including *pgp7/8* (Figure S5) and *parA/pgp5*. As this constitutes more than 65% of all plasmid genes in these species, this likely indicates acquisition of a complete plasmid, as suggested above in our analysis of conserved plasmid-encoded genes (Figure 1). The direction of this inter-species plasmid transfer cannot be reliably inferred, but the better fit of the *P. massiliensis* plasmid to its host's chromosomal signature in terms of GC content and trinucleotide signature—as opposed to *C. sequanensis* and its plasmid—suggests a fairly recent transfer from *P. massiliensis* to *C. sequanensis* (Figure S2C; Data S2A).

Two other notable sets of transfer events involve the T4SS-associated genes and the putative prophage, both previously identified as major building blocks of chlamydial plasmids

Please cite this article in press as: Köstlbacher et al., Coevolving Plasmids Drive Gene Flow and Genome Plasticity in Host-Associated Intracellular Bacteria, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.10.030

Current Biology
Article

CellPress
OPEN ACCESS

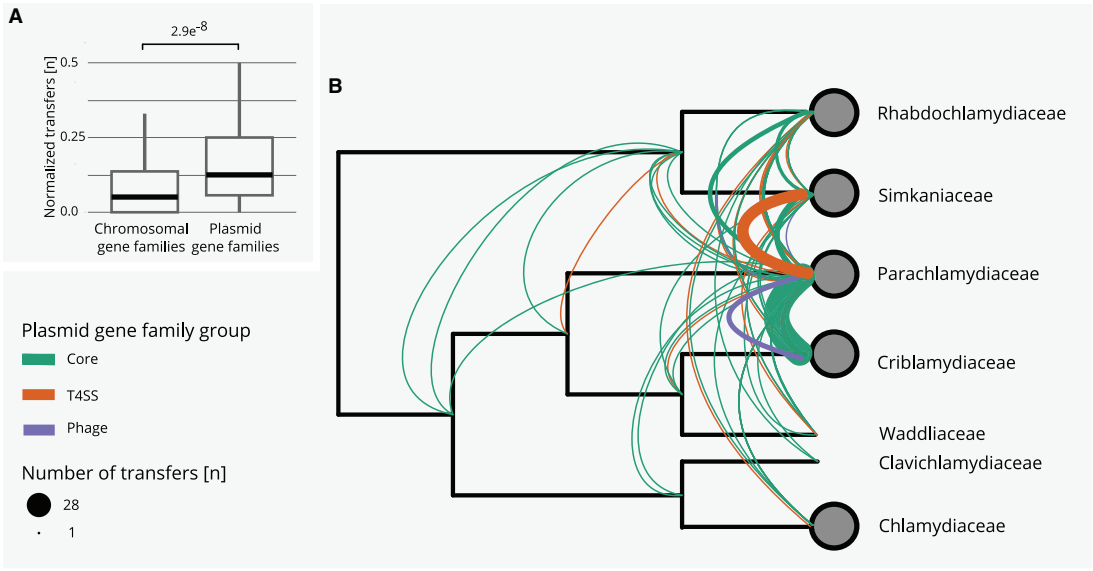


Figure 5. Increased Mobility of Plasmid Gene Families and Inter-family Transfer Events of Plasmid Genes

(A) Boxplot showing the number of normalized transfer events per gene family as inferred from gene tree-species tree reconciliations using 2,950 chromosomal and 141 plasmid gene families. The p value was calculated using the Wilcoxon signed-rank test. Outliers are not shown but are included in the statistical analysis. (B) Transfer events of plasmid genes superimposed on a schematic chlamydial species tree collapsed at the family level. The transfer of T4SS-associated genes between the Simkaniaceae and the Parachlamydiaceae is indicated in orange. Core plasmid gene transfers between multiple families and a potential whole plasmid transfer from *P. massiliensis* to *C. sequanensis* are shown in green. The inferred transfer of a prophage is indicated in purple. See also Figure S5.

(Figure 3). Gene tree-species tree reconciliation indicates that these gene sets were transferred between the LCAs of the Simkaniaceae and Parachlamydiaceae, and between the Parachlamydiaceae and Criblamydiaceae (Figure 5B).

Collectively, gene tree-species tree reconciliations revealed chlamydial plasmids as important facilitators of HGT. Plasmid-encoded gene families are more frequently transferred than chromosomal gene families, and there is evidence for interspecies transmission of complete plasmids and large functional units, such as the chlamydial T4SS. HGT is a major driver of microbial genome evolution, promoting the adaptation to novel environmental conditions.[78] It is considered particularly important for strictly intracellular bacteria as it provides another means to escape Muller's ratchet.[73,74]

A Scenario for Evolutionary Trajectories of Chlamydial Plasmids

Combining our comprehensive phylogenetic analysis and evidence from gene-tree species-tree reconciliation results in an evolutionary scenario for a common origin of extant chlamydial plasmids and a shared evolutionary history with their bacterial hosts. We base this scenario on the findings of (1) the acquisition of the host chromosome trinucleotide signature of chlamydial plasmids, (2) the presence of a set of co-occurring core chlamydial plasmid genes, (3) the monophyly of the key chlamydial plasmid genes *pgp5/parA* and *pgp7/8* and their inferred extrachromosomal origin, (4) the high prevalence of chlamydial plasmid genes on chromosomes, and (5) the predominantly vertical inheritance of pgp7/8. We derived the gene content of putative ancestral plasmids using the gene tree-species tree reconciliations of plasmid enriched gene families.

The reconstructed ancestral plasmid last common ancestor (plasmid LCA or pLCA) present in the LCA of all chlamydiae contained 11 plasmid gene families (Figure 6; Table S2), including *parA/pgp5*, the helicase *pgp1*, and *pgp6*, the two latter of which are essential for the maintenance of extant Chlamydiaceae plasmids.[57] Molecular dating of the chlamydiae LCA estimated an age of 700 My to one billion years,[23,24] which likely places the chlamydiae pLCA at approximately the same time.

Next, the pLCA of the Parachlamydiales-Chlamydiales ancestor presumably acquired an integrase from a phage donor related to the Siphoviridae, which subsequently underwent gene duplication (Figure 3). Most chlamydial plasmids retained only one copy, while both genes diverged to give rise to *pgp7* and *pgp8* in current Chlamydiaceae and Clavichlamydia plasmids (Figure 6; Figure S5). Consistent with this, the almost entirely vertical transmission of this gene family has been observed earlier for *C. trachomatis* strains[42] and the genus *Chlamydia* in general.[40]

A decisive event occurred during the divergence of the ancestor of the Parachlamydiaceae, Criblamydiaceae, and Waddliaceae, and the ancestor of the Chlamydiaceae and Clavichlamydiaceae (Figure 1). The ancestral plasmid of the latter gained *pgp4*, which today is a key plasmid specific transcription factor of virulence genes for *in vivo* pathogenicity in the Chlamydiaceae.[79] This event likely contributed to niche differentiation

CURBIO 16985

Please cite this article in press as: Köstlbacher et al., Coevolving Plasmids Drive Gene Flow and Genome Plasticity in Host-Associated Intracellular Bacteria, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.10.030

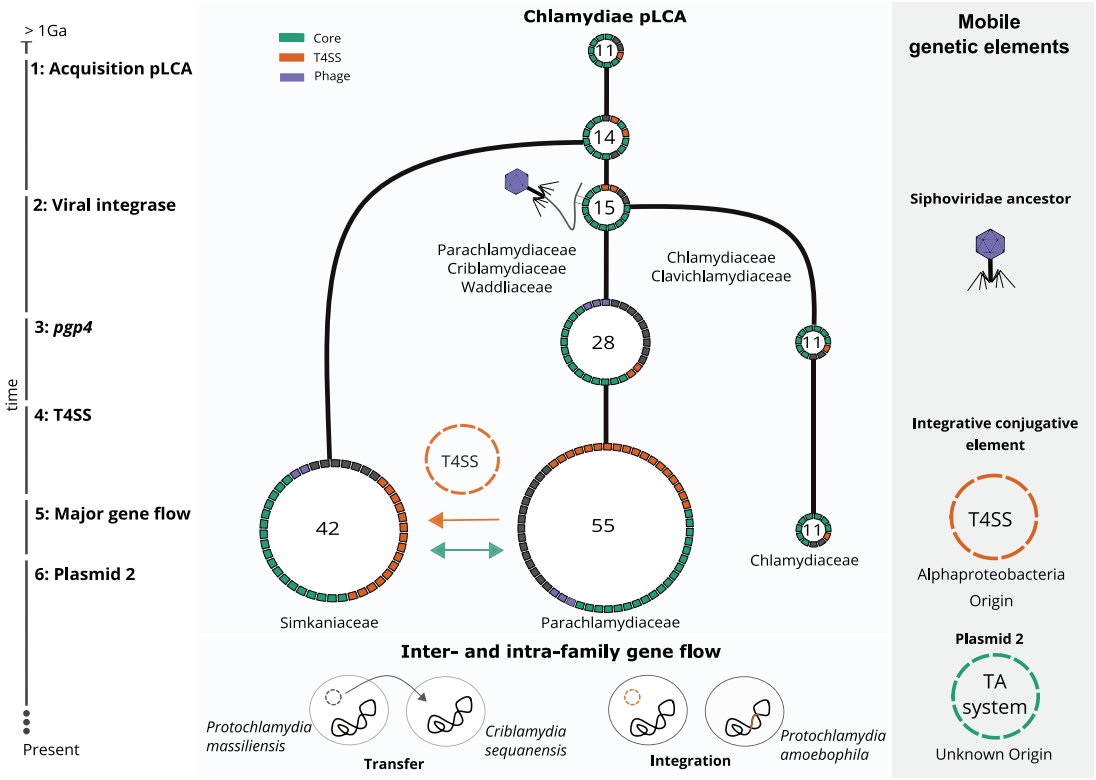


Figure 6. A Scenario for the Evolutionary History of Chlamydial Plasmids

Reconstructed ancestral plasmids (pLCAs; middle panel) are shown as rings along a schematic timeline of evolutionary events over an estimated period of 1 billion years (left). Ring segments indicate plasmid-encoded genes colored by functional groups (green, chlamydial core plasmid; yellow, T4SS genes; purple, phage genes). The numbers in the rings refer to the number of gene families present on the ancestral plasmids. Major events include 1: acquisition of the original Chlamydiae pLCA by the last common chlamydial ancestor from an unknown donor; 2: acquisition of the viral integrase pgp7/8; 3: acquisition of the transcriptional regulator pgp4 in the Chlamydiaceae/Clavichlamydiaceae pLCA; 4: acquisition of the T4SS by the Parachlamydiaceae ancestor from an Alphaproteobacteria ancestor; 5: transfer of the T4SS and pgp7/8 from the Parachlamydiaceae pLCA to the Simkaniaceae pLCA; 6: acquisition of a second plasmid in the Parachlamydiaceae LCA that encodes a TA system; 7: inter- and intra-family plasmid gene flow, such as plasmid transfer from P. *massiliensis* to *C. sequanensis* or plasmid integration in *P. amoebophila*. See also Figure S6 and Table S2.

and the infection of higher animals including humans, as loss of the plasmid has, in some *Chlamydia* species, been shown to lead to attenuated infection.[80,81] At this point, the plasmid already included seven of the eight plasmid gene families encoded in the extant Chlamydiaceae plasmid (Figure 6).

In the Parachlamydiaceae/Criblamydiaceae/Waddliaceae lineage, which includes a large number of diverse species that live as symbionts of amoeba in the environment,[82,83] the ancestral plasmid underwent major expansions through several independent gene acquisitions and almost doubled in gene content (from 28 to 55 gene families). A T4SS was acquired from an Alphaproteobacteria donor[50] and integrated into the plasmid (Figure 6; Figure S4). Intriguingly, the T4SS does not appear to originate from a conjugative plasmid but is likely an ICE[84] as the closest relatives are extant *Rickettsia* ICEs.[85] In close temporal proximity, another plasmid entered the Parachlamydiaceae ancestor, bringing a set of Parachlamydiaceae plasmid specific genes, including a TA system (Figure S6). Together this gene set forms the backbone for extant plasmids in members of the Parachlamydiaceae. The Parachlamydiaceae T4SS was subsequently acquired together with a number of accessory genes by the plasmid in the Simkaniaceae ancestor (Figures 5 and 6) and (partially) integrated in the chromosome in some Parachlamydiaceae members. Throughout this series of evolutionary events and during the long coevolution of chlamydiae with their plasmids, chromosomal integration of plasmid genes and mobilization of chromosomal genes contributed to shaping the chlamydial genome (Figure 5).

In summary, plasmids are well known for their contribution to the adaptation and evolution of microbes. Yet, coevolution of plasmids with their hosts has mostly been studied using experimental evolution approaches[14–17] or evolutionary genomics for closely related microorganisms.[86–88] Plasmids depend on host resources for maintenance and evolve toward a reduction of

Please cite this article in press as: Köstlbacher et al., Coevolving Plasmids Drive Gene Flow and Genome Plasticity in Host-Associated Intracellular Bacteria, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.10.030

metabolic costs and/or an increased persistence.[12,89,90] Additionally, adaptation on the host side can, given selective pressure for a period of time or mitigating environmental conditions, reduce the cost of plasmid carriage.[10,16,17,91,92] Here, we provided evidence that, in the phylum Chlamydiae, this has led to an unmatched intimate evolutionary relationship, in which an ancient acquisition of an ancestral plasmid and subsequent gene gains and losses gave rise to a collection of extant plasmids in a highly diverse range of bacterial hosts. These plasmids are crucial for the virulence of modern human and animal pathogens[79,93–95] and widespread among their environmental representatives. Chlamydial plasmids have promoted inter-species gene transfer, which in concert with the ancient and strictly intracellular lifestyle of chlamydiae has likely contributed to the maintenance and persistence of the plasmid over extended evolutionary time periods.[96] Plasmids may have provided a means for this group of strictly intracellular microbes to ameliorate the degenerative effects of Muller's ratchet by promoting HGT.[97] To the best of our knowledge, we documented the presumably oldest known system of host-plasmid coexistence and coevolution, with a shared history of around one billion years.[23,24]

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead Contact
  - Material Availability
  - Data and Code Availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Comparison of trinucleotide signatures of plasmids and chromosomes
  - Generation of a dereplicated plasmid dataset
  - Mapping to clusters of orthologous groups (COGs)
  - Mapping to viral orthology database
  - Identification of gene families by *de novo* clustering of orthologous groups (OGs)
  - Partial correlation network analysis
  - Phylogenetic analysis of COG and VOG-based datasets
  - Species tree reconstruction
  - Gene tree-species tree reconciliation
  - Reconstruction of ancestral chlamydial plasmids and estimation of gene transfer frequencies
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cub.2020.10.030.

ACKNOWLEDGMENTS

We want to thank Masaki Shintani for advice and discussions concerning plasmid biology and Craig Herbold for advice and discussions about gene tree-species tree reconstructions. The Life Science Compute Cluster (LiSC; http://cube.univie.ac.at/lisc) was used for computational analysis. This project was supported by the European Research Council ERC (EVOCHLAMY, grant no. 281633 to M.H.), the Austrian Science Fund FWF (projects DOC 69-B to M.H. and P 32112 to A.C.), and the University of Vienna (uni:docs fellowship to T.H.).

AUTHOR CONTRIBUTIONS

S.K., A.C., D.D., and M.H. conceptualized the study. S.K. and A.C. performed comparative genomic analysis. S.K. performed phylogenetic analyses and gene tree-species tree reconciliation analyses. S.K., A.C., T.H., and M.H. interpreted the results. All authors wrote and edited the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 21, 2020
Revised: September 30, 2020
Accepted: October 9, 2020
Published: November 5, 2020

REFERENCES

1. Summers, D.K. (1996). The Biology of Plasmids (Blackwell Science Ltd).
2. Shintani, M., Sanchez, Z.K., and Kimbara, K. (2015). Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. Front. Microbiol. *6*, 242.
3. Smillie, C., Garcillán-Barcia, M.P., Francia, M.V., Rocha, E.P.C., and de la Cruz, F. (2010). Mobility of plasmids. Microbiol. Mol. Biol. Rev. *74*, 434–452.
4. Johnson, T.J., and Nolan, L.K. (2009). Pathogenomics of the virulence plasmids of Escherichia coli. Microbiol. Mol. Biol. Rev. *73*, 750–774.
5. San Millan, A. (2018). Evolution of Plasmid-Mediated Antibiotic Resistance in the Clinical Context. Trends Microbiol. *26*, 978–985.
6. Harms, A., Brodersen, D.E., Mitarai, N., and Gerdes, K. (2018). Toxins, Targets, and Triggers: An Overview of Toxin-Antitoxin Biology. Mol. Cell *70*, 768–784.
7. Diaz Ricci, J.C., and Hernández, M.E. (2000). Plasmid effects on Escherichia coli metabolism. Crit. Rev. Biotechnol. *20*, 79–108.
8. Rozkov, A., Avignone-Rossa, C.A., Ertl, P.F., Jones, P., O'Kennedy, R.D., Smith, J.J., Dale, J.W., and Bushell, M.E. (2004). Characterization of the metabolic burden on Escherichia coli DH1 cells imposed by the presence of a plasmid containing a gene therapy sequence. Biotechnol. Bioeng. *88*, 909–915.
9. Bergstrom, C.T., Lipsitch, M., and Levin, B.R. (2000). Natural selection, infectious transfer and the existence conditions for bacterial plasmids. Genetics *155*, 1505–1519.
10. San Millan, A., Peña-Miller, R., Toll-Riera, M., Halbert, Z.V., McLean, A.R., Cooper, B.S., and MacLean, R.C. (2014). Positive selection and compensatory adaptation interact to stabilize non-transmissible plasmids. Nat. Commun. *5*, 5208.
11. Yano, H., Wegrzyn, K., Loftie-Eaton, W., Johnson, J., Deckert, G.E., Rogers, L.M., Konieczny, I., and Top, E.M. (2016). Evolved plasmid-host interactions reduce plasmid interference cost. Mol. Microbiol. *101*, 743–756.
12. Porse, A., Schønning, K., Munck, C., and Sommer, M.O.A. (2016). Survival and Evolution of a Large Multidrug Resistance Plasmid in New Clinical Bacterial Hosts. Mol. Biol. Evol. *33*, 2860–2873.
13. Krupovic, M., Makarova, K.S., Wolf, Y.I., Medvedeva, S., Prangishvili, D., Forterre, P., and Koonin, E.V. (2019). Integrated mobile genetic elements in Thaumarchaeota. Environ. Microbiol. *21*, 2056–2078.
14. Bottery, M.J., Wood, A.J., and Brockhurst, M.A. (2017). Adaptive modulation of antibiotic resistance through intragenomic coevolution. Nat. Ecol. Evol. *1*, 1364–1369.

Please cite this article in press as: Köstlbacher et al., Coevolving Plasmids Drive Gene Flow and Genome Plasticity in Host-Associated Intracellular Bacteria, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.10.030

15. Bottery, M.J., Wood, A.J., and Brockhurst, M.A. (2019). Temporal dynamics of bacteria-plasmid coevolution under antibiotic selection. ISME J. *13*, 559–562.
16. Jordt, H., Stalder, T., Kosterlitz, O., Ponciano, J.M., Top, E.M., and Kerr, B. (2020). Coevolution of host-plasmid pairs facilitates the emergence of novel multidrug resistance. Nat. Ecol. Evol. *4*, 863–869.
17. Stalder, T., Rogers, L.M., Renfrow, C., Yano, H., Smith, Z., and Top, E.M. (2017). Emerging patterns of plasmid-host coevolution that stabilize antibiotic resistance. Sci. Rep. *7*, 4853.
18. Harrison, E., and Brockhurst, M.A. (2012). Plasmid-mediated horizontal gene transfer is a coevolutionary process. Trends Microbiol. *20*, 262–267.
19. Hülter, N., Ilhan, J., Wein, T., Kadibalban, A.S., Hammerschmidt, K., and Dagan, T. (2017). An evolutionary perspective on plasmid lifestyle modes. Curr. Opin. Microbiol. *38*, 74–80.
20. Frost, L.S., Leplae, R., Summers, A.O., and Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. Nat. Rev. Microbiol. *3*, 722–732.
21. Wernegreen, J.J., and Moran, N.A. (2001). Vertical transmission of biosynthetic plasmids in aphid endosymbionts (Buchnera). J. Bacteriol. *183*, 785–790.
22. Boyd, B.M., Allen, J.M., Nguyen, N.-P., Vachaspati, P., Quicksall, Z.S., Warnow, T., Mugisha, L., Johnson, K.P., and Reed, D.L. (2017). Primates, Lice and Bacteria: Speciation and Genome Evolution in the Symbionts of Hominid Lice. Mol. Biol. Evol. *34*, 1743–1757.
23. Horn, M., Collingro, A., Schmitz-Esser, S., Beier, C.L., Purkhold, U., Fartmann, B., Brandt, P., Nyakatura, G.J., Droege, M., Frishman, D., et al. (2004). Illuminating the evolutionary history of chlamydiae. Science *304*, 728–730.
24. Kamneva, O.K., Knight, S.J., Liberles, D.A., and Ward, N.L. (2012). Analysis of genome content evolution in pvc bacterial super-phylum: assessment of candidate genes associated with cellular organization and lifestyle. Genome Biol. Evol. *4*, 1375–1390.
25. Greub, G., and Raoult, D. (2003). History of the ADP/ATP-translocase-encoding gene, a parasitism gene transferred from a Chlamydiales ancestor to plants 1 billion years ago. Appl. Environ. Microbiol. *69*, 5530–5535.
26. McCutcheon, J.P., and Moran, N.A. (2011). Extreme genome reduction in symbiotic bacteria. Nat. Rev. Microbiol. *10*, 13–26.
27. Sabater-Muñoz, B., Toft, C., Alvarez-Ponce, D., and Fares, M.A. (2017). Chance and necessity in the genome evolution of endosymbiotic bacteria of insects. ISME J. *11*, 1291–1304.
28. Andersson, S.G.E., Alsmark, C., Canbäck, B., Davids, W., Frank, C., Karlberg, O., Klasson, L., Antoine-Legault, B., Mira, A., and Tamas, I. (2002). Comparative genomics of microbial pathogens and symbionts. Bioinformatics *18* (*Suppl 2*), S17.
29. Bordenstein, S.R., and Reznikoff, W.S. (2005). Mobile DNA in obligate intracellular bacteria. Nat. Rev. Microbiol. *3*, 688–699.
30. Thomas, N.S., Lusher, M., Storey, C.C., and Clarke, I.N. (1997). Plasmid diversity in Chlamydia. Microbiology (Reading) *143*, 1847–1854.
31. Pearce, B.J., Fahr, M.J., Hatch, T.P., and Sriprakash, K.S. (1991). A chlamydial plasmid is differentially transcribed during the life cycle of Chlamydia trachomatis. Plasmid *26*, 116–122.
32. Jones, C.A., Hadfield, J., Thomson, N.R., Cleary, D.W., Marsh, P., Clarke, I.N., and O'Neill, C.E. (2020). The Nature and Extent of Plasmid Variation in *Chlamydia trachomatis*. Microorganisms *8*, 373.
33. Shima, K., Wanker, M., Skilton, R.J., Cutcliffe, L.T., Schnee, C., Kohl, T.A., Niemann, S., Geijo, J., Klinger, M., Timms, P., et al. (2018). The Genetic Transformation of Chlamydia pneumoniae. MSphere *3*, e00412-18.
34. Pickett, M.A., Everson, J.S., Pead, P.J., and Clarke, I.N. (2005). The plasmids of Chlamydia trachomatis and Chlamydophila pneumoniae (N16): accurate determination of copy number and the paradoxical effect of plasmid-curing agents. Microbiology (Reading) *151*, 893–903.
35. O'Connell, C.M., and Nicks, K.M. (2006). A plasmid-cured Chlamydia muridarum strain displays altered plaque morphology and reduced infectivity in cell culture. Microbiology (Reading) *152*, 1601–1607.
36. Patton, M.J., Chen, C.-Y., Yang, C., McCorrister, S., Grant, C., Westmacott, G., Yuan, X.-Y., Ochoa, E., Fariss, R., Whitmire, W.M., et al. (2018). Plasmid Negative Regulation of CPAF Expression Is Pgp4 Independent and Restricted to Invasive *Chlamydia trachomatis* Biovars. MBio *9*, e02164–e17.
37. Russell, M., Darville, T., Chandra-Kuntal, K., Smith, B., Andrews, C.W., Jr., and O'Connell, C.M. (2011). Infectivity acts as in vivo selection for maintenance of the chlamydial cryptic plasmid. Infect. Immun. *79*, 98–107.
38. Carlson, J.H., Whitmire, W.M., Crane, D.D., Wicke, L., Virtaneva, K., Sturdevant, D.E., Kupko, J.J., 3rd, Porcella, S.F., Martinez-Orengo, N., Heinzen, R.A., et al. (2008). The Chlamydia trachomatis plasmid is a transcriptional regulator of chromosomal genes and a virulence factor. Infect. Immun. *76*, 2273–2283.
39. Seth-Smith, H.M.B., Harris, S.R., Persson, K., Marsh, P., Barron, A., Bignell, A., Bjartling, C., Clark, L., Cutcliffe, L.T., Lambden, P.R., et al. (2009). Co-evolution of genomes and plasmids within Chlamydia trachomatis and the emergence in Sweden of a new variant strain. BMC Genomics *10*, 239.
40. Szabo, K.V., O'Neill, C.E., and Clarke, I.N. (2020). Diversity in Chlamydial plasmids. PLoS ONE *15*, e0233298.
41. Versteeg, B., Bruisten, S.M., Pannekoek, Y., Jolley, K.A., Maiden, M.C.J., van der Ende, A., and Harrison, O.B. (2018). Genomic analyses of the Chlamydia trachomatis core genome show an association between chromosomal genome, plasmid type and disease. BMC Genomics *19*, 130.
42. Hadfield, J., Harris, S.R., Seth-Smith, H.M.B., Parmar, S., Andersson, P., Giffard, P.M., Schachter, J., Moncada, J., Ellison, L., Vaulet, M.L.G., et al. (2017). Comprehensive global genome dynamics of *Chlamydia trachomatis* show ancient diversification followed by contemporary mixing and recent lineage expansion. Genome Res. *27*, 1220–1229.
43. Demars, R., Weinfurter, J., Guex, E., Lin, J., and Potucek, Y. (2007). Lateral gene transfer in vitro in the intracellular pathogen Chlamydia trachomatis. J. Bacteriol. *189*, 991–1003.
44. DeMars, R., and Weinfurter, J. (2008). Interstrain gene transfer in Chlamydia trachomatis in vitro: mechanism and significance. J. Bacteriol. *190*, 1605–1614.
45. Suchland, R.J., Carrell, S.J., Wang, Y., Hybiske, K., Kim, D.B., Dimond, Z.E., Hefty, P.S., and Rockey, D.D. (2019). Chromosomal Recombination Targets in *Chlamydia* Interspecies Lateral Gene Transfer. J. Bacteriol. *201*, e00365-19.
46. Bertelli, C., Cissé, O.H., Rusconi, B., Kebbi-Beghdadi, C., Croxatto, A., Goesmann, A., Collyn, F., and Greub, G. (2016). CRISPR System Acquisition and Evolution of an Obligate Intracellular Chlamydia-Related Bacterium. Genome Biol. Evol. *8*, 2376–2386.
47. Bou Khalil, J.Y., Benamar, S., Baudoin, J.-P., Croce, O., Blanc-Tailleur, C., Pagnier, I., Raoult, D., and La Scola, B. (2016). Developmental Cycle and Genome Analysis of "Rubidus massiliensis," a New Vermamoeba vermiformis Pathogen. Front. Cell. Infect. Microbiol. *6*, 31.
48. Benamar, S., Bou Khalil, J.Y., Blanc-Tailleur, C., Bilen, M., Barrassi, L., and La Scola, B. (2017). Developmental Cycle and Genome Analysis of *Protochlamydia massiliensis* sp. nov. a New Species in the *Parachlamydiacae* Family. Front. Cell. Infect. Microbiol. *7*, 385.
49. Bertelli, C., Goesmann, A., and Greub, G. (2014). Criblamydia sequanensis Harbors a Megaplasmid Encoding Arsenite Resistance. Genome Announc. *2*, e00949–e14.
50. Collingro, A., Tischler, P., Weinmaier, T., Penz, T., Heinz, E., Brunham, R.C., Read, T.D., Bavoil, P.M., Sachse, K., Kahane, S., et al. (2011). Unity in variety—the pan-genome of the Chlamydiae. Mol. Biol. Evol. *28*, 3253–3270.
51. Bertelli, C., Aeby, S., Chassot, B., Clulow, J., Hilfiker, O., Rappo, S., Ritzmann, S., Schumacher, P., Terrettaz, C., Benaglio, P., et al. (2015).

Please cite this article in press as: Köstlbacher et al., Coevolving Plasmids Drive Gene Flow and Genome Plasticity in Host-Associated Intracellular Bacteria, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.10.030

Sequencing and characterizing the genome of Estrella lausannensis as an undergraduate project: training students and biological insights. Front. Microbiol. *6*, 101.

52. Bertelli, C., Collyn, F., Croxatto, A., Rückert, C., Polkinghorne, A., Kebbi-Beghdadi, C., Goesmann, A., Vaughan, L., and Greub, G. (2010). The Waddlia genome: a window into chlamydial biology. PLoS ONE *5*, e10890.
53. Dharamshi, J.E., Tamarit, D., Eme, L., Stairs, C.W., Martijn, J., Homa, F., Jørgensen, S.L., Spang, A., and Ettema, T.J.G. (2020). Marine Sediments Illuminate Chlamydiae Diversity and Evolution. Curr. Biol. *30*, 1032–1048.
54. Rocha, E.P.C., and Danchin, A. (2002). Base composition bias might result from competition for metabolic resources. Trends Genet. *18*, 291–294.
55. Nishida, H. (2012). Comparative analyses of base compositions, DNA sizes, and dinucleotide frequency profiles in archaeal and bacterial chromosomes and plasmids. Int. J. Evol. Biol. *2012*, 342482.
56. Suzuki, H., Yano, H., Brown, C.J., and Top, E.M. (2010). Predicting plasmid promiscuity based on genomic signature. J. Bacteriol. *192*, 6045–6055.
57. Gong, S., Yang, Z., Lei, L., Shen, L., and Zhong, G. (2013). Characterization of Chlamydia trachomatis plasmid-encoded open reading frames. J. Bacteriol. *195*, 3819–3826.
58. Zhong, G. (2017). Chlamydial Plasmid-Dependent Pathogenicity. Trends Microbiol. *25*, 141–152.
59. Liu, Y., Huang, Y., Yang, Z., Sun, Y., Gong, S., Hou, S., Chen, C., Li, Z., Liu, Q., Wu, Y., et al. (2014). Plasmid-encoded Pgp3 is a major virulence factor for Chlamydia muridarum to induce hydrosalpinx in mice. Infect. Immun. *82*, 5327–5335.
60. Albrecht, M., Sharma, C.M., Reinhardt, R., Vogel, J., and Rudel, T. (2010). Deep sequencing-based discovery of the Chlamydia trachomatis transcriptome. Nucleic Acids Res. *38*, 868–877.
61. Ferreira, R., Borges, V., Nunes, A., Borrego, M.J., and Gomes, J.P. (2013). Assessment of the load and transcriptional dynamics of Chlamydia trachomatis plasmid according to strains' tissue tropism. Microbiol. Res. *168*, 333–339.
62. Ringgaard, S., van Zon, J., Howard, M., and Gerdes, K. (2009). Movement and equipositioning of plasmids by ParA filament disassembly. Proc. Natl. Acad. Sci. USA *106*, 19369–19374.
63. Motallebi-Veshareh, M., Rouch, D.A., and Thomas, C.M. (1990). A family of ATPases involved in active partitioning of diverse bacterial plasmids. Mol. Microbiol. *4*, 1455–1463.
64. Bignell, C., and Thomas, C.M. (2001). The bacterial ParA-ParB partitioning proteins. J. Biotechnol. *91*, 1–34.
65. Quisel, J.D., and Grossman, A.D. (2000). Control of sporulation gene expression in Bacillus subtilis by the chromosome partitioning proteins Soj (ParA) and Spo0J (ParB). J. Bacteriol. *182*, 3446–3451.
66. Lee, P.S., and Grossman, A.D. (2006). The chromosome partitioning proteins Soj (ParA) and Spo0J (ParB) contribute to accurate chromosome partitioning, separation of replicated sister origins, and regulation of replication initiation in Bacillus subtilis. Mol. Microbiol. *60*, 853–869.
67. Roux, S., Hallam, S.J., Woyke, T., and Sullivan, M.B. (2015). Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. eLife *4*, e08490.
68. Dugan, J., Rockey, D.D., Jones, L., and Andersen, A.A. (2004). Tetracycline resistance in Chlamydia suis mediated by genomic islands inserted into the chlamydial inv-like gene. Antimicrob. Agents Chemother. *48*, 3989–3995.
69. Zheng, J., Guan, Z., Cao, S., Peng, D., Ruan, L., Jiang, D., and Sun, M. (2015). Plasmids are vectors for redundant chromosomal genes in the Bacillus cereus group. BMC Genomics *16*, 6.
70. Hall, J.P.J., Williams, D., Paterson, S., Harrison, E., and Brockhurst, M.A. (2017). Positive selection inhibits gene mobilisation and transfer in soil bacterial communities. Nat. Ecol. Evol. *1*, 1348–1353.
71. Corel, E., Méheust, R., Watson, A.K., McInerney, J.O., Lopez, P., and Bapteste, E. (2018). Bipartite Network Analysis of Gene Sharings in the Microbial World. Mol. Biol. Evol. *35*, 899–913.
72. Taylor, J.S., and Raes, J. (2004). Duplication and divergence: the evolution of new genes and old ideas. Annu. Rev. Genet. *38*, 615–643.
73. Takeuchi, N., Kaneko, K., and Koonin, E.V. (2014). Horizontal gene transfer can rescue prokaryotes from Muller's ratchet: benefit of DNA from dead cells and population subdivision. G3 (Bethesda) *4*, 325–339.
74. Naito, M., and Pawlowska, T.E. (2016). Defying Muller's Ratchet: Ancient Heritable Endobacteria Escape Extinction through Retention of Recombination and Genome Plasticity. MBio *7*, e02057–e15.
75. Maciver, S.K. (2016). Asexual Amoebae Escape Muller's Ratchet through Polyploidy. Trends Parasitol. *32*, 855–862.
76. Jacox, E., Chauve, C., Szöllősi, G.J., Ponty, Y., and Scornavacca, C. (2016). ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony. Bioinformatics *32*, 2056–2058.
77. Scornavacca, C., Jacox, E., and Szöllősi, G.J. (2015). Joint amalgamation of most parsimonious reconciled gene trees. Bioinformatics *31*, 841–848.
78. Treangen, T.J., and Rocha, E.P.C. (2011). Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet. *7*, e1001284.
79. Song, L., Carlson, J.H., Whitmire, W.M., Kari, L., Virtaneva, K., Sturdevant, D.E., Watkins, H., Zhou, B., Sturdevant, G.L., Porcella, S.F., et al. (2013). Chlamydia trachomatis plasmid-encoded Pgp4 is a transcriptional regulator of virulence-associated genes. Infect. Immun. *81*, 636–644.
80. Kari, L., Whitmire, W.M., Olivares-Zavaleta, N., Goheen, M.M., Taylor, L.D., Carlson, J.H., Sturdevant, G.L., Lu, C., Bakios, L.E., Randall, L.B., et al. (2011). A live-attenuated chlamydial vaccine protects against trachoma in nonhuman primates. J. Exp. Med. *208*, 2217–2223.
81. O'Connell, C.M., Ingalls, R.R., Andrews, C.W., Jr., Scurlock, A.M., and Darville, T. (2007). Plasmid-deficient Chlamydia muridarum fail to induce immune pathology and protect against oviduct disease. J. Immunol. *179*, 4027–4034.
82. Horn, M. (2008). Chlamydiae as symbionts in eukaryotes. Annu. Rev. Microbiol. *62*, 113–131.
83. Collingro, A., Köstlbacher, S., and Horn, M. (2020). Chlamydiae in the Environment. Trends Microbiol. *28*, 877–888.
84. Guglielmini, J., Quintais, L., Garcillán-Barcia, M.P., de la Cruz, F., and Rocha, E.P.C. (2011). The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. PLoS Genet. *7*, e1002222.
85. Nakayama, K., Yamashita, A., Kurokawa, K., Morimoto, T., Ogawa, M., Fukuhara, M., Urakami, H., Ohnishi, M., Uchiyama, I., Ogura, Y., et al. (2008). The Whole-genome sequencing of the obligate intracellular bacterium Orientia tsutsugamushi revealed massive gene amplification during reductive genome evolution. DNA Res. *15*, 185–199.
86. Zheng, J., Peng, D., Ruan, L., and Sun, M. (2013). Evolution and dynamics of megaplasmids with genome sizes larger than 100 kb in the Bacillus cereus group. BMC Evol. Biol. *13*, 262.
87. Gillespie, J.J., Beier, M.S., Rahman, M.S., Ammerman, N.C., Shallom, J.M., Purkayastha, A., Sobral, B.S., and Azad, A.F. (2007). Plasmids and rickettsial evolution: insight from Rickettsia felis. PLoS ONE *2*, e266.
88. Gil, R., Sabater-Muñoz, B., Perez-Brocal, V., Silva, F.J., and Latorre, A. (2006). Plasmids in the aphid endosymbiont Buchnera aphidicola with the smallest genomes. A puzzling evolutionary story. Gene *370*, 17–25.
89. Levin, B.R. (1993). The accessory genetic elements of bacteria: existence conditions and (co)evolution. Curr. Opin. Genet. Dev. *3*, 849–854.
90. Dietel, A.-K., Kaltenpoth, M., and Kost, C. (2018). Convergent Evolution in Intracellular Elements: Plasmids as Model Endosymbionts. Trends Microbiol. *26*, 755–768.
91. Bouma, J.E., and Lenski, R.E. (1988). Evolution of a bacteria/plasmid association. Nature *335*, 351–352.

Please cite this article in press as: Köstlbacher et al., Coevolving Plasmids Drive Gene Flow and Genome Plasticity in Host-Associated Intracellular Bacteria, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.10.030

92. Wein, T., Hülter, N.F., Mizrahi, I., and Dagan, T. (2019). Emergence of plasmid stability under non-selective conditions maintains antibiotic resistance. Nat. Commun. *10*, 2595.

93. Skilton, R.J., Wang, Y., O'Neill, C., Filardo, S., Marsh, P., Bénard, A., Thomson, N.R., Ramsey, K.H., and Clarke, I.N. (2018). The *Chlamydia muridarum* plasmid revisited : new insights into growth kinetics. Wellcome Open Res. *3*, 25.

94. Shao, L., Melero, J., Zhang, N., Arulanandam, B., Baseman, J., Liu, Q., and Zhong, G. (2017). The cryptic plasmid is more important for Chlamydia muridarum to colonize the mouse gastrointestinal tract than to infect the genital tract. PLoS ONE *12*, e0177691.

95. Rockey, D.D. (2011). Unraveling the basic biology and clinical significance of the chlamydial plasmid. J. Exp. Med. *208*, 2159–2162.

96. Hall, J.P.J., Wood, A.J., Harrison, E., and Brockhurst, M.A. (2016). Source-sink plasmid transfer dynamics maintain gene mobility in soil bacterial communities. Proc. Natl. Acad. Sci. USA *113*, 8260–8265.

97. Koonin, E.V. (2016). Horizontal gene transfer: essentiality and evolvability in prokaryotes, and roles in evolutionary transitions. F1000Res. 5, F1000 Faculty Rev–1805.

98. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. *25*, 1043–1055.

99. R Core Team (2018). R: A Language and Environment for Statistical Computing, Available at. https://www.R-project.org/.

100. Charif, D., and Lobry, J.R. (2007). SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis (Springer-Verlag Berlin Heidelberg).

101. Olm, M.R., Brown, C.T., Brooks, B., and Banfield, J.F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. *11*, 2864–2868.

102. Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M.C., Rattei, T., Mende, D.R., Sunagawa, S., Kuhn, M., et al. (2016). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. *44* (D1), D286–D293.

103. Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., von Mering, C., and Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. Mol. Biol. Evol. *34*, 2115–2122.

104. Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLoS Comput. Biol. *7*, e1002195.

105. Letunic, I., and Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. *47* (W1), W256–W259.

106. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

107. Emms, D.M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. *16*, 157.

108. Opgen-Rhein, R., and Strimmer, K. (2007). From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. BMC Syst. Biol. *1*, 37.

109. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504.

110. Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. Nat. Methods *9*, 471–472.

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

112. Dress, A.W.M., Flamm, C., Fritzsch, G., Grünewald, S., Kruspe, M., Prohaska, S.J., and Stadler, P.F. (2008). Noisy: identification of problematic columns in multiple sequence alignments. Algorithms Mol. Biol. *3*, 7.

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

114. Minh, B.Q., Nguyen, M.A.T., and von Haeseler, A. (2013). Ultrafast approximation for phylogenetic bootstrap. Mol. Biol. Evol. *30*, 1188–1195.

115. Lartillot, N., Rodrigue, N., Stubbs, D., and Richer, J. (2013). PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst. Biol. *62*, 611–615.

116. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44* (D1), D733–D745.

117. Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2016). GenBank. Nucleic Acids Res. *44* (D1), D67–D72.

118. Suzuki, H., Sota, M., Brown, C.J., and Top, E.M. (2008). Using Mahalanobis distance to compare genomic signatures between bacterial plasmids and chromosomes. Nucleic Acids Res. *36*, e147.

119. Le, S.Q., and Gascuel, O. (2008). An improved general amino acid replacement matrix. Mol. Biol. Evol. *25*, 1307–1320.

120. Quang, S., Gascuel, O., and Lartillot, N. (2008). Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics *24*, 2317–2323.

121. Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol. Biol. Evol. *35*, 518–522.

122. Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. *59*, 307–321.

123. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods *14*, 587–589.

124. Wang, H.-C., Minh, B.Q., Susko, E., and Roger, A.J. (2018). Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. Syst. Biol. *67*, 216–235.

125. Lartillot, N., Lepage, T., and Blanquart, S. (2009). PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics *25*, 2286–2288.

126. David, L.A., and Alm, E.J. (2011). Rapid evolutionary innovation during an Archaean genetic expansion. Nature *469*, 93–96.

127. To, T.-H., Jacox, E., Ranwez, V., and Scornavacca, C. (2015). A fast method for calculating reliable event supports in tree reconciliations via Pareto optimality. BMC Bioinformatics *16*, 384.

128. Kachitvichyanukul, V., and Schmeiser, B. (1985). Computer generation of hypergeometric random variates. J. Stat. Comput. Simul. *22*, 127–145.

129. Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. B *57*, 289–300.

Please cite this article in press as: Köstlbacher et al., Coevolving Plasmids Drive Gene Flow and Genome Plasticity in Host-Associated Intracellular Bacteria, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.10.030

Current Biology
Article

CellPress
OPEN ACCESS

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFER
Software and Algorithms		
checkM v1.0.7	98	https://ecogenomics.github.io/CheckM/
R v3.5.1	99	www.r-project.org/
'seqinr' package	100	https://cran.r-project.org/web/packages/seqinr/index.html
Drep v1.4.3	101	https://github.com/MrOlm/drep
eggNOG v4.5.1	102	http://eggnog45.embl.de/#/app/home
eggNOG-mapper v1.0.1	103	https://github.com/eggnogdb/eggnog-mapper
HMMER suite v3.1b2	104	http://hmmer.org/
ITOL v4	105	https://itol.embl.de/
BLAST suite v2.5.0+	106	https://blast.ncbi.nlm.nih.gov/Blast.cgi
OrthoFinder 2.0	107	https://github.com/davidemms/OrthoFinder
GeneNet 1.2.13 package	108	https://cran.r-project.org/web/packages/GeneNet/index.html
Cytoscape 3.7.0	109	https://cytoscape.org/index.html
ClusterONE 1.0 plugin	110	https://paccanarolab.org/cluster-one/
VOGDB v72	NA	http://vogdb.org/
MAFFT v7.222	111	https://mafft.cbrc.jp/alignment/software/
Noisy v1.5.12	112	http://www.bioinf.uni-leipzig.de/Software/noisy/
trimAl v1.4.1	113	https://github.com/scapella/trimal
IQ-TREE 1.6.2	114	http://www.iqtree.org/
PhyloBayesMPI 1.7a	115	https://github.com/bayesiancook/pbmpi
ecceTERA v1.2.4	76	https://mbb.univ-montp2.fr/MBB/download_sources/16__ecceTERA

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Matthias Horn (matthias.horn@univie.ac.at).

Material Availability

This study did not generate new unique reagents.

Please cite this article in press as: Köstlbacher et al., Coevolving Plasmids Drive Gene Flow and Genome Plasticity in Host-Associated Intracellular Bacteria, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.10.030

Data and Code Availability

Alignment files, tree files, and the python script are available at zenodo (https://zenodo.org/record/3859863).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

To assemble a comprehensive genome sequence dataset, we collected 26 publicly available Chlamydiae genomes from GenBank and 28 genomes of members of the PVC superphylum from the NCBI RefSeq database (Data S1 and Figure S1)[116,117]. All genomes were checked for completeness and contamination with checkM v1.0.7[98] using the "taxonomy_wf" setting and the marker gene set for bacteria. We included only genomes with greater than 85% completeness and lower than 5% contamination.

METHOD DETAILS

Comparison of trinucleotide signatures of plasmids and chromosomes

Genomic signatures of chlamydial plasmids and chromosomes were calculated as described in[56,118]. Briefly, we cut chromosomal sequences into non-overlapping 10,000 bp segments and calculated the occurrence of trinucleotides on both strands with the 'seqinr' package[100] in R 3.5.1[99]. We then calculated δ-distance and Mahalanobis distance for plasmid sequences against the mean chromosome signature. We calculated the probability of the distance of the plasmid signature to the mean chromosomal signature to be smaller than that of the chromosomal segments, here referred to as P (δ) or P (Mahalanobis)). We calculated a median probability of 0.65 (P(Mahalanobis), IQR 0.27- 0.82; Data S2A) and set a P(Mahalanobis) cutoff of 0.6 for defining highly similar plasmid and chromosomal pairs as proposed by[56].

Generation of a dereplicated plasmid dataset

To be able to assemble comprehensive datasets for phylogenetic analysis, which includes all relevant plasmid homologs we first generated a dereplicated RefSeq plasmid dataset. All 13,200 plasmids present in NCBI RefSeq[116] (July 2018, ftp://ftp.ncbi.nlm.nih.gov/refseq/release/plasmid/) were clustered with Drep v1.4.3[101] at a 90% ANI cutoff with primary clustering resulting in 4,736 representative plasmids. We then extracted the associated proteome of representative plasmids to generate a query database for plasmid-associated protein sequences.

Mapping to clusters of orthologous groups (COGs)

We mapped all proteins of our genome sequence dataset to eggNOG 4.5.1[102] to receive Clusters of Orthologous Groups (COG) classifications. We used eggNOG-mapper v1.0.1[103] with the bacteria optimized database using the "–database bact" option and default settings. For chlamydial plasmid encoded genes of interest with COG assignments we used the eggNOG provided HMM (hidden markov model) to screen the dereplicated RefSeq plasmid proteome for homologs. Using the *hmmsearch* program of the HMMER suite v3.1b2[104] with an e-value cutoff of 10^{-1} we first identified potential homologs which we then assigned to COGs with eggNOG-mapper as described above.

Mapping to viral orthology database

To be able to include homologs from virus genomes in our analysis, we downloaded all virus orthologous groups (VOGs) from VOGDB v72 (http://vogdb.org/). Using the *hmmpress* program of HMMER suite v3.1b2[104] we created a HMM database of all VOG HMMs. We searched plasmid encoded genes with the *hmmsearch* program with an e-value cutoff of 10^{-5} and selected the hits with the highest bitscore to assign VOGs for each gene.

Identification of gene families by *de novo* clustering of orthologous groups (OGs)

To infer gene relationships also for genes lacking representatives in public databases we performed *de novo* clustering of all proteins in our genome dataset. Protein sequences were aligned using the "*blastp*" program (BLAST suite v2.5.0+[106]) to compute sequence similarity scores between sequences with an expectation value cutoff of 10^{-3}. Using OrthoFinder 2.0[107] we clustered the proteins into orthogroups (OGs), referred to as gene families.

Partial correlation network analysis

To study co-occurrence of the most conserved gene families, i.e., those that were present on at least two plasmids, we performed correlation network analysis. We included all chlamydial plasmids but only used one representative of the Chlamydiaceae (*C. trachomatis* A/HAR-13) due to the high redundancy of members of this family with respect to plasmid gene content. A partial correlation network of conserved plasmid gene families was inferred using R 3.5.1[99] with the GeneNet 1.2.13 package[108] with default settings based on presence/absence patterns of 151 conserved plasmid gene families. Only statistically significant correlations with an FDR corrected p value ≤ 0.05 were retained. Gene families were clustered into groups in Cytoscape 3.7.0[109] with the ClusterONE 1.0 plugin[110] with default settings, except an overlap threshold of 10-3. Significant groups had a p value ≤ 0.05.

Please cite this article in press as: Köstlbacher et al., Coevolving Plasmids Drive Gene Flow and Genome Plasticity in Host-Associated Intracellular Bacteria, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.10.030

Current Biology
Article

CellPress
OPEN ACCESS

Phylogenetic analysis of COG and VOG-based datasets

For a detailed phylogenetic analysis of datasets assembled by mapping chlamydial proteins to COGs and VOGs, protein sequences were either aligned with MAFFT 7.222[111] using the "–localpair" and "–maxiterate 1000" parameter, or in the case of VOGs with the VOGDB-provided HMM. The ENOG4105C2U alignment was trimmed with Noisy v1.5.12[112], the ENOG4107QJE and VOG000016 alignments were trimmed with trimAl "-gappyout" to reduce the gap rate[113]. Identical sequences were removed prior alignment. Maximum likelihood phylogenies were calculated with IQ-TREE 1.6.2[114] under the empirical LG model[119]. We applied the same model testing regiment as proposed by Dharamshi et al.[53] with the empirical mixture models C10 to C60[120]. Because of the large number of sequences in the ENOG4107QJE dataset (n = 1,738), mixture model testing was restricted to C10 only. Support values were inferred from 1000 ultrafast bootstrap replicates[121] with the "-bnni" option for bootstrap tree optimization and from 1000 replicates of the (Shimodaira-Hasegawa) SH-like approximate likelihood ratio test[122]. Trees were visualized and edited using the Interactive Tree Of Life v4[105].

Species tree reconstruction

Species tree reconstruction was performed with the entire genome sequence dataset (Data S1). 43 conserved marker genes were extracted and aligned in checkM v1.0.7 with the "tree" workflow[98]. Bayesian tree samples with five MCMC chains in parallel (n = 10,000 each) were inferred using the CAT+GTR model[120] with 4 discrete gamma categories in PhyloBayesMPI 1.7a[115]. Convergence was assumed once the discrepancies in bipartition frequencies dropped below 0.1 and the effective sample sizes for continuous parameters were greater 100 (according to the *bpcomp* and *tracecomp* commands in PhyloBayes, respectively) after burnin (n = 2,500). Species tree was rooted according to Kamneva et al.[24] at the base of the Planctomycetes.

Gene tree-species tree reconciliation

We aligned all gene families (OGs) calculated with OrthoFinder using MAFFT 7.222[111] using the "–localpair" and "–maxiterate 1000" parameter. The protein alignments were trimmed with Noisy[112]. For each family with more than three sequences (n = 5,184) we reconstructed unrooted phylogenies with IQ-TREE 1.6.2[114] using the implemented ModelFinder[123] to find the appropriate model. The best fit model in combination with posterior mean site frequencies to model site heterogeneity[124] under the C20 model[125] was used to calculate 1,000 ultra-fast bootstrap samples for the downstream amalgamation procedure (n = 4964). 220 gene families had four or more sequences in total, but less than 4 unique sequences. There the only unrooted topology was used. We then performed gene tree-species tree reconciliation with ecceTERA v1.2.4[76], a program that implements a generic parsimony reconciliation algorithm, which accounts for duplications, losses and transfers, as well as speciation, and can accurately estimate species-tree aware gene trees using amalgamation[77]. We used the undated species tree mode "dated=0" without transfer from unsampled lineages "compute.TD=false." We calculated the average genome size flux[126] between ancestors for all fixed combinations of HGT cost 1-10 and duplication cost 1-10. For the ten cost vectors with minimal flux we calculated the mean support values of the symmetric median reconciliations for all gene trees (as proposed in[127]). We proceeded with cost settings of HGT = 3 and duplication = 1 (highest average support) for 4,624 gene families. For 542 gene families we used one of the alternative cost settings from the ten cost vectors with minimal genome size flux, if they were better supported.

Reconstruction of ancestral chlamydial plasmids and estimation of gene transfer frequencies

We used a custom python script to integrate over the computed gene family phylogenies. Briefly, we extracted the presence/absence information for all gene families and their evolutionary events from root to leaves of the species tree for the ecceTERA symmetric median reconciliations. We then summarized the reconstructed sets of gene families that were present in chlamydial LCAs and tracked speciation, duplication, and loss events, as well as horizontal transfers. To identify chlamydial gene families that are predominantly encoded on plasmids we analyzed the number of occurrences of each gene family on chlamydial chromosomes and plasmids (n = 3,091 with more than one chlamydial sequence), respectively. We used hypergeometric tests in the R base package phyper[128] with "lower.tail=T" to identify gene families that are significantly enriched on plasmids with a "BH"[129] corrected p value $\leq$ 0.05 using the R base package "p.adjust." pLCAs were then reconstructed based on these plasmid enriched gene families present in chlamydial LCAs. We calculated normalized gene transfers per gene family by dividing transfer events inferred by ecceTERA by the number of chlamydial branches in the gene tree (number of branches: 2 x (number of leafs - 1)). We then used a two-sided Wilcoxon signed rank test using the R base function "wilcox.test" to test for statistical significance.

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical tests and data analysis were performed in R version 3.5.1[99] and are described in the method details.

Current Biology, Volume *31*

Supplemental Information

Coevolving Plasmids Drive Gene Flow and Genome Plasticity in Host-Associated Intracellular Bacteria

Stephan Köstlbacher, Astrid Collingro, Tamara Halter, Daryl Domman, and Matthias Horn

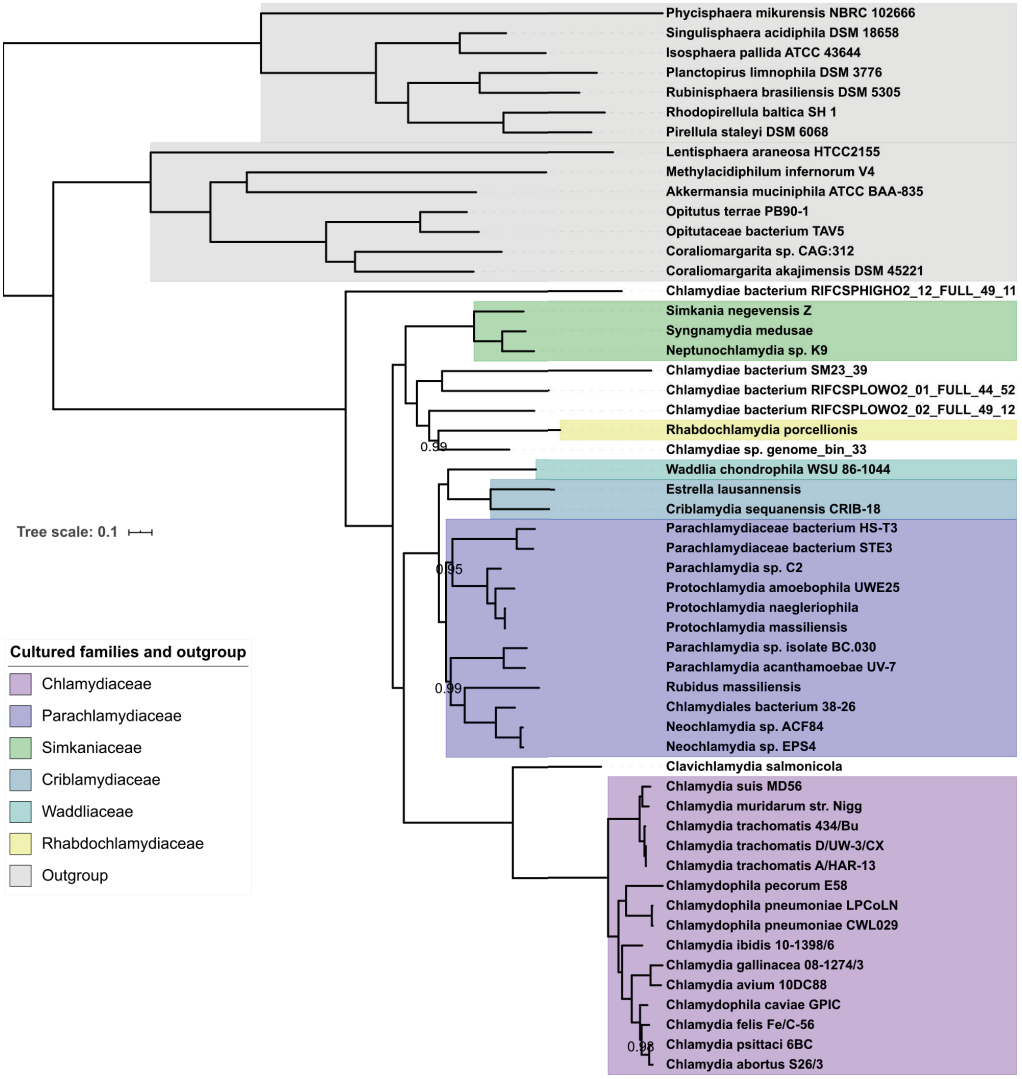


Figure S1: Chlamydial species tree based on 43 conserved marker genes. Related to Figure 1. Chlamydial families with cultured representatives are indicated, outgroup colored in grey. Species tree is rooted according to Kamneva et al. [S1] at the base of the Planctomycetes. Bayesian phylogeny with the CAT profile mixture model and GTR model of substitution, based on 3 converged, independent chains. Only posterior probabilities < 1 are indicated as numbers at splits.

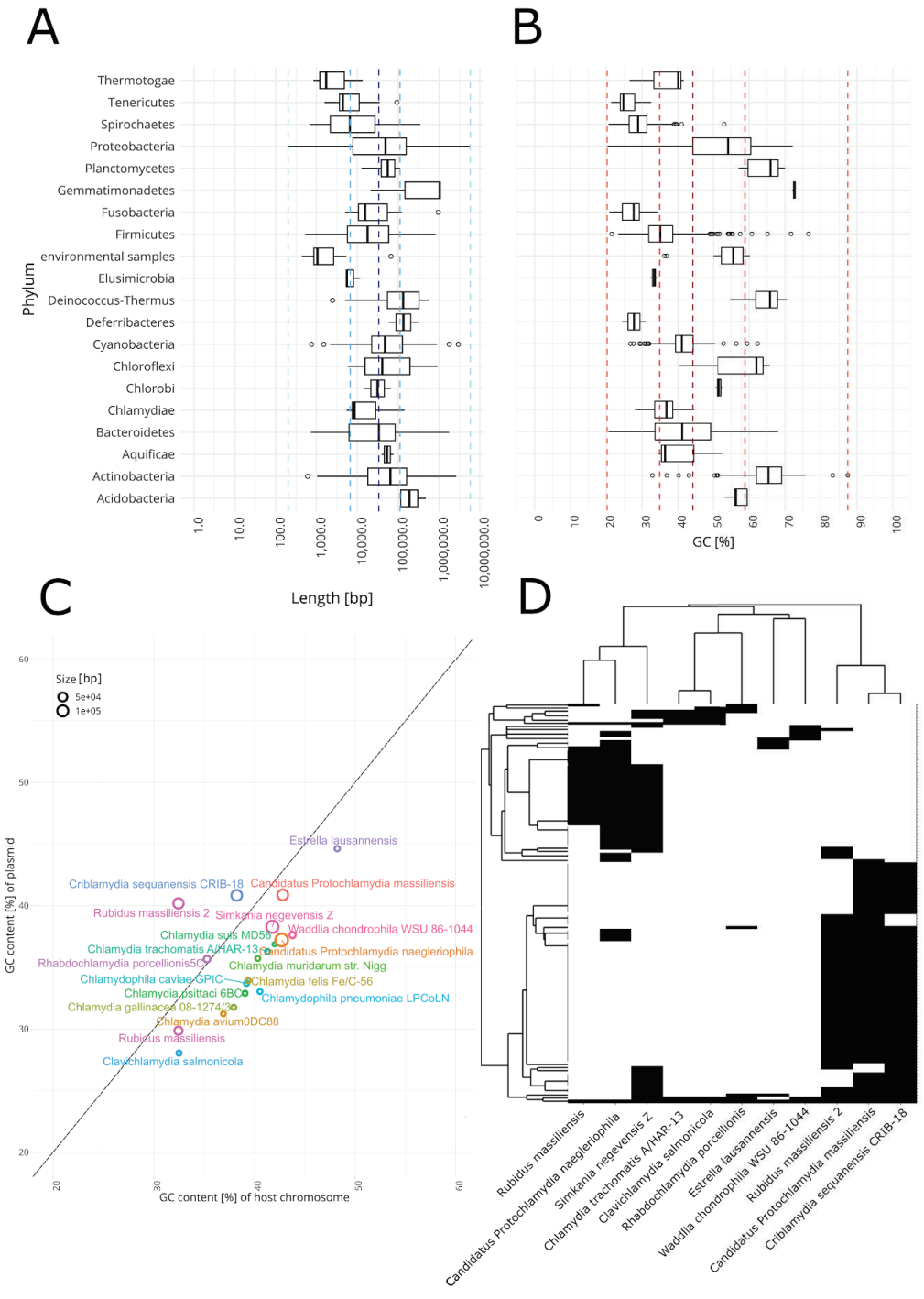


Figure S2: Comparison of chlamydial plasmids with other phyla, the plasmid host organisms, and phyletic pattern within chlamydiae. Related to Figures 1 and 2. (A)

Size distribution of dereplicated RefSeq plasmids in logarithmic scale plotted as boxplots with whiskers by host phylum. The central dark blue dashed line indicates median size of all plasmids. Lighter blue lines next to median indicates the middle 50% range, light blue lines indicate lowest or highest 25% of the data. Chlamydiae plasmids fall with sizes from 7,510 (*Chlamydia trachomatis* A/HAR-13) - 145,285 bp (*Protochlamydia naegleriophila*) into the interquartile range (IQR; range between first and third quartile) of bacterial plasmids (IQR 7- 110 kb, median 50 kb). (**B**) GC distribution of dereplicated RefSeq plasmids plotted as boxplots with whiskers by host phylum. The central dark red dashed line indicates the median size of all plasmids. Lighter red lines next to median indicate the middle 50 % range, light red lines indicate lowest or highest 25% of the data. The GC content of chlamydial plasmids ranges from 28 % (*Clavichlamydia salmonicola*) to 44.6 % (*Estrella lausannensis*) and is thus slightly lower than in most other bacterial phyla (IQR 35- 58 % GC, median 44 % GC). (**C**) %GC of host chromosomes plotted against %GC of plasmids. Most chlamydial plasmids have a lower GC content than the respective host chromosome and are therefore likely coevolving with the host for a prolonged time period. The dashed diagonal line indicates an equal %GC of host chromosome and plasmid. Like host dependent bacteria, plasmids tend to have a lower GC content than their hosts [S2] and the differences are highly correlated between plasmids and host chromosomes [S3]. Chlamydiae plasmid GC content is significantly correlated with the host chromosome and is also on average 4.8 % lower (Pearson's correlation coefficient r = 0.603, p-value = 0.005). **(D)** Binary map representing chlamydial plasmids on the y-axis and the 216 conserved plasmid gene families on the x-axis**.** Plasmid dendrogram based on the binary jaccard distance of gene family presence and absence.

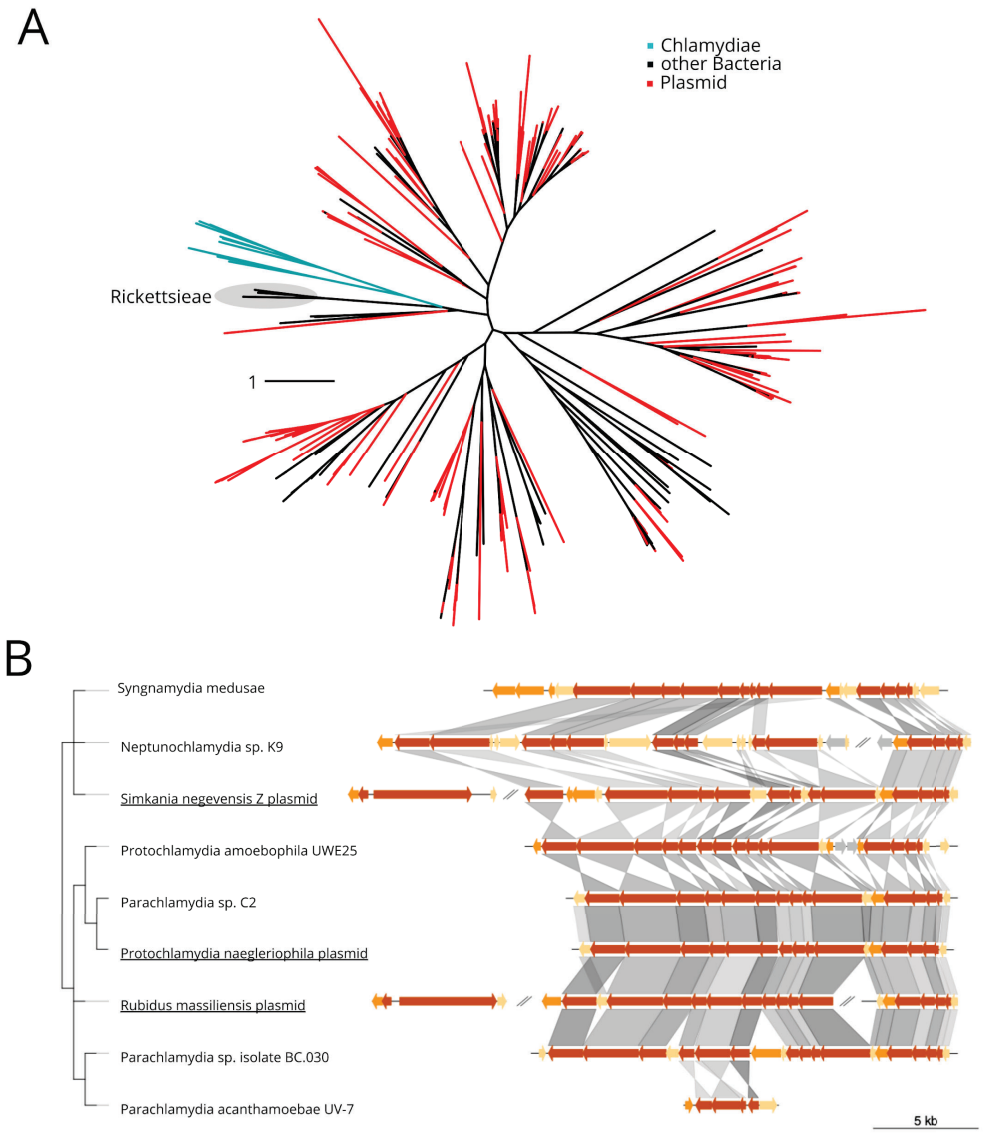


Figure S3: The chlamydial T4SS is monophyletic and is conserved on multiple plasmids and chromosomes. Related to Figure 2. (**A**) Approximate maximum likelihood (FastTree) phylogenetic tree of *traN* (OG0002252) with LG model with 1000 parametric bootstraps. Bootstrap support for monophyly of chlamydial chlade and monophyly with tribe Rickettsieae (Alphaproteobacteria) ≥ 0.95. Turquoise indicates chlamydial branches, black other bacterial branches, and red plasmid genes from the dereplicated RefSeq plasmid dataset. (**B**) Gene organization of the T4SS in chlamydiae on the backbone of the species tree. Underlined species names indicate plasmid encoded T4SS loci. Marked loci encode tra genes, if not named otherwise.

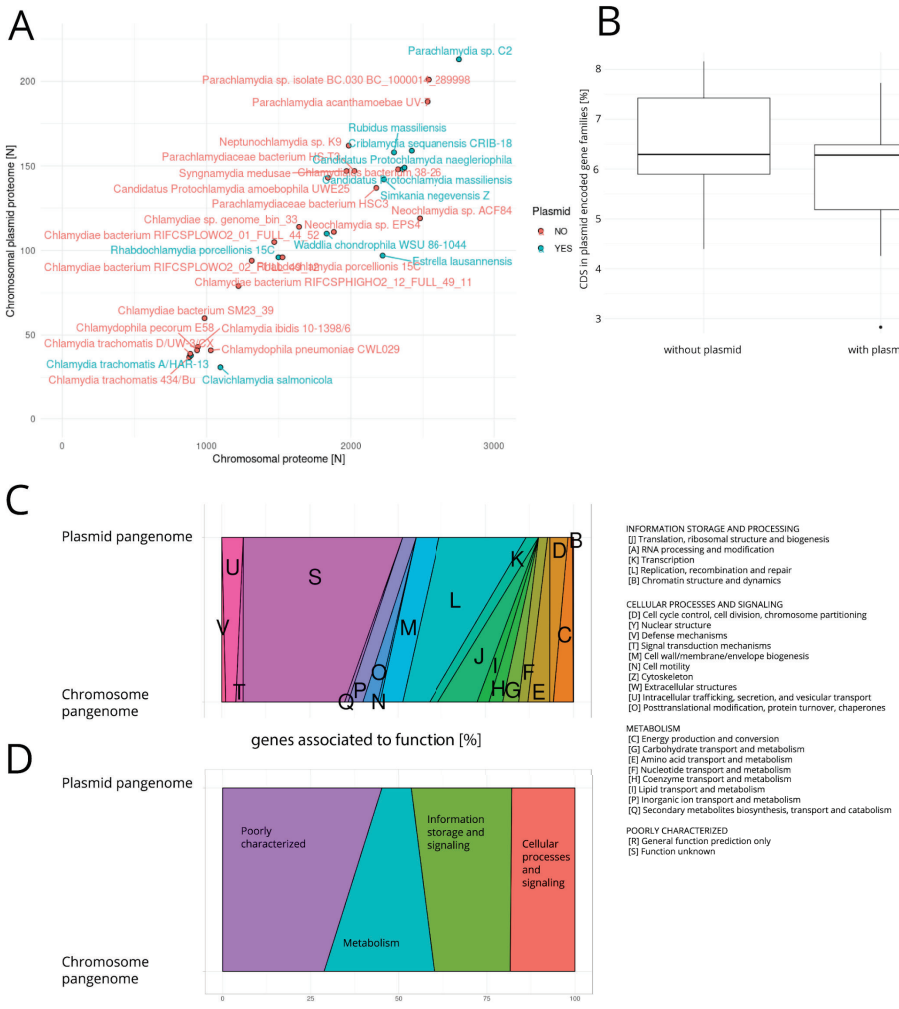


Figure S4: Plasmid gene content on chlamydial chromosomes and comparison of plasmid and chromosome functional profiles. Related to Figure 4. (**A**) Scatter plot comparing chromosomal proteome size and number of chromosomally encoded plasmid gene families. (**B**) Boxplot comparing plasmid carrying and plasmidless organisms (excluding metagenome assembled genomes, MAGs) by the number of chromosomally encoded CDS belonging to plasmid gene families. (**C**,**D**) Functional profiles of plasmid vs. chromosomal pangenome gene families based on (**C**) single eggNOG functional categories or (**D**) larger functional groups.

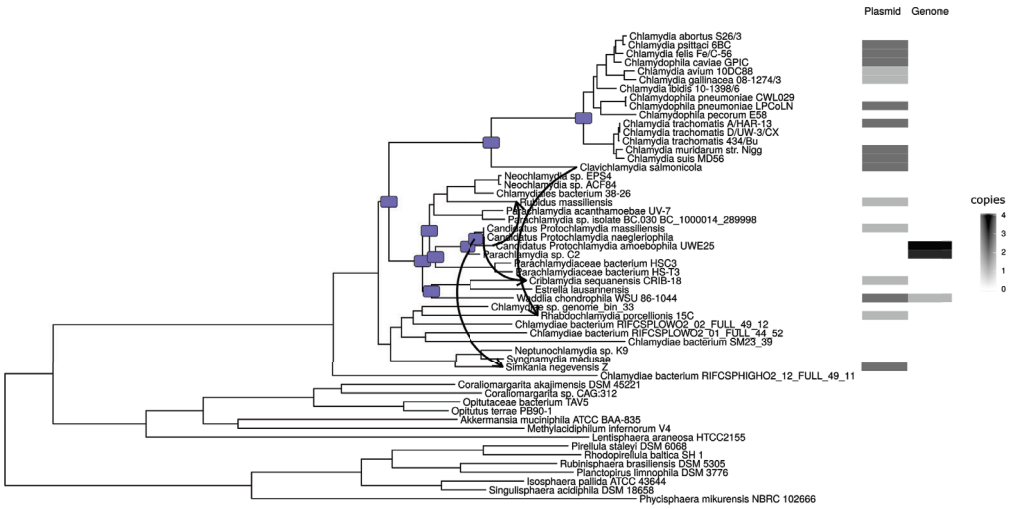


Figure S5: Gene tree reconciled evolutionary history of the integrase Pgp7/8 mapped on species tree on the left shows four transfers, but otherwise vertical transmission. Related to Figure 5. Violet squares on nodes on the tree indicate presence of Pgp7/8 in the ancestor. Bars on the right show copy number on plasmids or chromosome.

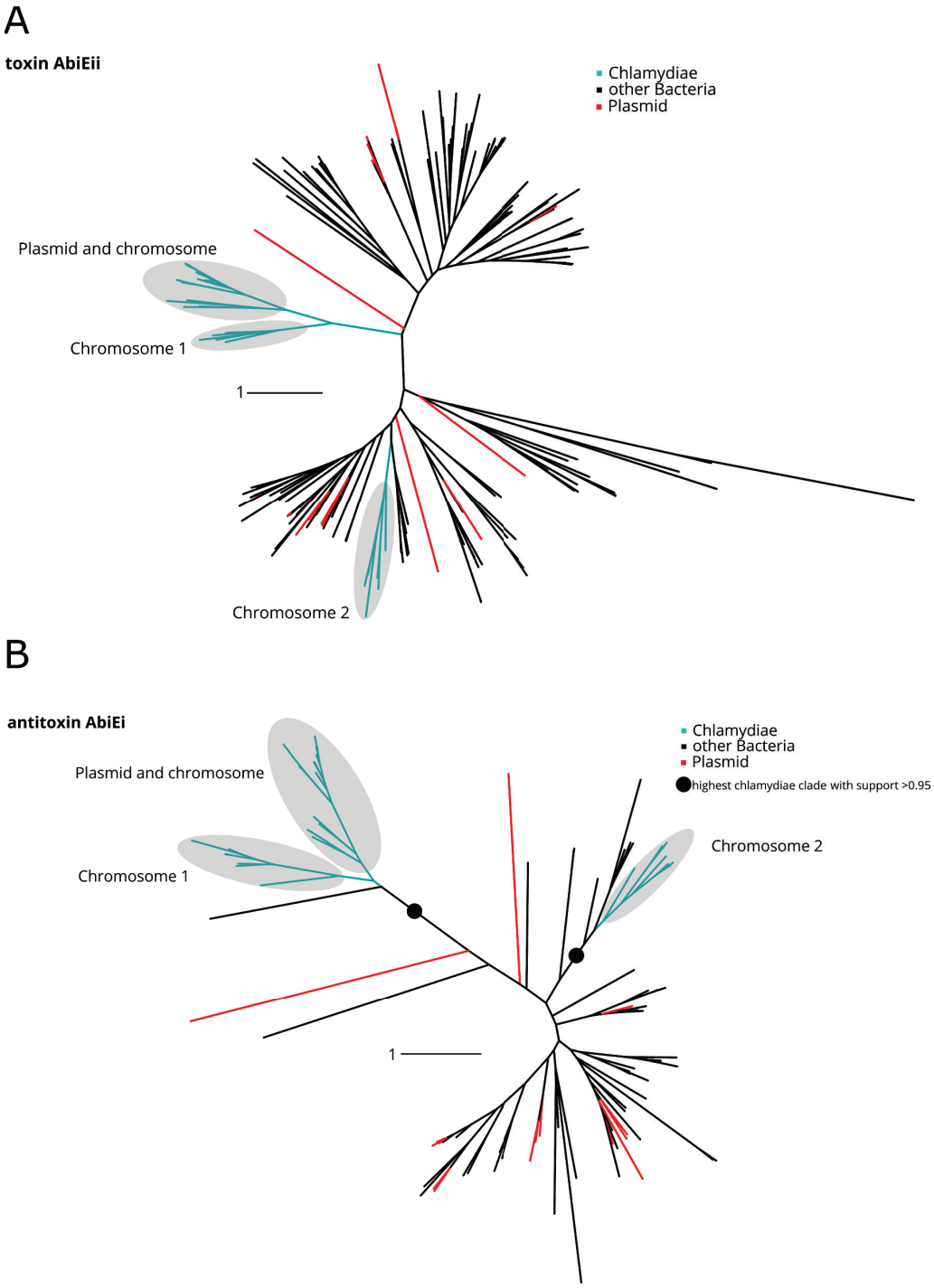


Figure S6: Independent acquisition of two toxin-antitoxin systems. Related to Figure 6. (**A**) Two monophyletic clades of toxin AbiEii have been acquired independently. Approximate maximum likelihood (FastTree) phylogenetic tree of type IV toxin-antitoxin

'innate immunity' bacterial abortive infection (Abi) system toxin AbiEii (eggNOG COG ENOG4105F2S, OG0000561) with LG model with 1000 parametric bootstraps. Bootstrap support for monophyly of chlamydial clades always ≥ 0.95. Turquoise indicates chlamydial branches, black other bacterial branches, and red plasmid genes from the dereplicated RefSeq plasmid dataset. (**B**) Two monophyletic clades of antitoxin AbiEi have been acquired independently. Approximate maximum likelihood (FastTree) phylogenetic tree of type IV toxin-antitoxin 'innate immunity' bacterial abortive infection (Abi) system antitoxin AbiEi (eggNOG COG ENOG4107VIN, OG0000768) with LG model with 1000 parametric bootstraps. Bootstrap support for monophyly of chlamydial chlades was < 0.95 so the first node traversing inwards the tree ≥ 0.95 was indicated with a black circle. Turquoise indicates chlamydial branches, black other bacterial branches, and red plasmid genes from the dereplicated RefSeq plasmid dataset. In addition to the monophyletic antitoxin encoded on both plasmid and chromosomes, an independent acquisition of a distantly related antitoxin occurred in some chlamydiae.

Plasmid gene family	Number of chromosomes*	Conservation on chromosomes [%]	Number of plasmids*	Conservation on plasmids[%]	EggNOG annotation
OG0000076	25	92.59	12	100.00	ParA like protein (Pgp5)
OG0000311	27	100.00	6	50.00	Virulence plasmid gene (Pgp6) - function unknown
OG0000038	27	100.00	3	25.00	Heavy metal translocating P-type ATPase
OG0000197	27	100.00	3	25.00	Histone-like DNA-binding protein
OG0000162	26	96.30	4	33.33	Replicative DNA helicase (Pgp1)
OG0000005	27	100.00	2	16.67	Short-chain dehydrogenase reductase Sdr
OG0000006	25	92.59	2	16.67	Methyltransferase
OG0000033	27	100.00	1	8.33	Dihydrolipoyl dehydrogenase
OG0000058	27	100.00	1	8.33	Phosphotransferase system, IIa
OG0000648	27	100.00	1	8.33	BAF60b domain protein
OG0000674	27	100.00	1	8.33	Function unknown
OG0000036	26	96.30	1	8.33	ABC transporter, ATP-binding protein
OG0000060	26	96.30	1	8.33	Aminotransferase
OG0000109	26	96.30	1	8.33	Peptidyl-prolyl cis-trans isomerase
OG0000196	25	92.59	1	8.33	Chaperone protein ClpB

Table S1: Plasmid gene families well-represented on chlamydial chromosomes. Related to Figure 4. *Total number of chlamydial genomes with plasmids in this comparison is 27. Number of plasmids is 12. Only two chlamydiaceae genomes were used for this table because of 100 % plasmid gene content redundancy (Chlamydia trachomatis A/HAR-13 and Chlamydophila pneumoniae LPCoLN).

OG	Function	bactNOG
OG0000076	Pgp5	ENOG4105C2U
OG0000162	Pgp1 Replicative dna helicase	ENOG4105CDU
OG0000311	Pgp6	ENOG4106MFV
OG0000031	YD repeat protein	ENOG4108ADK
OG0000038	heavy metal translocating p-type ATPase	ENOG4105C59
OG0000197	Histone-like DNA-binding protein which is capable of wrapping DNA to stabilize it, and thus to prevent its denaturation under extreme environmental conditions (By similarity)	ENOG41082SS
OG0000621	Efflux transporter rnd family, mfp subunit	ENOG4105EDC
OG0000707	GrpB protein	ENOG4105RHY
OG0001837	Conjugal transfer ATPase	ENOG4105EJX
OG0001972	Hypothetical protein	ENOG4106587
OG0002052	Outer membrane efflux protein	ENOG4107XZU

Table S2: Gene families reconstructed in chlamydiae pLCA. Related to Figure 6.

Supplemental references

S1. Kamneva, O.K., Knight, S.J., Liberles, D.A., and Ward, N.L. (2012). Analysis of genome content evolution in pvc bacterial super-phylum: assessment of candidate genes associated with cellular organization and lifestyle. Genome Biol. Evol. *4*, 1375–1390.

S2. Rocha, E.P.C., and Danchin, A. (2002). Base composition bias might result from competition for metabolic resources. Trends Genet. *18*, 291–294.

S3. Nishida, H. (2012). Comparative analyses of base compositions, DNA sizes, and dinucleotide frequency profiles in archaeal and bacterial chromosomes and plasmids. Int. J. Evol. Biol. *2012*, 342482.

CHAPTER IV

Pangenomics reveals alternative environmental lifestyles among chlamydiae

Authors:

Stephan Köstlbacher, Astrid Collingro, Tamara Halter, Frederik Schulz, Sean P. Jungbluth, and Matthias Horn

Published in:

Nature Communications (2021)

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-24294-3 OPEN

Pangenomics reveals alternative environmental lifestyles among chlamydiae

Stephan Köstlbacher [1,3], Astrid Collingro [1], Tamara Halter [1], Frederik Schulz [2], Sean P. Jungbluth [2] & Matthias Horn [1✉]

Chlamydiae are highly successful strictly intracellular bacteria associated with diverse eukaryotic hosts. Here we analyzed metagenome-assembled genomes of the "Genomes from Earth's Microbiomes" initiative from diverse environmental samples, which almost double the known phylogenetic diversity of the phylum and facilitate a highly resolved view at the chlamydial pangenome. Chlamydiae are defined by a relatively large core genome indicative of an intracellular lifestyle, and a highly dynamic accessory genome of environmental lineages. We observe chlamydial lineages that encode enzymes of the reductive tricarboxylic acid cycle and for light-driven ATP synthesis. We show a widespread potential for anaerobic energy generation through pyruvate fermentation or the arginine deiminase pathway, and we add lineages capable of molecular hydrogen production. Genome-informed analysis of environmental distribution revealed lineage-specific niches and a high abundance of chlamydiae in some habitats. Together, our data provide an extended perspective of the variability of chlamydial biology and the ecology of this phylum of intracellular microbes.

[1] Centre for Microbiology and Environmental Systems Science, University of Vienna, Vienna, Austria. [2] DOE Joint Genome Institute, Berkeley, CA, USA. [3] Present address: Laboratory of Microbiology, Wageningen University and Research, Wageningen, The Netherlands. ✉email: matthias.horn@univie.ac.at

Microbes specialized to live inside eukaryotic cells are diverse and have emerged independently among various bacterial and archaeal taxa. This includes pathogens of humans as well as beneficial symbionts of animals, overall with a major impact on the life around us[1]. Intracellular bacteria are generally studied in the context of a particular host, e.g., with respect to a disease or nutritional interactions, and focused on groups of closely related microorganisms. One of the most diverse, successful, and ancient bacterial lineages intimately associated with eukaryotes is the phylum Chlamydiae[2,3]. Studying these microbes has the potential to understand the variability and evolution of the intracellular lifestyle in a much broader context, across an array of different eukaryotic hosts, environments, and over extended evolutionary time scales.

The Chlamydiae are part of the Planctomycetes-Verrucomicrobia-Chlamydiae (PVC) superphylum, a group that, apart from Chlamydiae, predominantly consists of free-living bacteria of environmental and biotechnological importance[4,5]. Chlamydiae were long thought to consist of a single family, the Chlamydiaceae, including several well-known human and animal pathogens[3,6]. Yet, molecular diversity surveys suggest the existence of hundreds of chlamydial families in a great range of different environments[7,8]. Our knowledge about these microbes, commonly referred to as environmental chlamydiae[2], is sparse, except that many of them are likely associated with protist hosts[7,9]. These unicellular eukaryotes are ubiquitous and make up more than twice the biomass on earth than all animals combined[10]. However, the isolation and cultivation of chlamydiae is challenging and was so far only successful for members of six chlamydial families[7,11]. Confounding factors include their strict dependence on eukaryotic host cells, the fact that the natural host is often unidentified[11], and unknown growth conditions aggravating the cultivation of protists. Despite the phylum-level diversity of chlamydiae, their intracellular lifestyle appears to be well-conserved as all cultured representatives share a unique developmental cycle consisting of alternation between an infectious extracellular stage, the elementary body (EB), and an intracellular replicative stage, the reticulate body (RB)[6].

In the face of the experimental challenges associated with the intracellular lifestyle and for a long time the lack of methods to genetically modify chlamydiae[12], genomics of cultured representatives has been of particular importance to understand chlamydial biology and host interaction[13–16]. Recent advances in metagenomics and single cell genomics enabled the recovery of single cell amplified genomes (SAGs) and metagenome-assembled genomes (MAGs) from yet uncultured chlamydiae despite their generally low abundance in complex microbial communities[16–22]. This revealed a number of surprising findings and provided a first glimpse at the genomic versatility of environmental chlamydiae[18,19]. For instance, marine chlamydial SAGs encoded a complete flagellar apparatus, while all known chlamydiae are non-motile[18]. Furthermore, chlamydial MAGs were strikingly abundant in anoxic marine deep sea sediments[19,23]. This was particularly unexpected as chlamydiae had been considered aerobic or microaerobic microbes. In contrast, the anoxic sediment MAGs showed features indicative for an anaerobic metabolism[19,23]. Previous studies have consistently described hundreds of genes conserved in all or nearly all members of the phylum Chlamydiae[15,16,20], denoting a large core genome[24]. The accessory genome, i.e., the set of genes encoded only in one or few representatives, indicates potential niche or host-specific adaptations and seems to be expanded in environmental chlamydiae—although comprehensive analyses are missing so far. More generally, the pangenome, i.e., the sum of core and accessory genome, can give insights into habitat specificity and evolutionary forces shaping microbial genomes[24].

Here we used the Chlamydiae as a model to study the variability of the intracellular lifestyle in the context of an entire bacterial phylum and a global genome sequence dataset. To this end, we capitalized on the Genomes from Earth's Microbiomes (GEM) initiative, which represents a comprehensive collection of MAGs from diverse environments worldwide[25] (https://genome.jgi.doe.gov/GEMs). Our analysis of chlamydial MAGs from this resource expands recognized chlamydial taxonomic richness based on genomic data by almost doubling representatives at the species and genus rank. We discovered additional chlamydial families and provide evidence for surprisingly widespread distribution of the potential for anaerobic metabolism as well as a number of other niche-specific adaptations. Genome-informed mining of public 16S ribosomal RNA (16S rRNA) gene data revealed distinct and lineage-specific environmental preferences, with many yet uncultured chlamydiae reaching high abundances and being found in diverse aquatic systems.

Results and discussion

A phylogenomic perspective on chlamydial diversity. In total, 82 MAGs from the GEM dataset were classified as members of the phylum Chlamydiae[25]. Phylogenomic analysis using a set of 43 conserved marker proteins confirmed that all MAGs are of chlamydial origin and distributed throughout the chlamydial species tree obtained with a reference dataset including published and few newly determined genome sequences (Fig. 1, Supplementary Data 1 and 2). In line with MIMAG standards (Minimal Information about a Metagenome-Assembled Genome)[26], 67 MAGs have medium quality corresponding to an estimated genome completeness over 50% and contamination lower than 10%. The remaining 15 MAGs are high quality with an estimated completeness of over 90%, contamination under 5%, a full-length 16S rRNA gene, and more than 18 tRNAs (Fig. 1, Supplementary Data 1).

Consistent with known chlamydial genomes, the 82 MAGs show a reduced (estimated) genome size (0.9–2.6 Mb, average 1.6 Mb) and a moderately low average GC content (42.6%, range 25.9–49.8%; Fig. 1, Supplementary Data 1). In general, chlamydiae associated with multicellular eukaryotes have smaller genomes, while chlamydial symbionts of protists show larger genome sizes[15,19]. The MAGs from this study might thus represent both animal and protist-associated chlamydiae.

Based on our de novo species tree (Fig. 1a), we estimated the level of taxon sampling in the chlamydiae by calculating phylogenetic diversity and phylogenetic gain, representing the sum of branch lengths in the tree and the added branch lengths by a group of taxa, respectively[27] (Supplementary Data 3). The added MAGs represented 39.5% of the total branch length in the chlamydial species tree, thus almost doubling the known chlamydial phylogenetic diversity.

Next, we inferred the environmental origin of the MAGs using metadata from the Integrated Microbial Genomes and Microbiome database IMG/M[28] supplemented by additional information from the literature (Fig. 1 and Supplementary Data 4). More than two-thirds of the MAGs are derived from aquatic sources and terrestrial habitats ($n = 38$ marine and freshwater microbiomes; $n = 24$ soil and plant microbiomes), further supporting a ubiquitous occurrence of chlamydiae in the environment[7]. These findings reflect 16S rRNA gene based studies, suggesting marine, freshwater, soil, and plant systems as environmental reservoirs of chlamydiae[8,29]. Most of the additional diversity observed here is due to MAGs from freshwater and marine environments (19.2% and 10.9% phylogenetic gain, respectively; Supplementary Data 5). Soils add only 2.5% phylogenetic gain, indicating that this environment already has been well sampled with respect to

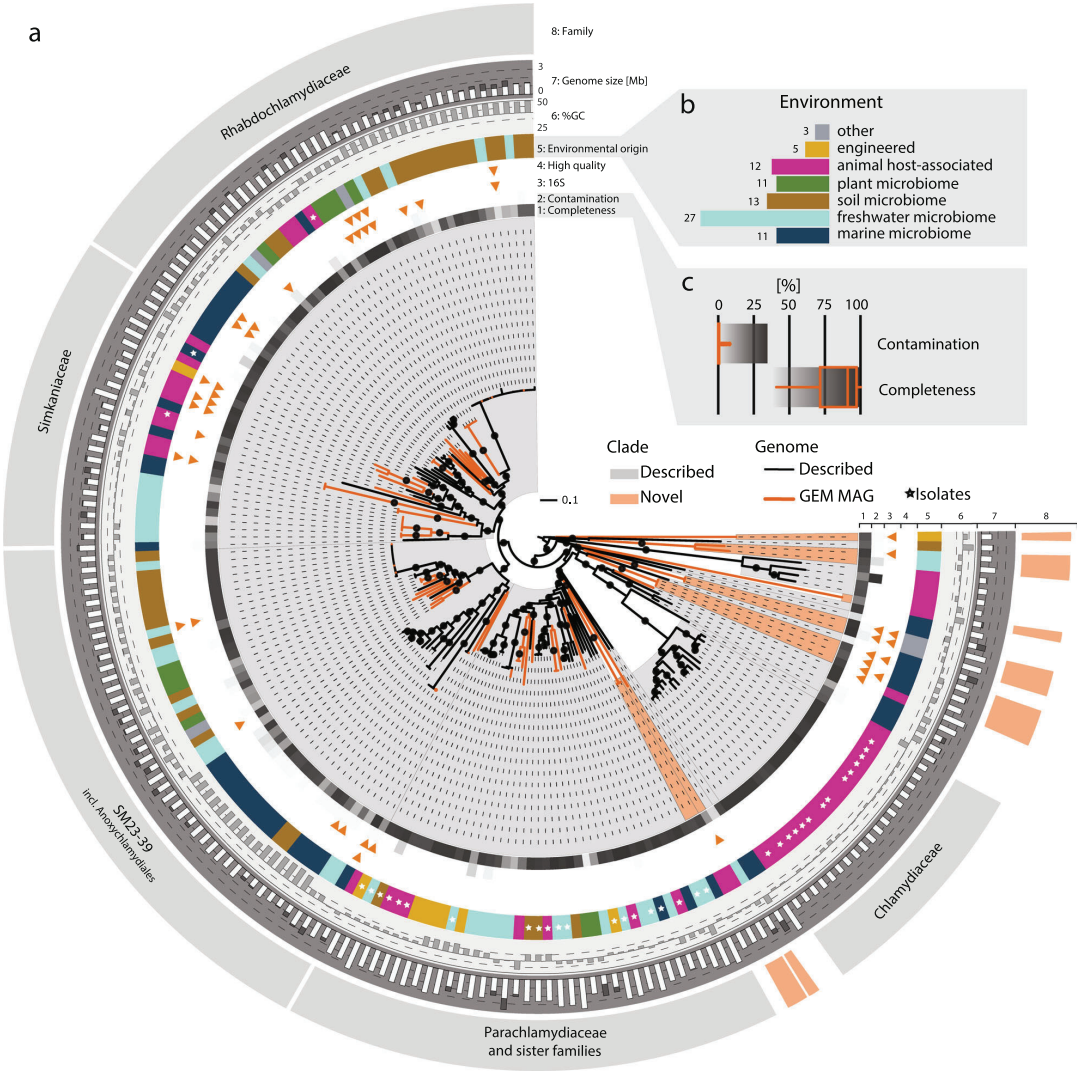


Fig. 1 MAGs from diverse environments expand known and add previously undescribed clades in the phylum Chlamydiae. a Maximum likelihood phylogenetic tree based on a concatenated set of 43 conserved marker proteins (5704 sites) in which published genomes and 82 MAGs generated in the GEM initiative are shown in black and orange, respectively. Chlamydial monophyly was supported by optimized ultrafast bootstrap and SH likelihood ratio test support with 100% for both. Previously established chlamydial families are shaded in light gray, previously undescribed families are shaded in orange. The tree was inferred under the LG + C60 + G4 + F model with the IQ-TREE software. Nodes with an optimized ultrafast bootstrap support ≥ 95% are labelled with black circles. Tree annotations from inside to outside: (1) completeness, (2) contamination, (3) MAG with 16S rRNA gene, (4) high quality MAGs, (5) environmental origin (white stars indicate genomes from cultured isolates), (6) %GC content, (7) assembly size and estimated genome size (stacked white and gray bars, respectively), and (8) names of chlamydial families represented by more than ten genomes and added metagenomic clades indicated by orange segments. Scale bar indicates 0.1 substitutions per position in the alignment. **b** Number of MAGs retrieved per environmental category. **c** Completeness and contamination estimates of chlamydial MAGs from the GEM dataset. Shaded gradients behind the completeness and contamination boxplots represent the values in the heatmap boxes in the tree.

chlamydial diversity. Notably, the second highest total phylogenetic gain (18.1%) was obtained from MAGs detected in host-associated animal microbiomes (Supplementary Data 1). There are no known chlamydiae infecting plants[7,30,31], and consistent with this, MAGs from plant microbiomes were mostly derived from rhizosphere, rhizoplane, and phyllosphere samples, with the exception of three MAGs originating from surface-sterilized *Populus* roots, i.e., the endosphere[32].

Metagenomics-driven discovery of taxa. To assign chlamydial MAGs to taxonomic units, we used the relative evolutionary distance (RED) approach of the Genome Taxonomy Database GTDB[33]. We classified all MAGs with the GTDB-tk toolkit and used our de novo species tree as additional reference and for refinement, as the GTDB framework only allows classification to known taxa in the database. Sixty-nine MAGs were assigned to five existing chlamydial families. All were confirmed by our

species tree except for three MAGs (1039677-28, 1039689-34, and 1039701-25), which represented a highly supported sister clade to the GTDB family GCA-270938. Consistent with this grouping RED values indicated that the three MAGs establish a separate family, for the purpose of this study referred to as Metagenomic Chlamydial Family MCF-E (Supplementary Data 7). In total, 13 MAGs represent seven previously undescribed family-rank clades, derived mostly from aquatic environments and denoted here as MCF-A to MCF-G (Fig. 2; Supplementary Data 1, 6, and 7; families MCF-D and MCF-E represented by high quality MAGs, the other families by medium quality MAGs according to MIMAG standards).

To better understand the taxonomic diversity within chlamydial families, we used a whole genome average nucleotide identity (ANI) and average amino acid identity (AAI) based clustering to resolve the species and genus rank, respectively (Supplementary Figs. 1–3, Supplementary Data 4). We identified 54 species in 44 genera among the 12 chlamydial families that contained the MAGs from the current study. The GEM dataset comprises more previously unknown than described chlamydial taxa on all taxonomic ranks analyzed, including 44 species and 34 genera (Fig. 2). Notably, the highest number of added genera was found in families whose members are traditionally considered environmental representatives of the phylum, often associated with amoeba or arthropods. This includes the Parachlamydiaceae, the Simkaniaceae, and the Rhabdochlamydiaceae (in GTDB v89 named Ga0074140) (Fig. 2, Supplementary Data 7). In addition, the recently described family SM23-39 (also referred to as Limichlamydiaceae or Anoxychlamydiales)[17,19,20], so far represented by MAGs exclusively, includes seven additional genera.

In total, after the addition of the MAGs from the GEM catalog, 117 chlamydial species, 94 genera, and 21 families are currently represented with genomic data, leading to an increase of 60%, 57%, and 50%, at the respective taxonomic rank (Fig. 2; Supplementary Data 7). Our analysis thus corroborates the large chlamydial diversity estimates inferred from 16S rRNA gene surveys. The additional genome data provides an important step toward understanding chlamydial diversity in the environment.

A stable lifestyle-reflecting core genome and genomic plasticity in environmental lineages. Genes shared across all genomes of a set of organisms, also referred to as the core genome, provide evidence for conserved biological features[24]. We de-replicated all 192 genomes of our dataset at 99% ANI to reduce redundancy and only included genomes that were the most complete (≥85%) and the least affected by contamination (≤5%). This resulted in a representative dataset for the phylum of 96 genome sequences (Supplementary Data 8). We next inferred non-supervised orthologous groups (NOGs) corresponding to gene families represented in the dataset[34,35]. 375 NOGs were conserved among more than 90% of all genomes, forming the chlamydial core genome (Supplementary Fig. 4). This amounts to a median of 25% of NOGs per genome (interquartile range 22–34%). The core

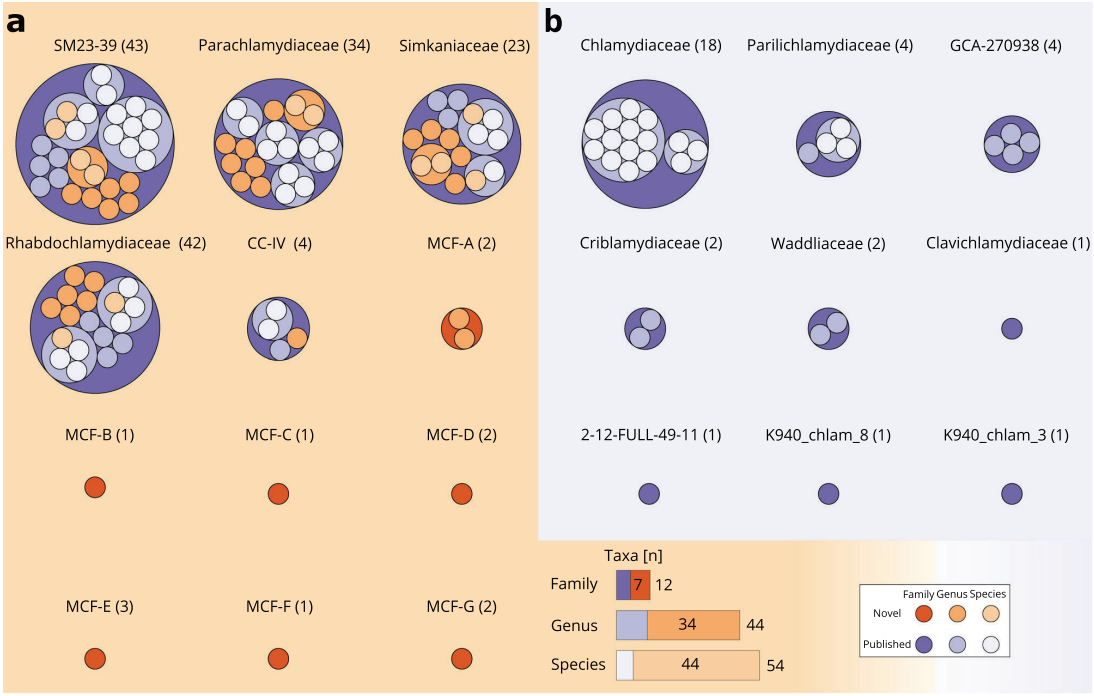


Fig. 2 MAGs from the GEM dataset broadly populate the taxonomy of the Chlamydiae at family-, genus-, and species rank. MAGs from the GEM catalog significantly extend known chlamydial taxa, including 7 additional families, 34 genera, and 44 species, highlighting the taxonomic heterogeneity of the phylum. Packed circles represent chlamydiae taxonomic ranks and their higher level taxonomic structure. From the outermost to the innermost circle the family, genus, and species ranks are depicted. Violet indicates lineages with previously known genome representatives (family, genus, species rank in dark, medium, and light violet, respectively), while added lineages are shown in orange (family, genus, species in dark, medium, and light orange, respectively). The number in brackets next to the family names indicates the number of genome sequences available. **a** Known and previously undescribed chlamydial families containing MAGs from the GEM catalog. Bar charts represent the number of families, genera, and species recruited in this study. **b** Families without genome sequences from this study.

genome size is thus in the range of early estimates including genomes of only ten cultured representatives ($n = 560$, core genome: conserved in all genomes)[15], and more recent analyses including a few MAGs and SAGs ($n = 108$; no quality filtering; core genome: conserved >90% of genomes)[20], or genomes of six chlamydial families, respectively ($n = 342$; core genome: conserved in all genomes)[16].

The chlamydial core genome determined with our dataset encompasses a number of recognized features facilitating the lifestyle of known chlamydiae and representing hallmarks of intracellular microbes (Fig. 3). This includes the non-flagellar type III secretion system (T3SS), a key virulence mechanism translocating effector proteins into chlamydial host cells[15,19,36,37]. Effector proteins interfere with host cellular pathways as

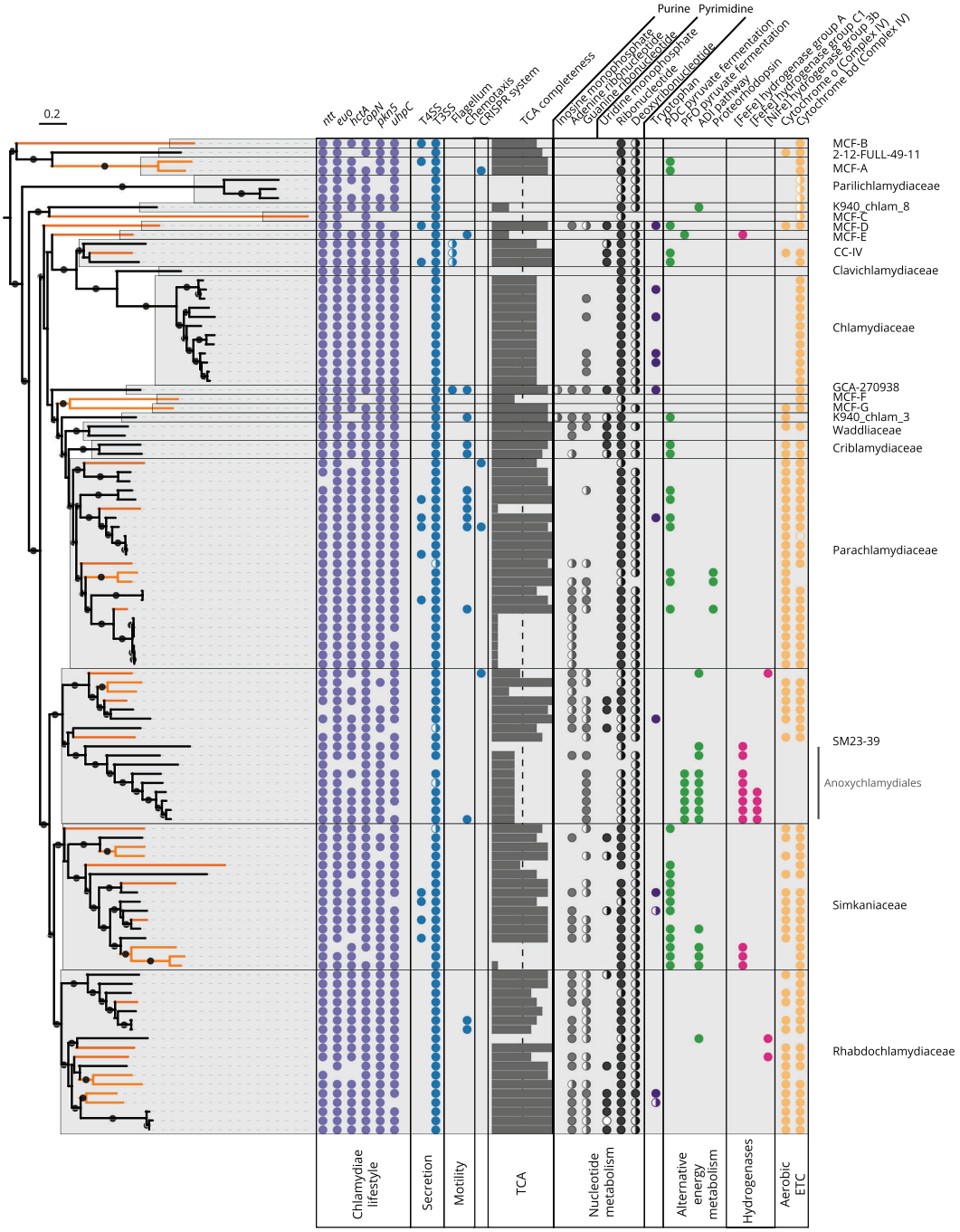


Fig. 3 Chlamydiae show conserved features of an intracellular lifestyle but versatility in oxygen adaptation. The presence of selected genes and pathways across the chlamydial core and accessory genome is depicted. The phylogenetic tree includes 96 high quality genomes used for pangenome analysis and additional representatives ($n = 109$ in total). The tree is based on a concatenated set of 43 conserved marker proteins (6268 sites) and was inferred under the LG + C60 + G4 + F derived PMSF approximation by the IQ-TREE software. Branch support values are based on 100 non-parametric bootstraps, support ≥ 70% is indicated as black circles. MAGs from the GEM catalog are indicated by orange branch colors. Colored circles show full or partial presence of selected genes or metabolic pathways. Pyruvate fermentation refers to the presence of the full pathway for pyruvate fermentation to acetate and is differentiated based on the presence of the enzyme for acetyl-CoA generation from acetate, i.e., pyruvate dehydrogenase complex (PDC), or pyruvate ferredoxin oxidoreductase (PFO) together with phosphate acetyltransferase and acetate kinase, or acetate-CoA ligase. The arginine deiminase (ADI) pathway is only indicated if arginine deiminase, ornithine carbamoyltransferase, and carbamate kinase were found. Bar chart shows the completeness of the tricarboxylic acid cycle (TCA). Genes encoding nucleotide transport proteins (*ntt*); early upstream ORF (*euo*), transcriptional regulator of the chlamydial developmental cycle; histone-like developmental protein (*hctA*); serine/threonine protein kinase CopN (*copN*); pseudokinase Pkn5 (*pkn5*); glucose 6-phosphate transporter (*uhpC*); type IV secretion system (T4SS); type III secretion system (T3SS).

exemplified by the conserved serine/threonine protein kinase CopN (present in 99% of dereplicated genomes)[38] and the pseudokinase Pkn5 (93%)[39]. All known chlamydiae rely on host-derived metabolites[40]. Our analysis suggested that glucose-6-phosphate can likely be scavenged by all chlamydiae using the glucose 6-phosphate transporter UhpC (97%, only missing in one MAG and two draft genomes)[41]. In addition, the core genome includes a suite of nucleotide transport proteins (98%, Ntp1; 96%, Ntp2) to import ATP and other nucleotides[42–44]. Of note, the master regulator of the unique chlamydial developmental cycle, EUO, is highly conserved (99%)[45]. To a lesser degree, this is also the case for the histone-like protein HctA (83%), which facilitates the conversion of RBs to the EB stage[46] (Fig. 3). Taken together, the chlamydial core genome includes both hallmarks of a conserved developmental cycle and an host-associated lifestyle.

Zooming in from the phylum level to the family rank, we next set out to investigate the pangenome of selected chlamydial families. Calculating the core genome for families with at least three members (Supplementary Fig. 5a), we retrieve a median size of 599 NOGs per family, which is considerably larger than the phylum core genome. Furthermore, the presence-absence hierarchical clustering of core NOGs reflected the grouping of chlamydial families in our phylogenetic analysis, together indicating selection of family-level traits over extended evolutionary time periods (Fig. 3, Supplementary Fig. 5b). The fish pathogenic Parilichlamydiaceae have the smallest core genome with 415 NOGs. They also show the so far most reduced chlamydial genomes with estimated genome sizes < 1 Mb and a pronounced reduction in metabolic capacity (Fig. 3; Supplementary Fig. 5)[16]. In comparison, the protist-associated Parachlamydiaceae have a core genome of 727 NOGs (Fig. 3; Supplementary Fig. 5a), in line with their larger genomes and generally more complete metabolic capabilities[15,40]. Genes that do occur in only some but not all genomes of a group of organisms are together referred to as accessory genome, often comprising niche or organism-specific features[24]. Prominent examples of the chlamydial accessory genome are the patchy nucleotide and amino acid synthesis pathways, and the variations in the tricarboxylic acid cycle (TCA) observed in our dataset (Fig. 3, Supplementary discussion 1, Supplementary Data 9).

The relative contribution of core and accessory genes to the pangenome can provide insights into genome evolution[24]. Such analysis is, however, inherently prone to differences in sample size, i.e., the number of available genomes per family. We therefore included only families represented by at least ten genomes and chose to analyze the genomic fluidity parameter, which was shown to be robust to small sample sizes[47] (see Methods). The parameter measures the dissimilarity of genomes at the gene level within a taxonomic rank by averaging the dissimilarity of genomes within this group—denoted as mean φ, where 0 means highly similar and 1 indicates dissimilar genomes, respectively[47]. We focused our analysis on the family pangenomes of the human and animal pathogens in the Chlamydiaceae and the protist-associated Parachlamydiaceae. As expected for a highly specialized intracellular pathogen like *Chlamydia trachomatis*, the Chlamydiaceae showed a low genomic fluidity, as their genomes are highly similar at the gene level (mean $\varphi = 0.1$; 41% core genes). This is consistent with *Chlamydia trachomatis* having a closed species pangenome, indicating generally small population sizes and limited impact of horizontal gene transfer (HGT)[48]. The protist-associated Parachlamydiaceae, on the other hand, showed a significantly more open pangenome compared to the Chlamydiaceae (mean $\varphi = 0.5$; false discovery rate adjusted p value of t-test < 0.001; 6% core genes; Supplementary Fig. 6). This suggests that genome evolution of members of these environmental chlamydiae was characterized by larger population sizes and more interactions with other microbes, e.g., through a larger host spectrum and contact to other (facultative) intracellular microbes within their environmental hosts. This might have facilitated adaptive evolution through HGT, which is consistent with the concept of protists as "melting pots" for the evolution of intracellular bacteria[49,50]. In line with the Parachlamydiaceae, all other chlamydial families that could be included in this analysis also showed open pangenomes, suggesting a great genotypic and phenotypic plasticity across several chlamydial clades (Supplementary Fig. 6).

Clade-specific potential for inorganic carbon fixation and light-driven ATP synthesis. Some environmental chlamydiae encode features that deviate from the generally highly conserved biology of the majority of known chlamydiae. Among these are gene sets for a flagellar apparatus and a chemosensing system[15,18,19,51,52], a conjugative type IV secretion system[15,53], and the CRISPR-Cas phage defense system[54,55]. We recovered all of these features in our extended genome dataset and found support for a more widespread occurrence among different chlamydial lineages (Fig. 3; Supplementary discussion 2, Supplementary Fig. 7, Supplementary Data 10 and 11).

The MAGs from the GEM catalog added 45% novel gene families (NOGs) to our dereplicated and quality filtered dataset—gene content that has not been associated with chlamydiae before. Among these, an unexpected finding was the presence of key enzymes of the reductive tricarboxylic acid cycle (rTCA), a pathway for carbon fixation in microoxic and anaerobic microbes. We detected genes encoding ATP-citrate lyase (AclA and AclB) in MAGs of the MCF-D family from antarctic saline lakes (Fig. 3, Supplementary Data 9)[56]. Based on the AclA phylogeny, the chlamydial enzyme is related to ATP-citrate lyases from Epsilonproteobacteria and Aquificae (Supplementary Fig. 8). Host-associated microbial photo- or chemoautotrophic carbon fixation is important in many marine invertebrates[57,58], yet the chlamydial MAGs lack the full potential for photo- or

chemoautotrophy (i.e., the ferredoxin-dependent pyruvate synthase and the 2-oxoglutarate synthase). The partial rTCA in these chlamydiae might instead function in a similar fashion as in the pathogen *Mycobacterium tuberculosis*, which uses the pathway to maintain proton gradient and red-ox balance for short-term survival of hypoxia[59].

Previously unknown Parachlamydiaceae genomes revealed evidence for light-driven ATP synthesis in chlamydiae. A member of the genus *Neochlamydia* from a wastewater bioreactor and three novel MAGs of a genus from microbial mats from antarctic freshwater lakes[60,61] encoded a complete proteorhodopsin gene cluster including enzymes for synthesis of the light-harvesting co-factor retinal (Figs. 3 and 4a). Phylogenetic analysis suggests the independent acquisition of the gene set in two chlamydial lineages, indicating lineage-specific adaptations (Fig. 4b), which is consistent with this trait known to being frequently subject to HGT. Proteorhodopsins are commonly found in marine microbes in the sunlit (euphotic) zone and represent a major mechanism for light-driven ATP synthesis in these systems[62,63]. A marine *Vibrio* strain that gained proteorhodopsin through HGT showed increased long-term survival under resource-limited conditions[64]. It is therefore tempting to speculate that proteorhodopsin in chlamydiae may function as a maintenance mechanism for EBs, prolonging extracellular survival and increasing the chance to encounter new protist hosts. Alternatively, proteorhodopsin-driven energy generation might alleviate the host cell burden during intracellular replication. Taken together, these findings demonstrate that our understanding of chlamydial biology is far from complete, not only with respect to only recently recognized lineages but even for those environmental chlamydiae with cultured representatives.

Widespread anaerobic and molecular hydrogen metabolism among chlamydiae. Chlamydial metabolism has long been understood as aerobic or microaerobic using substrate-level phosphorylation in combination with oxidative phosphorylation[15,40,65]. Yet, recent metagenomic findings in marine deep sea sediments have uncovered a clade of chlamydiae with a specialized anaerobic lifestyle[19,23]. The Anoxychlamydiales (family SM23-39) have the genetic potential to carry out acetogenic fermentation and use the arginine deiminase (ADI) pathway to produce ATP[23]. Like other anaerobic microorganisms, these chlamydiae show an incomplete respiratory chain and a truncated TCA cycle[23] (Fig. 3). In order to investigate the prevalence of the potential for anaerobic substrate-level phosphorylation, we screened all chlamydial genomes for the presence of known anaerobic pathways and classified them using MetaCyc[66]. Of note, we discovered complete pyruvate fermentation to acetate in 43% of all chlamydial families investigated (9 out of 21; Figs. 3 and 5).

Like the Anoxychlamydiales, members of the family MCF-E have the genetic potential to convert pyruvate to acetyl-CoA using pyruvate ferredoxin oxidoreductase (PFO; Fig. 3, Supplementary Data 9). Phylogenetic analysis suggests that the PFO of these chlamydiae has been independently acquired through HGT. This is consistent with MCF-E members using acetate-CoA ligase[67] for ATP generation from acetyl-CoA as an alternative to phosphate acetyltransferase (Atp) and acetate kinase (AckA) employed in the Anoxychlamydiales.

The most prevalent pathway of acetogenic fermentation among chlamydiae, however, is acetyl-CoA generation from pyruvate via the pyruvate dehydrogenase complex (PDC), followed by ATP generation and acetate production through the concerted action of Atp and AckA. Genes for these key enzymes are found in representatives of seven families, including cultured members of the Criblamydiaceae, Simkaniaceae, and Parachlamydiaceae, as well as in a number of MAGs from the families CC-IV, MCF-A, MCF-D, and K940_chlam_3 (Figs. 3 and 5). Unlike the Anoxychlamydiales and MCF-E, pyruvate-fermenting chlamydiae

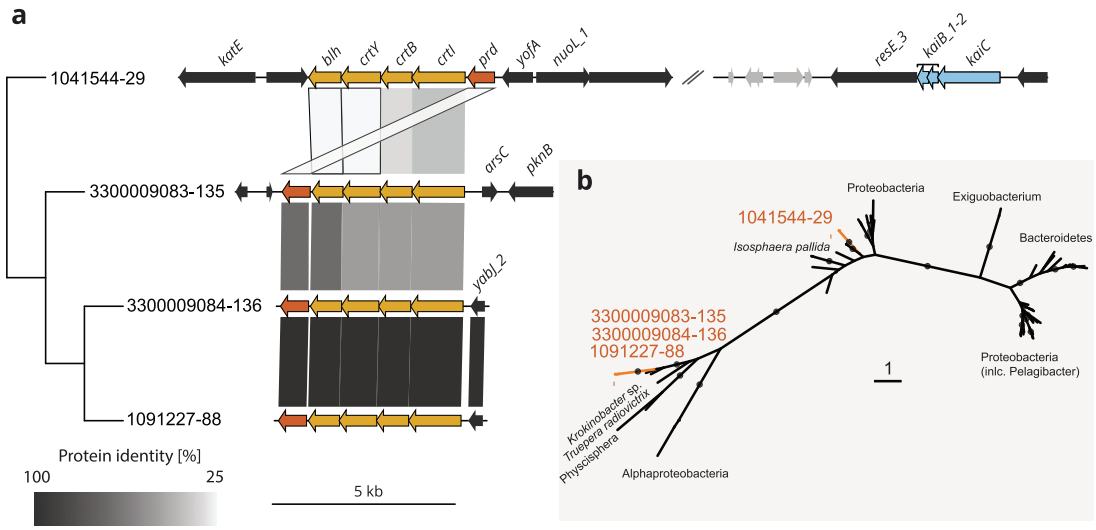


Fig. 4 Independent acquisition of light-driven ATP synthesis in two potentially amoeba-associated clades. a Gene synteny plot of proteorhodopsin related gene clusters in Parachlamydiaceae MAGs. Comparisons are ordered according to the phylogenomic species tree in Fig. 1. Arrows colored in orange, yellow, and blue represent proteorhodopsin (*prd*), carotene biosynthesis, and circadian clock genes, respectively. Black arrows indicate genes with chlamydial homologs. Bands connect homologs and are colored according to their protein identity. All other proteins of contigs encoding proteorhodopsin gene clusters were blasted against the NCBI non-redundant (nr) database to confirm the chlamydial origin of the contig. **b** Maximum likelihood phylogenetic tree of proteorhodopsin (Prd) (ENOG4105CSB) with chlamydial sequences showing two distinct clades. Maximum likelihood tree was inferred under LG + C30 + G + F model with 1000 improved ultrafast bootstraps and 1000 replicates of the SH-like approximate likelihood ratio test. Filled circles at nodes indicate a bootstrap support ≥ 95%. Scale bar indicates the number of substitutions per site.

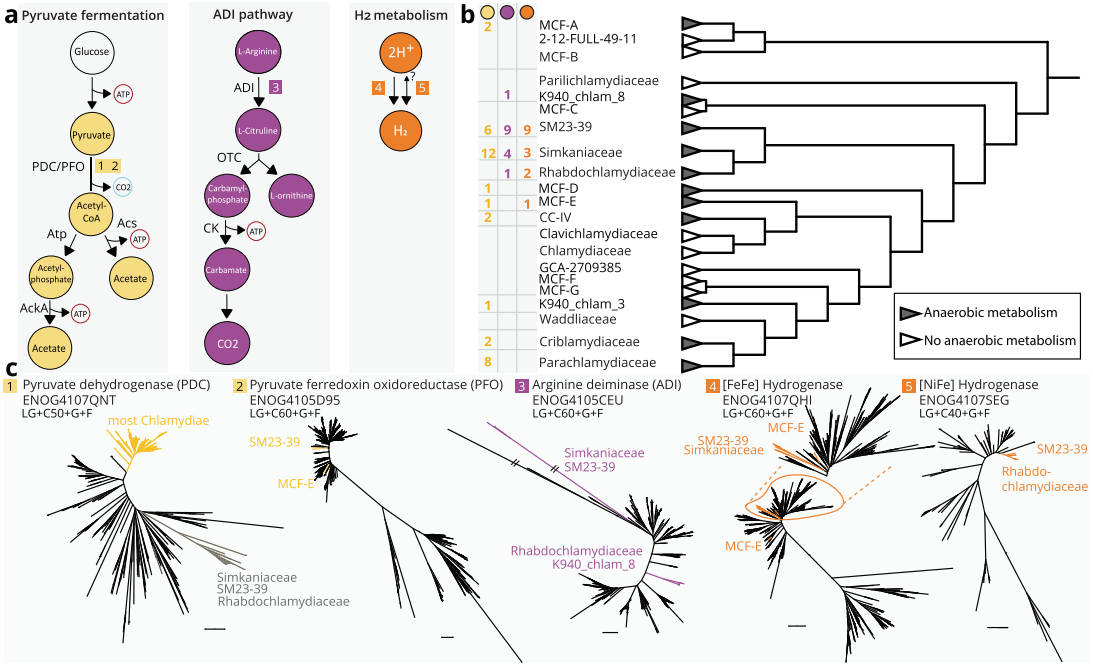


Fig. 5 Widespread fermentation pathways and molecular hydrogen production in chlamydiae. a Representation of putative anaerobic pathways for ATP generation and molecular hydrogen metabolism in chlamydiae. Labels next to enzymatic reactions indicate the associated enzymes. Numbers in squares correspond to phylogenetic trees in (**c**). Colors indicate affiliation with different pathways—pyruvate fermentation (yellow), ADI pathway (violet), hydrogen metabolism (orange). **b** Species tree of chlamydial representative genomes as in Fig. 3 collapsed at the family rank. Branch support values are based on 100 non-parametric bootstraps, support ≥ 70% is indicated as black circles. Box next to family names indicates the number of non-redundant genomes in a family with the respective color coded metabolic pathway. Pyruvate fermentation to acetate was only counted if genes encoding the complete pathway were present, i.e., pyruvate dehydrogenase complex (PDC), or pyruvate ferredoxin oxidoreductase (PFO) together with phosphate acetyltransferase (Atp) and acetate kinase (AckA), or acetate-CoA ligase (Acs). Likewise, arginine deaminase (ADI) pathway was only counted if genes encoding arginine deaminase (ADI), ornithine carbamoyltransferase (OTC), and carbamate kinase (CK) were present in a genome. **c** Unrooted maximum likelihood phylogenetic trees with best fit models numbered and colored according to (**a**) with 1000 optimized ultrafast bootstrap and 1000 SH-like approximate likelihood ratio test support. Best fit models per gene are indicated under the gene name and clades are named by family.

with PDC (Fig. 5c) also encode respiratory chain complex IV (cytochrome o and/or cytochrome bd; Fig. 3). This complex is typically associated with aerobic metabolism, but the additional presence of fermentation-related enzymes indicates a facultative anaerobic lifestyle[40]. Well-known protist hosts of environmental chlamydiae, such as *Acanthamoeba*, show a preference for low oxygen conditions and have specialized mitochondria for anaerobic ATP generation[68,69]. Chlamydial lineages infecting these protists may encounter anaerobic conditions, in which the ability to ferment could represent a selective advantage.

An alternative means for anaerobic ATP formation is the ADI pathway, in which arginine is converted to ornithine, ammonia, and carbon dioxide, generating ATP. We find the complete ADI pathway—indicated by the presence of ADI, ornithine carbamoyltransferase, and carbamate kinase—in four chlamydial families, including the Anoxychlamydiales (Figs. 3 and 5)[23]. Our phylogenetic analysis of the key enzyme arginine deaminase retrieved two distinct chlamydial clades (Fig. 5c). This suggests a polyphyletic origin of this gene among chlamydiae, which would be consistent with the ADI pathway being subject to frequent HGT[70].

In the Anoxychlamydiales fermentation is thought to be coupled to hydrogen production, a strategy to dump electrons in the absence of oxygen or alternative electron acceptors also used by other microbes[23,71]. To investigate the potential for hydrogen metabolism among all chlamydiae, we identified putative hydrogenases based on conserved protein domains and classified them with HydDB[72]. We found in total 40 hydrogenases in 34 genomes, classified as [FeFe] hydrogenases or [NiFe] hydrogenases, respectively (Supplementary Data 12).

[FeFe] hydrogenases previously described in Anoxychlamydiales MAGs are also present in the putative anoxic family MCF-E and a lineage of three Simkaniaceae MAGs associated with gutless marine oligochaetes of geographically distant origin (Figs. 3 and 5; Supplementary Data 12)[73]. All three chlamydial groups encode oxygen-sensitive trimeric [FeFe] hydrogenases to synergistically oxidize NADH and ferredoxin to produce molecular hydrogen[74] (Supplementary Fig. 9). While these hydrogenases are functionally similar, phylogenetic analysis recovers two separate monophyletic clades, suggesting they have been acquired independently (Fig. 5c). One additional [FeFe] hydrogenase is only present in one MCF-E member and is only distantly related to the two other clades (Fig. 5c, Supplementary Fig. 9).

Of note, we also identified oxygen-tolerant hydrogenases in chlamydial genomes. Type 3b [NiFe] hydrogenases are present in two members of the Rhabdochlamydiaceae, and one member of SM23-39. These cytosolic hydrogenases directly couple oxidation

of NADPH to hydrogen evolution but might also catalyze the reverse reaction[75] (Fig. 5a, reverse reaction annotated with a question mark). All chlamydial homologs are monophyletic and group with the methanotroph *Methylacidiphilum infernorum*, a member of the Verrucomicrobia[76] (Fig. 5c). Some obligate aerobic mycobacteria use these types of hydrogenases under low oxygen conditions when there is a lack of other terminal electron acceptors[71,77], suggesting a similar function in chlamydiae.

Molecular hydrogen metabolism is a widespread yet often poorly studied feature in pathogenic bacteria and protists, which is often critical for growth and virulence[71,77], not only for strict anaerobes such as *Clostridium perfringens*[78] or the parasite *Trichomonas vaginalis*[79], but also for the microaerophilic *Helicobacter pylori* and the facultative anaerobe *Campylobacter jejuni*[71,77].

In summary, our findings reveal surprisingly widespread traits of an anaerobic lifestyle among chlamydiae. This includes apparently strictly anaerobic lineages such as the Anoxychlamydiales and MCF-E, as well as putative facultative anaerobes in the Simkaniaceae, Rhabdochlamydiaceae, Criblamydiaceae, Parachlamydiaceae, and other families (Figs. 3 and 5b). The patchy distribution of fermentation pathways and hydrogenases indicates a complex scenario for the evolutionary relationship of chlamydiae with oxygen.

Family-specific habitat preferences. We next used our genome sequence dataset to investigate the abundance and distribution of chlamydiae in the environment. Of the chlamydial genomes in this study, 84 of 192 (32/82 MAGs of the GEM catalog) encode near full-length 16S rRNA genes ≥ 1300 nt, covering 15 of 21 chlamydial families with genome representatives. We used these sequences together with all publicly available near full-length 16S rRNA sequences and dereplicated the dataset at 99% sequence identity[80], yielding 310 chlamydial species representatives. Phylogenetic analysis confirmed the monophyly of all chlamydiae with high support (Fig. 6a), and the 16S rRNA gene tree corroborated the genome-based classification of chlamydial families (Fig. 1). While most sequences from the GEM catalog are part of chlamydial families identified earlier, sequences of the putative families MCF-A, MCF-B, MCF-D, MCF-E, and MCF-F represent yet unrecognized lineages in the 16S rRNA-based tree (Figs. 2 and 6a).

We queried all chlamydial 16S rRNA sequences for which a genome sequence is available against the integrated microbial next-generation sequencing (IMNGS) database[81] with 99% identity to estimate environmental species-level distribution and abundance. We obtained chlamydial amplicons matching the 16S rRNA gene of genomic representatives from 3,261 samples. Consistent with previous 16S rRNA meta-analysis, chlamydiae can be found in all major environments, as well as in a multitude of eukaryotic microbiomes[7,8]. Presence and relative abundance (RA) information was summarized at the family rank in order to investigate habitat preferences of the major lineages in our genome dataset. We obtained amplicon hits for 13 of the 15 families, only missing the fish pathogens Clavichlamydiaceae and Parilichlamydiaceae. Indeed, members of these families have only been found in fish gills so far[82,83], for which no public microbiome studies are available (https://www.imngs.org/; April 2020). This suggests that the Clavichlamydiaceae and Parilichlamydiaceae might be limited to these vertebrate hosts and not associated with microbial eukaryotes.

Unsurprisingly, Chlamydiaceae show a highly significant association with animal host-associated microbiomes (Fig. 6b). If present, members of the Chlamydiaceae reach RA values in the bacterial community of up to 79% in a variety of human and animal microbiomes[84–88] (Supplementary Fig. 10, Supplementary Data 13). For chlamydial families with cultured representatives in protists, we observe significant enrichment in soil (Parachlamydiaceae, Criblamydiaceae, Waddliaceae) and engineered environments (Parachlamydiaceae, Criblamydiaceae, Simkaniaceae), respectively (Fig. 6b), which is coherent with the origin of the majority of isolates from these families.

Families without cultured representatives on the other hand show significant enrichment in marine environments, including MCF-D (water column and sediment), MCF-E (water column), and K940_chlam_8 (sediment). This illustrates that these environments are still undersampled with respect to chlamydiae (Fig. 6b). So far, the Simkaniaceae members *Neptunochlamydia vexilliferae* and *Syngnamydia salmonis* are the only marine isolates available[89,90]. Accordingly, the Simkaniaceae are significantly enriched in marine environments. Even though some evidence for the clinical relevance of the third cultivated representative of this family, *Simkania negevensis*[91], has been reported[92], it is found with up to 1.7% RA in coral microbiomes[93] and at 0.5% RA in anaerobic digesters (Supplementary Fig. 10, Supplementary Data 13). This supports the existence of an environmental niche for *S. negevensis* and corroborates our finding of anaerobic metabolic potential for this and other members of the Simkaniaceae (Figs. 3 and 5).

For family SM23-39, which contains the anaerobic Anoxychlamydiales[17,19,23,94], the IMNGS query only yielded hits for those members that lack anaerobic pathways or appear to be facultative anaerobes (Fig. 3), and these are enriched in freshwater sediment environments (Fig. 6b, Supplementary Fig. 10, Supplementary Data 13). Owing to the lack of the 16S rRNA gene in all but one Anoxychlamydiales MAG, we could not further assess the environmental distribution of this group. However, the second anaerobic lineage, family MCF-E, is found in marine water column habitats and can reach up to 1% RA (Fig. 6b and 6c, Supplementary Fig. 10, Supplementary Data 13). All 16S rRNA gene sequences from this family originate from samples from Saanich Inlet, a seasonally anoxic fjord at the coast of Vancouver Island, British Columbia, Canada[95,96]. We related RA in samples containing MCF-E amplicons to oxygen concentration and sampling depth and observed the highest abundance of chlamydiae below the oxycline, i.e., in the deeper, anoxic layers of the water column (Fig. 6c). Potential hosts of these chlamydiae are microaerophilic or anaerobic protists, which are known to occur in the anaerobic water column and may, together with methanogenic endosymbiotic bacteria, be important for the biochemical cycling of methane[97].

In summary, the comprehensive analysis of chlamydial MAGs in this study provides novel insights into the genomic diversity of a bacterial phylum of strictly intracellular microbes, revealing a surprising variation with respect to their biology. Our analysis expands the known phylogenetic diversity of chlamydiae by 40%. We show that the chlamydial core genome comprises the toolbox for an host-associated intracellular lifestyle, while the accessory genome varies strongly in size between different families, reflecting adaptation to various environments and diverse hosts. We found evidence for light-driven ATP synthesis and key genes for the rTCA cycle in chlamydial organisms, and we show that members of several lineages have the genetic potential for anaerobic and hydrogen metabolism. Our genome-informed diversity survey revealed the presence of these chlamydiae in various anaerobic environments and provided further evidence for a ubiquitous occurrence of chlamydiae, sometimes at surprisingly high abundance. Targeted metagenomics and isolation approaches using diverse protist hosts will be important to further investigate those chlamydial groups that are only poorly

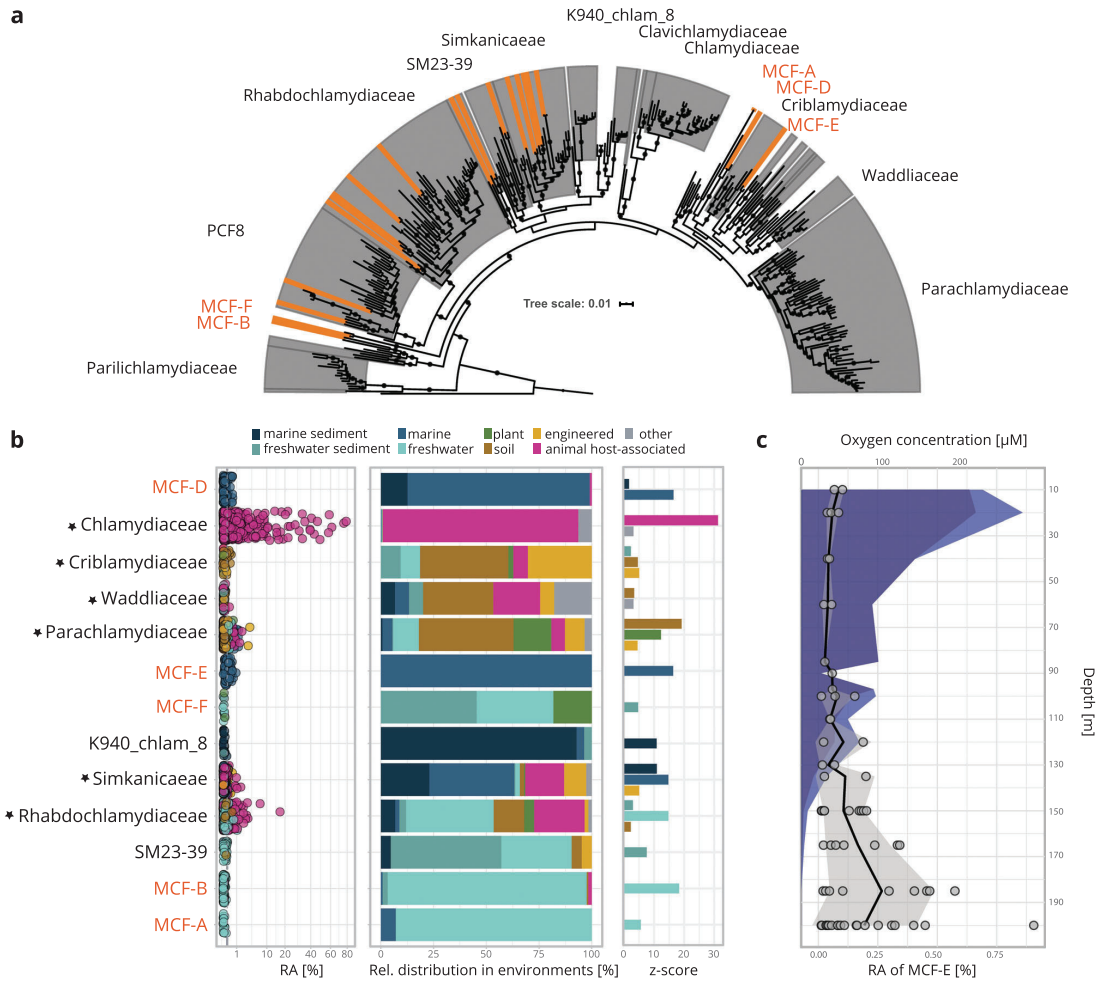


Fig. 6 Members of novel chlamydial families predominantly occur in freshwater and marine environments. a 16S rRNA gene maximum likelihood phylogenetic tree using near full-length sequences de-replicated at 99% under the SYM + R10 model inferred with IQTREE. Support was inferred from transfer bootstrap expectation (TBE) based on 100 non-parametric bootstraps. Circles at nodes indicate TBE support ≥70. The tree is pruned and does not include the outgroup. Chlamydial families are highlighted by gray background, 16S rRNA genes from novel MAGs in this study are indicated by orange shading. **b** All chlamydial full-length 16S rRNA genes in chlamydial genomes were used as a query against IMNGS with an identity cutoff of 1% to ensure species-specificity and summarized at the family level. Stars next to names indicate families with cultured representatives. Environmental categories "marine" and "freshwater" represent samples originating from the water column. The scatter plot on the left shows the relative abundance of chlamydial 16S rRNA gene amplicons. The bar plot in the middle shows the relative distribution of family members across diverse environments. The bar plot on the right indicates significant enrichment (adjusted *p* value ≤ 0.05) in environments based on one-sided Fisher's exact test with false discovery rate adjusted *p* values expressed as *z*-scores. **c** Relationship of relative abundance of the anaerobic family MCF-E with oxygen concentration and depth in samples from Saanich Inlet. *Y*-axis depicts depth in meters below surface, top *x*-axis indicates molarity of oxygen and bottom *x*-axis indicates relative abundance in percent of total 16S rRNA amplicons. Dark blue and light blue areas depict mean oxygen concentration and standard deviation, respectively. Gray filled points, black line, and gray area represent relative abundance in a sample, mean relative abundance, and standard deviation, respectively.

represented in our datasets so far. Ultimately, this will contribute to a better understanding of how an entire bacterial phylum that engaged in an intracellular lifestyle early on during evolution has emerged, and how niche specialization and adaptation to novel hosts have taken place.

Methods

Genome sequencing. The genome sequences of four Parachlamydiaceae symbionts of *Acanthamoeba* spp. isolated from fish gills in Thailand in 2014 have been determined in the context of this study. *Acanthamoeba* isolation and cultivation were carried out as described in Köstlbacher et al.[98]. Briefly, for isolation of genomic DNA amoeba cells were lysed, and host DNA was digested using 10 units DNase I (Thermo Fisher Scientific) at 37 °C for 30 min. DNase digestion was inactivated as recommended by the manufacturer. Bacterial DNA was purified using the DNeasy blood and tissue kit (Qiagen) as recommended by the manufacturer. Sequencing libraries were prepared using the Nextera XT kit (Illumina) and sequenced on an Illumina HiSeq 2000 platform. Trimming and quality control of reads were conducted with BBMap v35.43 (https://sourceforge.net/projects/bbmap/) (bbduk minlen = 50, qtrim = rl, trimq = 25, ktrim = r, k = 25, mink = 11, hdist = 1) and FastQC v0.11.4 (https://www.bioinformatics.babraham.ac.uk/

projects/fastqc/). Assemblies were performed with SPAdes v3.5.0[99], screened for contamination with CheckM[100], and annotated with prokka v1.14[101].

Dataset compilation and quality control. We used 38 MAGs from the published GEM catalog and added 44 MAGs from the GEM project that affiliated with the phylum chlamydiae[25]. Basic MAG features (size, GC content, N50 value, etc.) were calculated with QUAST v5.0.2[102]. Initial gene calling and annotation was performed with prokka v1.14[101] with the flags"—mincontiglen 200" to call genes only on contigs larger than 200 nt and "—gram neg" for usage of the gram negative database.

In addition to the 82 MAGs of the GEM catalog[25], we collected all publicly available chlamydial genomes ($n = 80$) on June 25, 2019 from NCBI Genbank and RefSeq, and we added the four Parachlamydiaceae draft genomes, one Rhabdochlamydiaceae MAG from a spider microbiome[103], one MAG from a metagenome of a marine worm from the genus *Xenoturbella*[104], and 24 MAGs from deep sea sediment samples[19] (Supplementary Data 2). As an outgroup, we added 15 non-chlamydial genome sequences of members of the PVC superphylum (Supplementary Data 2). We estimated completeness and contamination of all genomes with CheckM v1.1.2[100] using general bacterial marker genes with 'taxonomy_wf domain Bacteria'. We assigned the environmental origin of the genomes based on publicly available metadata. Organisms with known protist hosts were associated with the host environment.

MAG phylogeny and species tree reconstruction. For comprehensive phylogenomic analysis including all chlamydial MAGs, the protein sequences of 43 conserved marker proteins were extracted and aligned in CheckM v1.1.2 with the 'tree' workflow[100]. Model testing and maximum likelihood phylogenies were performed with IQ-TREE 1.6.2[105] under the empirical LG model[106]. The optimal model was determined with the "-m TESTNEW" procedure[107]. We added the empirical mixture models C10–C60 with the "-madd" option (Best model: C60 + LG + G + F)[108]. Support values were inferred from 1000 ultrafast bootstrap replicates[109] with the "-bnni" option for bootstrap tree optimization and from 1000 replicates of the SH-like approximate likelihood ratio test[110]. Trees were visualized and edited using the Interactive Tree Of Life v4[111]. We calculated phylogenetic diversity and phylogenetic gain for GEM chlamydiae MAGs in the context of the chlamydial species tree with the GenomeTreeTk v0.1.6 (https://github.com/dparks1134/GenomeTreeTk). We used the MAGs from the GEM catalog as the ingroup and all other chlamydiae were used as outgroup.

To establish a robust species tree, we removed redundancy and low quality MAGs by de-replicating all genomes at 99% ANI with dRep v1.4.3 using default parameters except for "—contamination 10" to remove highly contaminated MAGs. We retained 109 chlamydial and 15 outgroup genomes of the PVC superphylum for downstream analysis[112]. As described above, we calculated a maximum likelihood species tree with IQ-TREE 1.6.2 using the de-replicated dataset (Best model: C60 + LG + G + F). We used this tree as a guide tree for the posterior mean site frequency (PMSF) model[113] for improved site heterogeneity modeling under the C60 + LG + G + F model and inferred 100 non-parametric bootstraps "-b 100".

Taxonomy assignment. Taxonomy was assigned to all genomes with GTDB-Tk v0.3.3[33] using the 'classify_wf' option based on database release version 89 (https://data.ace.uq.edu.au/public/gtdb/data/releases/release89/). Taxonomic novelty for genus and species was inferred based on GTDB. The GTDB-Tk infers RED values for nodes by phylogenetically placing marker protein sequences into the reference species tree (Fig. 1). However, the accuracy of phylogenetic placement decreases with increasing phylogenetic distance[114]. To account for this, we enforced the additional rule that genomes had to be monophyletic (UF-bootstrap ≥ 95%) with the reference genomes at the family rank in the species tree (Fig. 1) in addition to the GTDB assignment. Due to paraphyly with the family rank representative GCA-270938, we therefore changed the taxonomic assignment of the monophyletic group of MAGs 1039677-28, 1039689-34, and 1039701-25 with RED values of 0.70–0.71 (GTDB family novelty below 0.77) to the family MCF-E.

To delineate genus rank clades, we calculated reciprocal best blast hit average amino acid identities (AAI) of chlamydial proteomes as described by Hausmann et al.[115]. We clustered genomes according to Konstantinidis et al.[116] into genus rank groups at the cutoff of 65% AAI (alignment fraction ≥ 35%). We visualized the density distribution of within family AAI between genomes with the "geom_density" function in the ggplot2 package[117] (Supplementary Fig. 2). Genus level clusters (AAI ≥ 95% and alignment fraction ≥ 35%) were illustrated with Cytoscape v3.7.0[118] (Supplementary Fig. 3). Accordingly, we separated species rank clades by calculating the whole genome ANI with FastANI v1.3[119] and clustering at the 95% ANI cutoff (alignment fraction ≥ 65%). We visualized species-level clusters (ANI ≥ 95% and alignment fraction ≥ 65%) with Cytoscape v3.7.0[118] (Supplementary Fig. 1). We named previously undetected families, genera, and species according to the MAG with the highest genome quality score (completeness - 5 × contamination) in the respective group. Taxonomic organization of chlamydiae on the family rank and above was performed in R with the ggraph package (https://cloud.r-project.org/package=ggraph).

Pangenome reconstruction. For pangenome reconstruction, we only considered the 96 genomes of the de-replicated dataset with an estimated completeness > 85% and contamination < 5%. To retrieve orthologous clusters we mapped all protein sequences against eggNOG v4.5.1[34] with emapper v1.0.1[120] against the bacterial database "-d bact" and proceeded using these NOGs. We performed an all-against-all blastp search of 45,717 (29.5% of all) unmapped proteins and clustered proteins based on hits with an E value < 0.001 with SiLiX[35] yielding 31,007 de novo NOGs (25,886 singletons). Combining eggNOG and de novo NOGs, the chlamydiae pangenome totaled at 37,380 NOGs (Supplementary Fig. 4). We calculated the chlamydial pangenome subcomponents with the following definitions: core—present in more than 90% of genomes; cloud—present in <15%; and shell—present in 15–90% of genomes[121]. The accessory genome is composed of the cloud and shell genome. We applied the same definitions for family-specific pangenome calculations (Supplementary Fig. 5). The exact numbers of gene families in the accessory genome are dependent on the clustering method and parameters used. However, the general trend of a pronounced difference between Chlamydiaceae and environmental representatives should be largely independent of the thresholds used.

We further analyzed pangenome features for chlamydial families with at least ten genome sequences (Supplementary Fig. 6) to ensure sufficient data points for resampling. We used the micropan[122] package in R version 3.5.1[123] genomic fluidity with the "fluidity" function using 100 simulations. We then tested whether the genomic fluidity of Chlamydiaceae is different from other environmental families in this analysis using two-sample t-test and corrected for false discoveries using the "p.adjust" function with the "BH"[124] method in R version 3.5.1[123].

Reconstruction of metabolic pathways, and identification of hydrogenases, defense, and secretion systems. We mapped all proteins to Kyoto Encyclopedia of Genes and Genomes (KEGG) orthologs (KOs) using GhostKOALA v2.2[125]. KO associated Enzyme Commission numbers (EC numbers) were used to reconstruct pathways of interest with MetaCyc[66] or KEGG (Supplementary Data 7). We identified conserved protein domains in all proteins and associated them to metabolic pathways and gene ontology terms using InterProScan v5.35-74.0[126] with the parameters "-dp—pathways—goterms" using hidden markov models from Pfam[127], TIGRFAM[128], and TMHMM[129] databases. Putative hydrogenases were identified based on conserved TIGRFAM (TIGR02512; [FeFe] hydrogenase, group A) or Pfam domains (PF00374; Nickel-dependent hydrogenase) and verified and classified using the web tool hydDB[72]. Gene synteny plots representing proteorhodopsin or [FeFe] hydrogenase gene clusters in chlamydiae genomes were visualized with genoplotR v0.8.9[130] (Supplementary Fig. 9). In addition, genomes were screened for the presence of secretion systems and CRISPR cas systems using MacSyFinder v1.0.5[131] with the "TXSScan" models[132] with "—db_type ordered_replicon all" and CRISPRCasFinder v2.0.2[133] with the parameters "-ccc 20000 -ccvRep -html -rcfowce -def S", respectively. We blasted all identified CRISPR spacers against the viral RefSeq database (https://blast.ncbi.nlm.nih.gov/Blast.cgi) on July 27, 2020.

Phylogenetic analysis of metabolic genes. For phylogenetic analysis of metabolic genes (including Supplementary Figs. 7 and 8) we downloaded the corresponding NOG protein sequences from eggNOG v4.5.1[34] and aligned them de novo using mafft v .427 "—maxiterate 1000—localpair". We trimmed the alignments using BMGE v1.12[134] using a gap rate of 0.2 "-g 0.2" and an entropy of 0.6 "-h 0.6". We calculated maximum likelihood phylogenetic trees with IQTREE v1.6.2[135] under the empirical LG model[106] using model testing "-m TESTNEW" including the empirical mixture models C10–C60[108] and "-seed 12345". Support values were obtained from 1000 ultrafast bootstraps with bootstrap tree optimization using "-bb 1000 -bnni"[105] and 1000 replicates of the SH-like approximate likelihood ratio test using "-alrt 1000"[110]. Trees were visualized and edited using the Interactive Tree Of Life v4[111].

16S rRNA gene phylogeny. All available unique near-full length 16S rRNA gene sequences of chlamydiae ($n = 233$) and other PVC members ($n = 205$) were downloaded from SILVA v138 SSU Ref NR 99[136]. An additional 79 near full-length chlamydial 16S rRNA gene sequences (97% identity OTU representatives) from Schulz et al.[29] were added to the dataset, in addition to 103 sequences from our reference genome dataset totaling 620 near full-length 16s rRNA sequences. Sequences were clustered at 99% sequence identity to reduce redundancy using USEARCH v11.0.667[137] with "-cluster_smallmem" resulting in 310 Chlamydiae and 198 non-chlamydial PVC members. We aligned the clustered sequences with SINA[138] and trimmed the alignment with trimAl "-gappyout"[139]. Model testing was performed with IQ-TREE 1.6.2[105] "-m TESTNEW" (Best model: SYM + R10), and initial support values were inferred from 100 non-parametric bootstraps using "-b 100". As Felsenstein's bootstrapping methods tend to yield very low support for large sequence datasets we additionally inferred transfer bootstrap expectation values based on the non-parametric bootstrap trees with booster (https://booster.pasteur.fr/; accessed in April 2020)[140].

Environmental distribution and abundance of chlamydiae. We queried all near full-length 16S rRNA gene sequences (≥1300 nt) present in MAGs from the GEM

catalog ($n = 46$) against the IMNGS database[81] (accessed March 5th, 2018), which systematically collects and preclusters amplicon studies deposited in the short read archive (SRA)[81]. We used a 99% identity cutoff to approximate retrieval of 16S rRNA gene amplicons at the species level to estimate the environmental distribution of chlamydial species with genome representatives. We accepted an SRA sample if at least three reads mapped to a chlamydial 16S rRNA gene query sequence. We classified SRA samples mirroring the IMG/M environmental nomenclature using SRA metadata (https://www.ncbi.nlm.nih.gov/sra)[28]. We tested for overrepresentation of chlamydial families in environments using Fisher's exact test with the "fisher.test" ("stats" package) with "alternative = greater" in R version 3.5.1[123] and corrected p values with "BH"[124] using the R base package function "p.adjust". We considered p values ≤ 0.05 as significant and transformed them into z-scores using the "qnorm" function in the stats package.

Statistical analysis. All statistical tests and data analysis were performed in R version 3.5.1[123].

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All metagenomic data, bins and annotations are available through the IMG/M portal (https://img.jgi.doe.gov/). Metagenome-assembled genome sequences from the Genomes from Earth's Microbiomes initiative[25] are available at https://genome.jgi.doe.gov/GEMs and https://portal.nersc.gov/GEM. Small subunit rRNA gene data used in this study are available via the SILVA database (https://www.arb-silva.de/)[136] and IMNGS database (https://www.imngs.org/)[81]. Metadata for data used from the IMNGS database can be accessed via the Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra)[28]. The collection of MAGs and proteomes used in this study, mapping files (pangenome NOGs, KEGG and Interpro), trimmed alignment files, and tree files are available at zenodo (https://doi.org/10.5281/zenodo.4318714). Accession numbers for reference genomes are available in Supplementary Table 2. Additional genome sequences generated in this study have been deposited in GenBank under the accession numbers JAEMUB000000000, JAEMUC000000000, JAEMUD000000000, and JAEMUE000000000.

Received: 4 February 2021; Accepted: 10 June 2021;
Published online: 29 June 2021

References

1. McFall-Ngai, M. et al. Animals in a bacterial world, a new imperative for the life sciences. *Proc. Natl Acad. Sci.* **110**, 3229–3236 (2013).
2. Horn, M. Chlamydiae as symbionts in eukaryotes. *Annu. Rev. Microbiol.* **62**, 113–131 (2008).
3. Taylor-Brown, A., Vaughan, L., Greub, G., Timms, P. & Polkinghorne, A. Twenty years of research into Chlamydia-like organisms: a revolution in our understanding of the biology and pathogenicity of members of the phylum Chlamydiae. *Pathog. Dis.* **73**, 1–15 (2015).
4. Wagner, M. & Horn, M. The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance. *Curr. Opin. Biotechnol.* **17**, 241–249 (2006).
5. Rivas-Marín, E. & Devos, D. P. The Paradigms They Are a-Changin': past, present and future of PVC bacteria research. *Antonie van. Leeuwenhoek* **111**, 785–799 (2018).
6. Elwell, C., Mirrashidi, K. & Engel, J. Chlamydia cell biology and pathogenesis. *Nat. Rev. Microbiol.* **14**, 385–400 (2016).
7. Collingro, A., Köstlbacher, S. & Horn, M. Chlamydiae in the Environment. *Trends Microbiol.* **28**, 877–888 (2020).
8. Lagkouvardos, I. et al. Integrating metagenomic and amplicon databases to resolve the phylogenetic and ecological diversity of the Chlamydiae. *ISME J.* **8**, 115–125 (2014).
9. Greub, G. & Raoult, D. Microorganisms resistant to free-living amoebae. *Clin. Microbiol. Rev.* **17**, 413–433 (2004).
10. Bar-On, Y. M., Phillips, R. & Milo, R. The biomass distribution on Earth. *Proc. Natl Acad. Sci.* **115**, 6506–6511 (2018).
11. Taylor-Brown, A., Madden, D. & Polkinghorne, A. Culture-independent approaches to chlamydial genomics. *Micro. Genom.* **4**, e000145 (2018).
12. Sixt, B. S. & Valdivia, R. H. Molecular Genetic Analysis of Chlamydia Species. *Annu. Rev. Microbiol.* **70**, 179–198 (2016).
13. Bachmann, N. L., Polkinghorne, A. & Timms, P. Chlamydia genomics: providing novel insights into chlamydial biology. *Trends Microbiol.* **22**, 464–472 (2014).
14. Subtil, A. & Dautry-Varsat, A. Chlamydia: five years A.G. (after genome). *Curr. Opin. Microbiol.* **7**, 85–92 (2004).
15. Collingro, A. et al. Unity in Variety—The Pan-Genome of the Chlamydiae. *Mol. Biol. Evol.* **28**, 3253–3270 (2011).
16. Taylor-Brown, A. et al. Metagenomic Analysis of Fish-Associated Ca. Parilichlamydiaceae Reveals Striking Metabolic Similarities to the Terrestrial Chlamydiaceae. *Genom. Biol. Evol.* **10**, 2587–2595 (2018).
17. Baker, B. J., Lazar, C. S., Teske, A. P. & Dick, G. J. Genomic resolution of linkages in carbon, nitrogen, and sulfur cycling among widespread estuary sediment bacteria. *Microbiome* **3**, 14 (2015).
18. Collingro, A. et al. Unexpected genomic features in widespread intracellular bacteria: evidence for motility of marine chlamydiae. *ISME J.* **11**, 2334–2344 (2017).
19. Dharamshi, J. E. et al. Marine Sediments Illuminate Chlamydiae Diversity and Evolution. *Curr. Biol.* **30**, 1032–1048.e7 (2020).
20. Pillonel, T., Bertelli, C. & Greub, G. Environmental Metagenomic Assemblies Reveal Seven New Highly Divergent Chlamydial Lineages and Hallmarks of a Conserved Intracellular Lifestyle. *Front. Microbiol.* **9**, 79 (2018).
21. Taylor-Brown, A., Bachmann, N. L., Borel, N. & Polkinghorne, A. Culture-independent genomic characterisation of Candidatus Chlamydia sanzinia, a novel uncultivated bacterium infecting snakes. *BMC Genom.* **17**, 710 (2016).
22. Taylor-Brown, A. et al. Culture-independent genomics of a novel chlamydial pathogen of fish provides new insight into host-specific adaptations utilized by these intracellular bacteria. *Environ. Microbiol.* **19**, 1899–1913 (2017).
23. Stairs, C. W. et al. Chlamydial contribution to anaerobic metabolism during eukaryotic evolution. *Sci. Adv.* **6**, eabb7258 (2020).
24. Brockhurst, M. A. et al. The Ecology and Evolution of Pangenomes. *Curr. Biol.* **29**, R1094–R1103 (2019).
25. Nayfach, S. et al. A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 499–509 (2020).
26. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
27. Wu, D. et al. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462**, 1056–1060 (2009).
28. Chen, I. A. et al. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.* **47**, D666-D677 (2019).
29. Schulz, F. et al. Towards a balanced view of the bacterial tree of life. *Microbiome* **5**, 140 (2017).
30. Subtil, A., Collingro, A. & Horn, M. Tracing the primordial Chlamydiae: extinct parasites of plants? *Trends Plant Sci.* **19**, 36–43 (2014).
31. Cenci, U. et al. Biotic Host-Pathogen Interactions As Major Drivers of Plastid Endosymbiosis. *Trends Plant Sci.* **22**, 316–328 (2017).
32. Blair, P. M. et al. Exploration of the Biosynthetic Potential of the *Populus* Microbiome. *mSystems* **3**, e00045-18 (2018).
33. Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
34. Huerta-Cepas, J. et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44**, D286–D293 (2016).
35. Miele, V., Penel, S. & Duret, L. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinforma.* **12**, 116 (2011).
36. Abby, S. S. & Rocha, E. P. C. The non-flagellar type III secretion system evolved from the bacterial flagellum and diversified into host-cell adapted systems. *PLoS Genet.* **8**, e1002983 (2012).
37. Peters, J., Wilson, D. P., Myers, G., Timms, P. & Bavoil, P. M. Type III secretion à la Chlamydia. *Trends Microbiol.* **15**, 241–251 (2007).
38. Archuleta, T. L. et al. The Chlamydia effector chlamydial outer protein N (CopN) sequesters tubulin and prevents microtubule assembly. *J. Biol. Chem.* **286**, 33992–33998 (2011).
39. Verma, A. & Maurelli, A. T. Identification of two eukaryote-like serine/threonine kinases encoded by Chlamydia trachomatis serovar L2 and characterization of interacting partners of Pkn1. *Infect. Immun.* **71**, 5772–5784 (2003).
40. Omsland, A., Sixt, B. S., Horn, M. & Hackstadt, T. Chlamydial metabolism revisited: interspecies metabolic variability and developmental stage-specific physiologic activities. *FEMS Microbiol. Rev.* **38**, 779–801 (2014).
41. Schwöppe, C., Winkler, H. H. & Neuhaus, H. E. Properties of the glucose-6-phosphate transporter from Chlamydia pneumoniae (HPTcp) and the glucose-6-phosphate sensor from Escherichia coli (UhpC). *J. Bacteriol.* **184**, 2108–2115 (2002).
42. Tjaden, J. et al. Two nucleotide transport proteins in Chlamydia trachomatis, one for net nucleoside triphosphate uptake and the other for transport of energy. *J. Bacteriol.* **181**, 1196–1202 (1999).
43. Schmitz-Esser, S. et al. ATP/ADP translocases: a common feature of obligate intracellular amoebal symbionts related to Chlamydiae and Rickettsiae. *J. Bacteriol.* **186**, 683–691 (2004).

44. Haferkamp, I. et al. Tapping the nucleotide pool of the host: novel nucleotide carrier proteins of Protochlamydia amoebophila. *Mol. Microbiol.* **60**, 1534–1545 (2006).
45. Rosario, C. J. & Tan, M. The early gene product EUO is a transcriptional repressor that selectively regulates promoters of Chlamydia late genes. *Mol. Microbiol.* **84**, 1097–1107 (2012).
46. Belland, R. J. et al. Genomic transcriptional profiling of the developmental cycle of Chlamydia trachomatis. *Proc. Natl Acad. Sci. U. S. A.* **100**, 8478–8483 (2003).
47. Kislyuk, A. O., Haegeman, B., Bergman, N. H. & Weitz, J. S. Genomic fluidity: an integrative view of gene diversity within microbial populations. *BMC Genom.* **12**, 32 (2011).
48. McInerney, J. O., McNally, A. & O'Connell, M. J. Why prokaryotes have pangenomes. *Nat. Microbiol.* **2**, 17040 (2017).
49. Wang, Z. & Wu, M. Comparative Genomic Analysis of Acanthamoeba Endosymbionts Highlights the Role of Amoebae as a 'Melting Pot' Shaping the Rickettsiales Evolution. *Genom. Biol. Evol.* **9**, 3214–3224 (2017).
50. Moliner, C., Fournier, P.-E. & Raoult, D. Genome analysis of microorganisms living in amoebae reveals a melting pot of evolution. *FEMS Microbiol. Rev.* **34**, 281–294 (2010).
51. Bertelli, C. et al. Sequencing and characterizing the genome of Estrella lausannensis as an undergraduate project: training students and biological insights. *Front. Microbiol.* **6**, 101 (2015).
52. Bertelli, C., Goesmann, A. & Greub, G. Criblamydia sequanensis Harbors a Megaplasmid Encoding Arsenite Resistance. *Genom. Announc.* **2**, e00949–14 (2014).
53. Köstlbacher, S., Collingro, A., Halter, T., Domman, D. & Horn, M. Coevolving Plasmids Drive Gene Flow and Genome Plasticity in Host-Associated Intracellular Bacteria. *Curr. Biol.* **31**, 346–357.e3 (2021).
54. Bertelli, C. et al. CRISPR System Acquisition and Evolution of an Obligate IntracellularChlamydia-Related Bacterium. *Genom. Biol. Evol.* **8**, 2376–2386 (2016).
55. Benamar, S. et al. Developmental Cycle and Genome Analysis of Protochlamydia massiliensis sp. nov. a New Species in the Parachlamydiacae Family. *Front. Cell. Infect. Microbiol.* **7**, 385 (2017).
56. Panwar, P. et al. Influence of the polar light cycle on seasonal dynamics of an Antarctic lake microbial community. *Microbiome* **8**, 116 (2020).
57. Venn, A. A., Loram, J. E. & Douglas, A. E. Photosynthetic symbioses in animals. *J. Exp. Bot.* **59**, 1069–1080 (2008).
58. Cavanaugh, C. M. Symbiotic chemoautotrophic bacteria in marine invertebrates from sulphide-rich habitats. *Nature* **302**, 58–61 (1983).
59. Hu, J., Jin, K., He, Z.-G. & Zhang, H. Citrate lyase CitE in Mycobacterium tuberculosis contributes to mycobacterial survival under hypoxic conditions. *PLoS ONE* **15**, e0230786 (2020).
60. Kantor, R. S. et al. Genome-Resolved Meta-Omics Ties Microbial Dynamics to Process Performance in Biotechnology for Thiocyanate Degradation. *Environ. Sci. Technol.* **51**, 2944–2953 (2017).
61. Wang, Z. et al. A new method for rapid genome classification, clustering, visualization, and novel taxa discovery from metagenome. https://doi.org/10.1101/812917.
62. Sabehi, G. et al. New insights into metabolic properties of marine bacteria encoding proteorhodopsins. *PLoS Biol.* **3**, e273 (2005).
63. Croitoru, K. Faculty Opinions recommendation of Environmental genome shotgun sequencing of the Sargasso Sea. *Faculty Opin.—Post-Publ. Peer Rev. Biomed. Lit.* (2014). https://doi.org/10.3410/f.1017813.793496370.
64. Gómez-Consarnau, L. et al. Proteorhodopsin phototrophy promotes survival of marine bacteria during starvation. *PLoS Biol.* **8**, e1000358 (2010).
65. Omsland, A., Sager, J., Nair, V., Sturdevant, D. E. & Hackstadt, T. Developmental stage-specific metabolic and transcriptional activity of Chlamydia trachomatis in an axenic medium. *Proc. Natl Acad. Sci.* **109**, 19781–19785 (2012).
66. Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes—a 2019 update. *Nucleic Acids Res.* **48**, D445–D453 (2020).
67. Glasemacher, J., Bock, A. K., Schmid, R. & Schønheit, P. Purification and Properties of acetyl-CoA Synthetase (ADP-forming), an Archaeal Enzyme of Acetate Formation and ATP Synthesis, From the Hyperthermophile Pyrococcus Furiosus. *Eur. J. Biochem.* **244**, 561–567 (1997).
68. Stairs, C. W., Leger, M. M. & Roger, A. J. Diversity and origins of anaerobic metabolism in mitochondria and related organelles. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140326 (2015).
69. Leger, M. M., Gawryluk, R. M. R., Gray, M. W. & Roger, A. J. Evidence for a hydrogenosomal-type anaerobic ATP generation pathway in Acanthamoeba castellanii. *PLoS ONE* **8**, e69532 (2013).
70. Novák, L. et al. Arginine deiminase pathway enzymes: evolutionary history in metamonads and other eukaryotes. *BMC Evol. Biol.* **16**, 197 (2016).
71. Benoit, S. L., Maier, R. J., Sawers, R. G. & Greening, C. Molecular Hydrogen Metabolism: a Widespread Trait of Pathogenic Bacteria and Protists. *Microbiol. Mol. Biol. Rev.* **84**, e00092–19 (2020).
72. Søndergaard, D., Pedersen, C. N. S. & Greening, C. HydDB: a web tool for hydrogenase classification and analysis. *Sci. Rep.* **6**, 34212 (2016).
73. Kleiner, M. et al. Metaproteomics of a gutless marine worm and its symbiotic microbial community reveal unusual pathways for carbon and energy use. *Proc. Natl Acad. Sci. U. S. A.* **109**, E1173–E1182 (2012).
74. Schut, G. J. & Adams, M. W. W. The iron-hydrogenase of Thermotoga maritima utilizes ferredoxin and NADH synergistically: a new perspective on anaerobic hydrogen production. *J. Bacteriol.* **191**, 4451–4457 (2009).
75. Greening, C. et al. Genomic and metagenomic surveys of hydrogenase distribution indicate H_2 is a widely utilised energy source for microbial growth and survival. *ISME J.* **10**, 761–777 (2016).
76. Hou, S. et al. Complete genome sequence of the extremely acidophilic methanotroph isolate V4, Methylacidiphilum infernorum, a representative of the bacterial phylum Verrucomicrobia. *Biol. Direct* **3**, 26 (2008).
77. Berney, M., Greening, C., Conrad, R., Jacobs, W. R. Jr & Cook, G. M. An obligately aerobic soil bacterium activates fermentative hydrogen production to survive reductive stress during hypoxia. *Proc. Natl Acad. Sci. U. S. A.* **111**, 11479–11484 (2014).
78. Kaji, M. et al. The hydA gene encoding the H(2)-evolving hydrogenase of Clostridium perfringens: molecular characterization and expression of the gene. *FEMS Microbiol. Lett.* **181**, 329–336 (1999).
79. Lindmark, D. G., Muller, M. & Shio, H. Hydrogenosomes in Trichomonas vaginalis. *J. Parasitol.* **61**, 552 (1975).
80. Edgar, R. C. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* **34**, 2371–2375 (2018).
81. Lagkouvardos, I. et al. IMNGS: A comprehensive open resource of processed 16S rRNA microbial profiles for ecology and diversity studies. *Sci. Rep.* **6**, 33721 (2016).
82. Stride, M. C. et al. Molecular characterization of 'Candidatus Parilichlamydia carangidicola,' a novel Chlamydia-like epitheliocystis agent in yellowtail kingfish, Seriola lalandi (Valenciennes), and the proposal of a new family, 'Candidatus Parilichlamydiaceae' fam. nov. (order Chlamydiales). *Appl. Environ. Microbiol.* **79**, 1590–1597 (2013).
83. Draghi, A. et al. Characterization of 'Candidatus Piscichlamydia salmonis' (Order Chlamydiales), a Chlamydia-Like Bacterium Associated With Epitheliocystis in Farmed Atlantic Salmon (Salmo salar). *J. Clin. Microbiol.* **42**, 5286–5297 (2004).
84. Neuendorf, E. et al. Chlamydia caviae infection alters abundance but not composition of the guinea pig vaginal microbiota. *Pathog. Dis.* **73**, ftv019 (2015).
85. Kelly, J. et al. Composition and diversity of mucosa-associated microbiota along the entire length of the pig gastrointestinal tract; dietary influences. *Environ. Microbiol.* **19**, 1425–1438 (2017).
86. Kelly, M. S. et al. The Nasopharyngeal Microbiota of Children With Respiratory Infections in Botswana. *Pediatr. Infect. Dis. J.* **36**, e211–e218 (2017).
87. Liechty, E. R. et al. The levonorgestrel-releasing intrauterine system is associated with delayed endocervical clearance of Chlamydia trachomatis without alterations in vaginal microbiota. *Pathog. Dis.* **73**, ftv070 (2015).
88. Ganz, H. H. et al. Community-Level Differences in the Microbiome of Healthy Wild Mallards and Those Infected by Influenza A Viruses. *mSystems* **2**, e00188–16 (2017).
89. Pizzetti, I. et al. Chlamydial seasonal dynamics and isolation of 'Candidatus Neptunochlamydia vexilliferae' from a Tyrrhenian coastal lake. *Environ. Microbiol.* **18**, 2405–2417 (2016).
90. Nylund, A. et al. Genotyping of Candidatus Syngnamydia salmonis (chlamydiales; Simkaniaceae) co-cultured in Paramoeba perurans (amoebozoa; Paramoebidae). *Arch. Microbiol.* **200**, 859–867 (2018).
91. Kahane, S., Gonen, R., Sayada, C., Elion, J. & Friedman, M. G. Description and partial characterization of a new Chlamydia-like microorganism. *FEMS Microbiol. Lett.* **109**, 329–333 (1993).
92. Vouga, M., Baud, D. & Greub, G. Simkania negevensis, an insight into the biology and clinical importance of a novel member of the Chlamydiales order. *Crit. Rev. Microbiol.* **43**, 62–80 (2017).
93. Ziegler, M. et al. Coral bacterial community structure responds to environmental change in a host-specific manner. *Nat. Commun.* **10**, 3092 (2019).
94. Anantharaman, K. et al. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* **7**, 13219 (2016).
95. Torres-Beltrán, M. et al. A compendium of geochemical information from the Saanich Inlet water column. *Sci. Data* **4**, 170159 (2017).
96. Hawley, A. K. et al. A compendium of multi-omic sequence information from the Saanich Inlet water column. *Sci. Data* **4**, 170160 (2017).
97. Orsi, W., Song, Y. C., Hallam, S. & Edgcomb, V. Effect of oxygen minimum zone formation on communities of marine protists. *ISME J.* **6**, 1586–1601 (2012).

98. Köstlbacher, S. et al. Draft Genome Sequences of Bacterium STE3 and sp. Strain AcF84. *Endosymbionts spp. Microbiol. Resour. Announc.* **9**, e00220–e00220 (2020).
99. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
100. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genom. Res.* **25**, 1043–1055 (2015).
101. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
102. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
103. Hendrickx, F. et al. A masculinizing supergene underlies an exaggerated male reproductive morph in a spider. https://doi.org/10.1101/2021.02.09.430505.
104. Philippe, H. et al. Mitigating Anticipated Effects of Systematic Errors Supports Sister-Group Relationship between Xenacoelomorpha and Ambulacraria. *Curr. Biol.* **29**, 1818–1826.e6 (2019).
105. Minh, B. Q., Nguyen, M. A. T. & von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* **30**, 1188–1195 (2013).
106. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008).
107. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
108. Quang, L. S., Gascuel, O. & Lartillot, N. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* **24**, 2317–2323 (2008).
109. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
110. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
111. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
112. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
113. Wang, H.-C., Minh, B. Q., Susko, E. & Roger, A. J. Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. *Syst. Biol.* **67**, 216–235 (2018).
114. Matsen, F. A., Kodner, R. B. & Armbrust, E. V. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinforma.* **11**, 538 (2010).
115. Hausmann, B. et al. Peatland Acidobacteria with a dissimilatory sulfur metabolism. *ISME J.* **12**, 1729–1742 (2018).
116. Konstantinidis, K. T., Rosselló-Móra, R. & Amann, R. Uncultivated microbes in need of their own taxonomy. *ISME J.* **11**, 2399–2406 (2017).
117. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer, 2016).
118. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genom. Res.* **13**, 2498–2504 (2003).
119. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).
120. Huerta-Cepas, J. et al. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
121. Maistrenko, O. M. et al. Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. *ISME J.* **14**, 1247–1259 (2020).
122. Snipen, L. & Liland, K. H. micropan: an R-package for microbial pan-genomics. *BMC Bioinforma.* **16**, 79 (2015).
123. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (2020).
124. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc.: Ser. B (Methodol.)* **57**, 289–300 (1995).
125. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J. Mol. Biol.* **428**, 726–731 (2016).
126. Quevillon, E. et al. InterProScan: protein domains identifier. *Nucleic Acids Res.* **33**, W116–W120 (2005).
127. El-Gebali, S. et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
128. Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371–373 (2003).
129. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).
130. Guy, L., Kultima, J. R. & Andersson, S. G. E. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26**, 2334–2335 (2010).
131. Abby, S. S. & Rocha, E. P. C. Identification of Protein Secretion Systems in Bacterial Genomes Using MacSyFinder. *Methods Mol. Biol.* **1615**, 1–21 (2017).
132. Abby, S. S. et al. Identification of protein secretion systems in bacterial genomes. *Sci. Rep.* **6**, 1–14 (2016).
133. Couvin, D. et al. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res.* **46**, W246–W251 (2018).
134. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).
135. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
136. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2012).
137. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
138. Pruesse, E., Peplies, J. & Glöckner, F. O. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**, 1823–1829 (2012).
139. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
140. Lemoine, F. et al. Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature* **556**, 452–456 (2018).

Acknowledgements

We want to thank the IMG/M Data Consortium for contributing metagenomic data. We thank Craig Herbold for discussion on pangenome and taxonomy analysis, Chaturong Putaporntip and Somchai Jongwutiwes for providing amoeba isolates for genome sequencing of chlamydial symbionts, Daryl Domman for genome assembly, and Chris Greening for feedback concerning the molecular hydrogen metabolism. We would like to thank Jennah Dharamshi, Thijs Ettema, and Frederik Hendrickx for early access to data from ongoing projects. The Life Science Compute Cluster (LiSC; http://cube.univie.ac.at/lisc) was used for computational analysis.

Author contributions

S.K. and M.H. conceptualized this study. F.S. and S.P.J. performed metagenome data mining. S.K. performed taxonomic, phylogenetic, pangenome, and 16S rRNA analyses. S.K., A.C., and T.H. performed comparative genomic analyses. S.K., A.C., T.H., and M.H. interpreted the results. All authors wrote and edited the paper.

Funding

This project has received funding from the European Research Council ERC (EVO-CHLAMY, grant no. 281633 to M.H.), and the Austrian Science Fund FWF (FunChlam, grant no. P32112 to A.C.; and doc.funds program DOC 69-B). Parts of this study were performed by the US Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, under Contract No. DE-AC02–05CH11231 and made use of resources of the National Energy Research Scientific Computing Center.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-24294-3.

Correspondence and requests for materials should be addressed to M.H.

Peer review information *Nature Communications* thanks Anders Andersson, Rolf Daniel, and Justin North for their contribution to the peer review of this work. Peer review reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

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

SUPPLEMENTARY INFORMATION

Pangenomics reveals alternative environmental lifestyles among chlamydiae

Stephan Köstlbacher[1], Astrid Collingro[1], Tamara Halter[1], Frederik Schulz[2], Sean P. Jungbluth[2], and Matthias Horn[1]*

[1] Centre for Microbiology and Environmental Systems Science, University of Vienna, Vienna, Austria

[2] DOE Joint Genome Institute, Berkeley, CA, USA

[3] Current address: Laboratory of Microbiology, Wageningen University and Research, Wageningen, The Netherlands

* Corresponding author: Matthias Horn, matthias.horn@univie.ac.at

Supplementary discussion 1: Patchy nucleotide and TCA metabolism as hallmarks of the accessory genome

Despite a general dependency of all chlamydiae on host resources, previous analysis suggested that environmental representatives retained more complete sets of central metabolic pathways than the pathogenic Chlamydiaceae[1–4]. This trend prevails in our extended dataset comprising a large number of chlamydiae from diverse environments. Consistent with the presence of nucleotide transport proteins for the uptake of ribonucleotides in all chlamydiae, only few genes involved in nucleotide synthesis are part of the phylum core genome. This includes for example the ribonucleotide reductase that catalyzes the reduction of ribonucleotide to deoxyribonucleotides (Figure 3, Supplementary Data 9). Yet, several chlamydiae including MAGs of aquatic origin encode complete *de novo* synthesis pathways for purines or pyrimidines[5–8] (Figure 3, Supplementary Data 9). We observed a similarly patchy distribution for genes functioning in amino acid synthesis, exemplified by the tryptophan synthesis pathway (Figure 3, Supplementary Data 9).

A reduced tricarboxylic acid (TCA) cycle is a hallmark of the Chlamydiaceae[1], which in these pathogens is supplemented by host derived intermediates during intracellular growth to produce biomass[9]. Many environmental chlamydiae encode a complete TCA cycle, but the pathway is truncated or nearly absent in the amoeba-associated *Neochlamydia* species, in members of the Anoxychlamydiales (a clade in the family SM23-39)[5,10], and completely absent in the fish pathogens *Clavichlamydia salmonicola* (Collingro et al, in preparation) and the Parilichlamydiaceae[11] (Figure 3). We found the entire TCA gene set in several environmental MAGs but observed a lack of the majority of the TCA cycle genes in the family MCF-C, and in novel MAGs of the Parachlamydiaceae and Simkaniaceae, several of which were retrieved from marine gutless worms or amoeba isolated from fish gills (Figure 3, Supplementary Data 9). The apparent lack of a TCA cycle in clades of chlamydiae whose members otherwise encode a (nearly) complete pathway points at lineage specific gene loss, potentially associated with major transition events such as the specialization to a new host. In fact, gene loss was proposed as a major driver of genotypic variation, shaping the accessory genome in pathogenic bacteria[12].

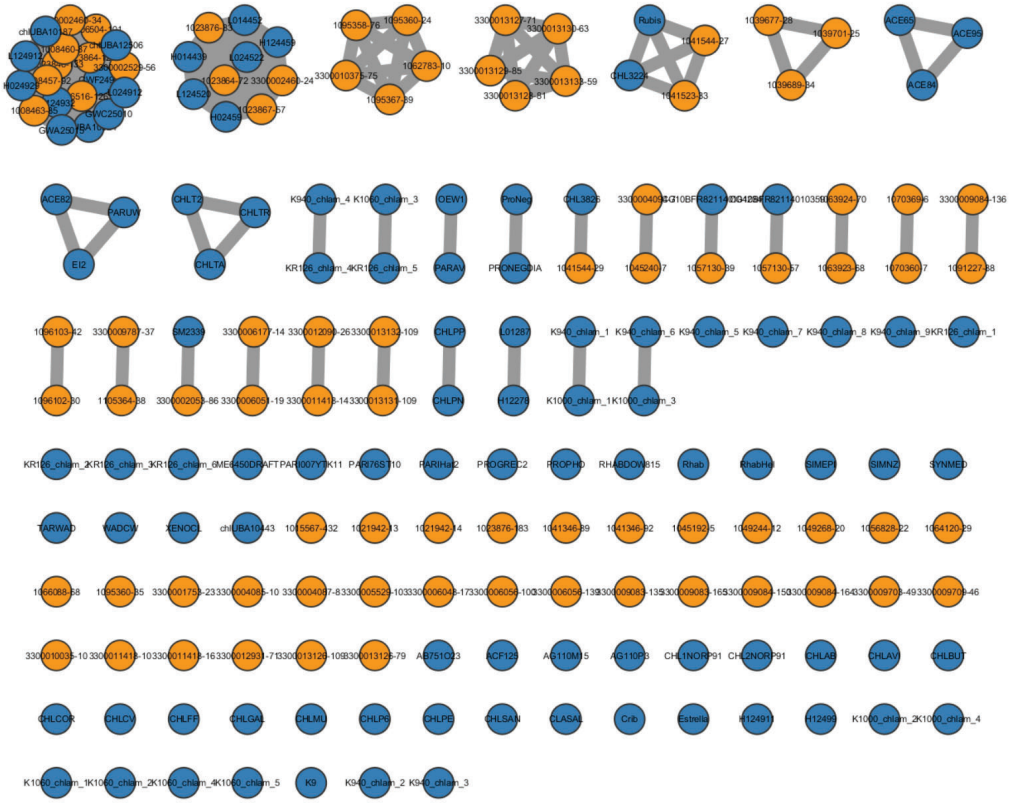
Supplementary discussion 2: Motility, mobile genetic elements, and antiviral defense

In addition to a greater metabolic versatility, environmental chlamydiae may encode a number of features unexpected and unusual in the context of the well-conserved chlamydial lifestyle[13]. Some amoeba-associated chlamydiae encode chemosensory systems proposed to regulate unknown cellular functions[2,7,14]. We find evidence for these systems in additional environmental MAGs (Supplementary Data 10). Further, several marine chlamydial SAGs and MAGs encode a semicomplete to complete gene set for a flagellar apparatus, which is generally regulated by chemosensory systems[5,8,15] (Figure 3, Supplementary Data 10). While motility likely represents an ancestral feature of chlamydiae that has been lost in many lineages[5,8], unusual features can also originate from gene gains, which have been shown to play an important role in chlamydial evolution[16–19].

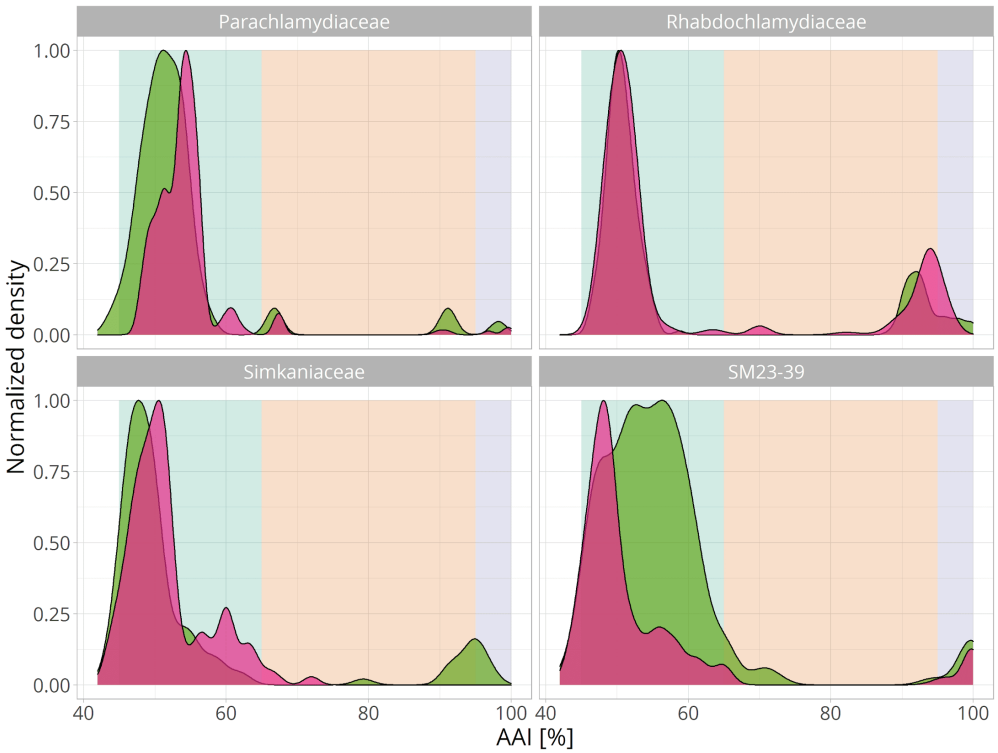
A driver of gene gain by HGT in chlamydiae are plasmids and the conjugative type IV secretion system (T4SS)[2,19,20]. We investigated the presence of the major ATPase VirB4, which is almost ubiquitously found in T4SS[21]**,** and detected homologs in six chlamydial families (Figure 3). Intriguingly, all except one VirB4 homologs (n = 21) are monophyletic and branch as a sister clade to rickettsiae, well-known intracellular alphaproteobacteria including human pathogens and amoeba-associated symbionts[2,19] (Supplementary Fig. 7). The role of the T4SS in the biology of extant chlamydiae is still unclear, but our findings provide further evidence for conjugative T4SS as a mechanism of inter-species HGT among chlamydiae[19].

Chlamydiae were considered to lack the phage defense CRISPR-Cas systems until recently[22]. Yet, a horizontally acquired CRISPR locus was discovered on plasmids of *Protochlamydia naegleriophila* KNic and *Protochlamydia massiliensis* (family Parachlamydiaceae)[23,24]. We screened the MAGs from the GEM catalogue using CRISPRCasFinder[25] and found DNA-targeting type I-C and I-E systems including six to 50 spacers in three genomes (Figure 3, Supplementary Data 11). Nucleotide blast against the NCBI viral refseq did not yield any significant hits (E-value < 0.001) for the spacers to known viral sequences. This illustrates our very limited knowledge about phages targeting chlamydiae, which have so far only been described for chlamydial pathogens in the Chlamydiaceae and sporadically for amoeba symbionts in the Parachlamydiaceae[26,27]. Of note, all chlamydiae with CRISPR-Cas systems stem from freshwater sources (Figure 1), potentially hinting at a higher relevance of phage predation in these environments.

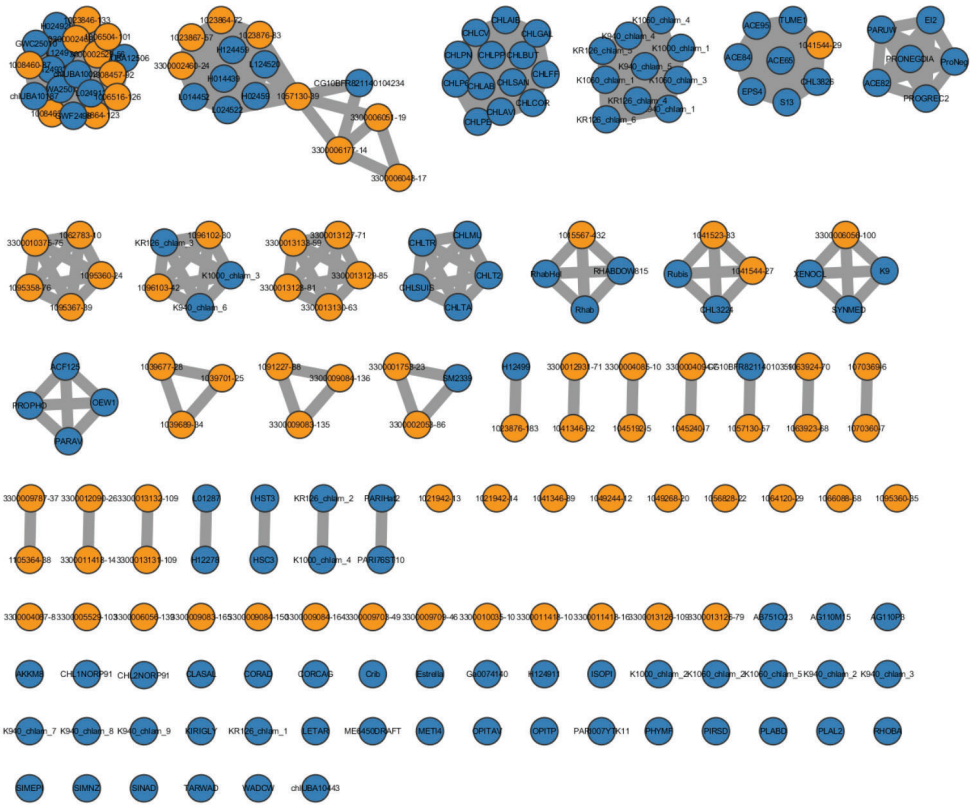
3



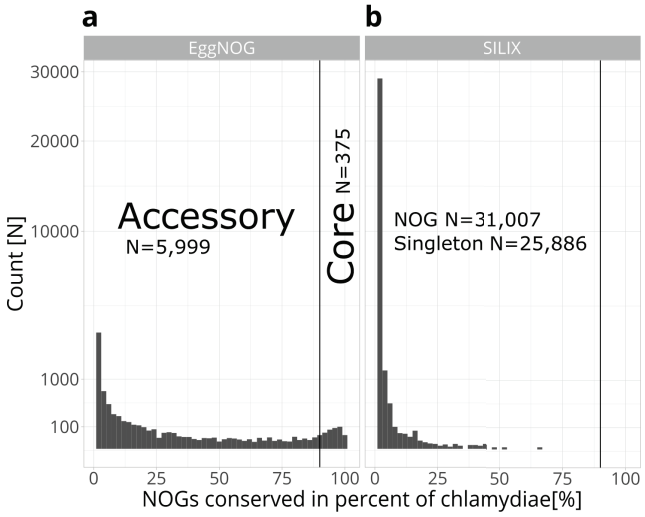
Supplementary Fig. 1: 117 species level clusters formed based on a 95% average nucleotide identity (ANI) cutoff. GEM MAGs are colored in orange and reference genomes in dark blue.



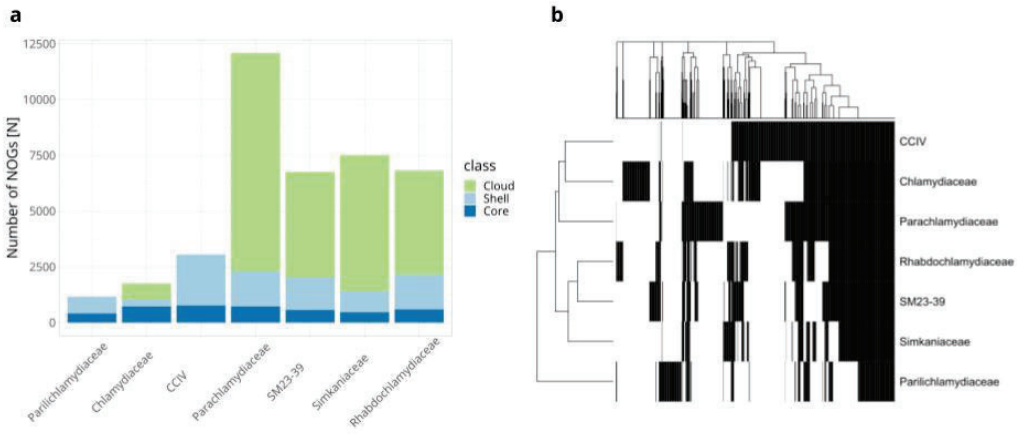
Supplementary Fig. 2: Density distribution of average amino acid identity (AAI) values of GEM MAGs. X-axis indicates AAI [%] and y-axis represents normalized density of AAI values within published chlamydial families. Curve area is colored by GEM AAI to reference (red) or other GEM MAGs (green) in the same described family. Vertical background highlights indicate different genus (45–65% AAI, light green), same genus (65-95% AAI light orange), or same species (> 95% AAI light purple) as defined by[28].



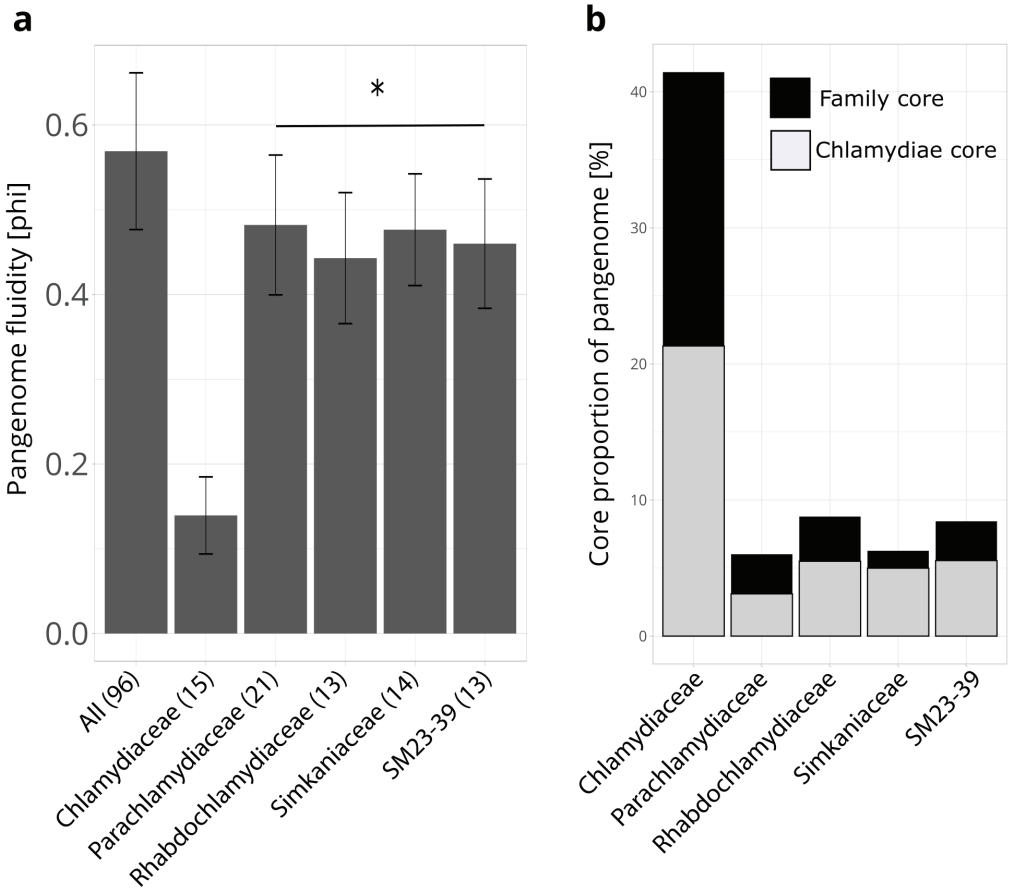
Supplementary Fig. 3: 94 genus level clusters formed based on a 65% average amino acid identity (AAI) cutoff. GEM MAGs are colored in orange and reference genomes in dark blue.



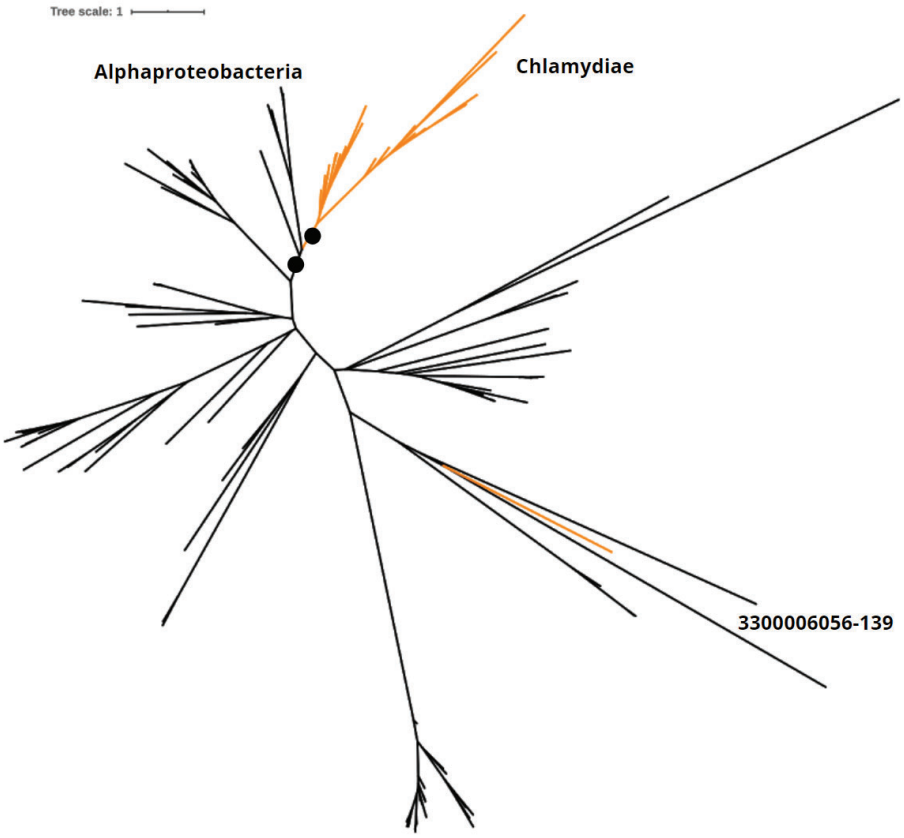
Supplementary Fig. 4: Chlamydial pangenome conservation of non-supervised orthologous groups (NOGs). NOGs were classified as core (conserved in > 90% of genomes), accessory genome (conserved in ≤90%). (**a**) NOGs derived from mapping to EggNOG 4.5 and (**b**) *de novo* clustering with SiLIX of unmapped genes.



Supplementary Fig. 5: Stable core genome and variable accessory genome sizes in chlamydial families with more than three members. (**a**) Core and accessory genome sizes of chlamydial families represented by at least three high quality genomes. (**b**) Clustering of the pangenome of chlamydial families based on presence/absence of genes.

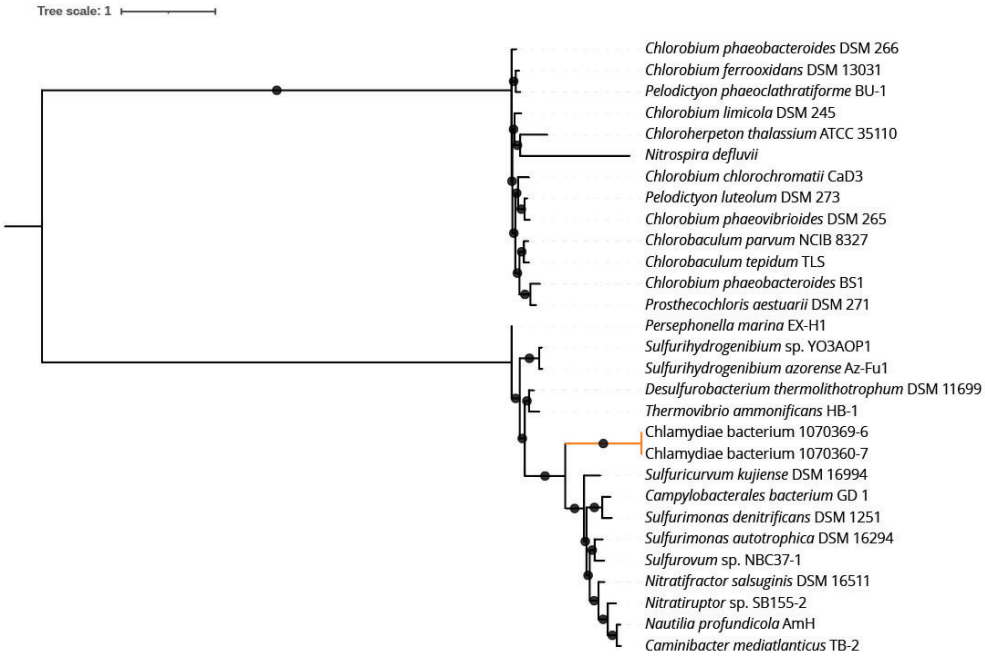


Supplementary Fig. 6: (a) Genomic fluidity of chlamydial families with at least 10 members. Bars represent mean genome fluidity between family representative genomes +/- standard deviation as error bars. Number of representative genomes used for comparison indicated in brackets next to the family label. 100 random genome pairs were compared to calculate the mean fluidity. The asterisk denotes families with a significantly different fluidity from Chlamydiaceae (p-values ≤ 0.05) based on two-sample t-tests (FDR adjusted p-values from left to right starting with Parachlamydiaceae: 3.8^{-13}, 3.9^{-10}, 8.0^{-11}, 5.4^{-10}). **(b)** Proportion of the core genome (chlamydiae- and family-specific, respectively) of the pangenome**.**

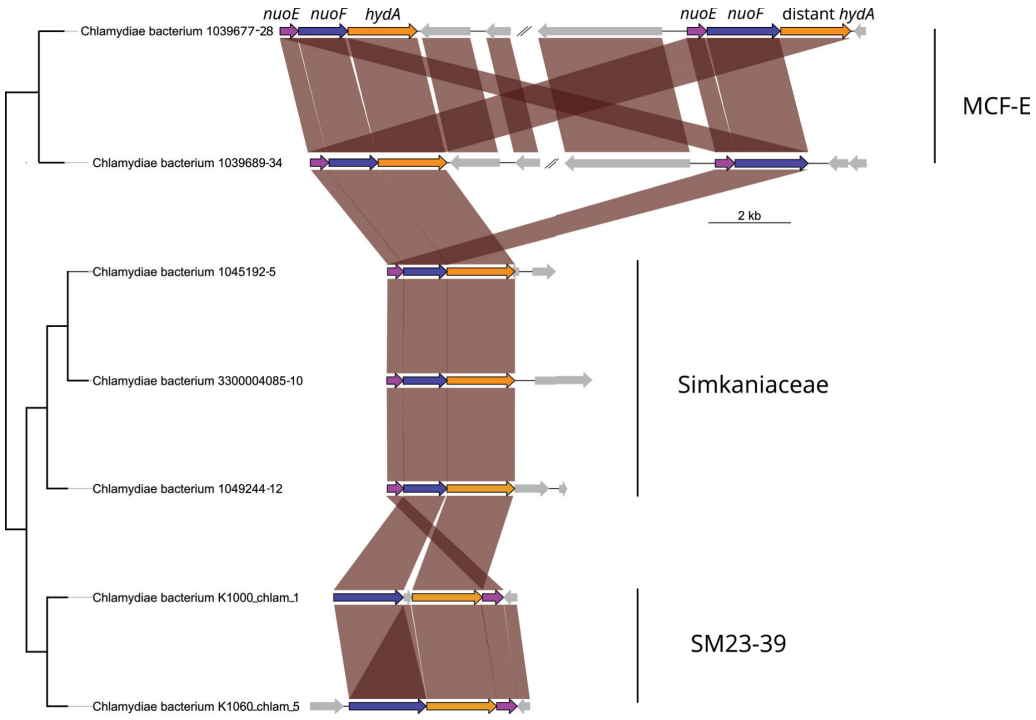


Supplementary Fig. 7: Maximum likelihood phylogenetic tree of ENOG4107S2X (VirB4) with chlamydial sequences. A large chlamydial VirB4 homologous clade (improved ultrafast bootstrap 98.2%, SH like test 97%) is sister to an alphaproteobacterial clade (improved ultrafast bootstrap 95.4%, SH like test 97%). Maximum likelihood tree was inferred under LG+C50+G+F model with 1,000 improved ultrafast bootstraps and 1,000 replicates of the SH-like approximate likelihood ratio test.

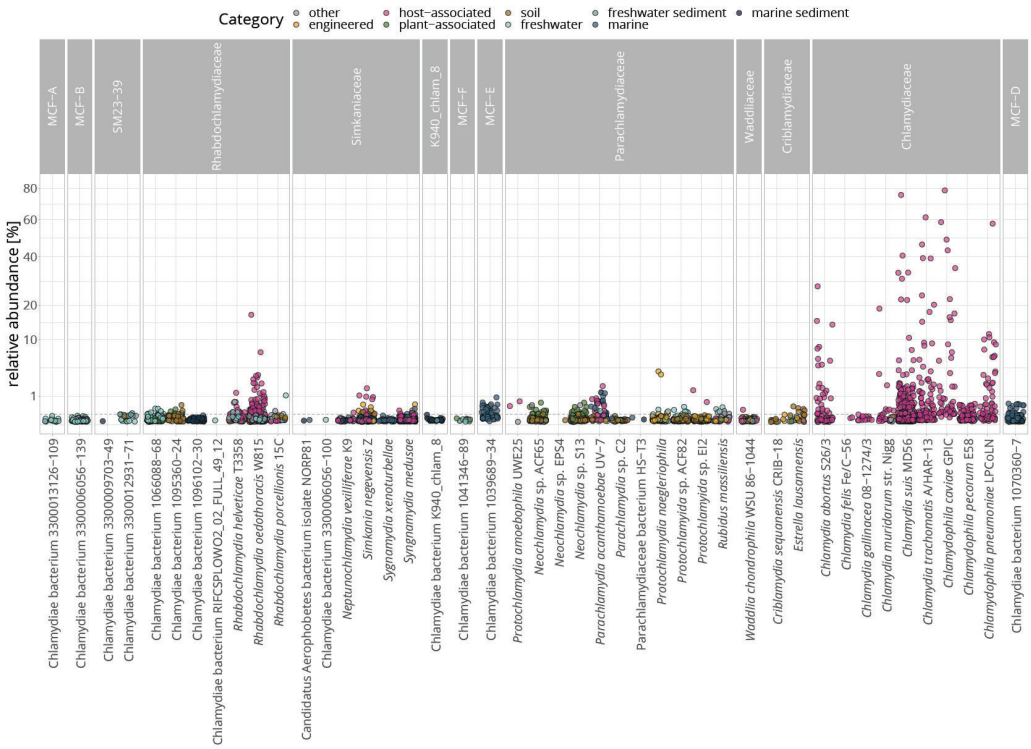
9



Supplementary Fig. 8: Maximum likelihood phylogenetic tree of ENOG4105C63 (AclA) with chlamydial sequences. Chlamydiae are monophyletic with a clade of campylobacterotal AclA (improved ultrafast bootstrap 100%, SH like test 100%). Maximum likelihood tree was inferred under LG+C20+G+F model with 1,000 improved ultrafast bootstraps and 1,000 replicates of the SH-like approximate likelihood ratio test. Filled circles at nodes indicate a bootstrap support > 95%.



Supplementary Fig. 9: Gene cluster structure of representatives of group A [FeFe]-hydrogenases encoded in members of the families SM23-39, Simkaniaceae, and MCF-E. The second gene cluster in MCF-E contains the distant copy of *hydA*. Comparisons are ordered according to phylogenomic species tree. Red bands indicate genes belonging to the respective NOG.



Supplementary Fig. 10: Relative abundance of chlamydial 16S rRNA gene amplicons in SRA samples per species representative with a sequenced genome.

Supplementary references

1. Omsland, A., Sixt, B. S., Horn, M. & Hackstadt, T. Chlamydial metabolism revisited: interspecies metabolic variability and developmental stage-specific physiologic activities. *FEMS Microbiol. Rev.* **38**, 779–801 (2014).
2. Collingro, A. *et al.* Unity in Variety--The Pan-Genome of the Chlamydiae. *Molecular Biology and Evolution* **28**, 3253–3270 (2011).
3. Bertelli, C. *et al.* The Waddlia Genome: A Window into Chlamydial Biology. *PLoS ONE* **5**, e10890 (2010).
4. Bertelli, C. *et al.* Sequencing and characterizing the genome of Estrella lausannensis as an undergraduate project: training students and biological insights. *Front. Microbiol.* **6**,

101 (2015).

5. Dharamshi, J. E. *et al.* Marine Sediments Illuminate Chlamydiae Diversity and Evolution. *Curr. Biol.* **30**, 1032–1048.e7 (2020).
6. Bertelli, C. *et al.* The Waddlia Genome: A Window into Chlamydial Biology. *PLoS ONE* **5**, e10890 (2010).
7. Bertelli, C. *et al.* Sequencing and characterizing the genome of Estrella lausannensis as an undergraduate project: training students and biological insights. *Front. Microbiol.* **6**, 101 (2015).
8. Collingro, A. *et al.* Unexpected genomic features in widespread intracellular bacteria: evidence for motility of marine chlamydiae. *ISME J.* **11**, 2334–2344 (2017).
9. Mehlitz, A. *et al.* Metabolic adaptation of Chlamydia trachomatis to mammalian host cells. *Mol. Microbiol.* **103**, 1004–1019 (2017).
10. Ishida, K. *et al.* Amoebal endosymbiont Neochlamydia genome sequence illuminates the bacterial role in the defense of the host amoebae against Legionella pneumophila. *PLoS One* **9**, e95166 (2014).
11. Taylor-Brown, A. *et al.* Metagenomic Analysis of Fish-Associated Ca. Parilichlamydiaceae Reveals Striking Metabolic Similarities to the Terrestrial Chlamydiaceae. *Genome Biol. Evol.* **10**, 2587–2595 (2018).
12. Bolotin, E. & Hershberg, R. Gene Loss Dominates As a Source of Genetic Variation within Clonal Pathogenic Bacterial Species. *Genome Biol. Evol.* **7**, 2173–2187 (2015).
13. Collingro, A., Köstlbacher, S. & Horn, M. Chlamydiae in the Environment. *Trends Microbiol.* (2020) doi:10.1016/j.tim.2020.05.020.
14. Bertelli, C., Goesmann, A. & Greub, G. Criblamydia sequanensis Harbors a Megaplasmid Encoding Arsenite Resistance. *Genome Announc.* **2**, (2014).
15. Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci Data* **5**, 170203 (2018).

16. Kim, H., Kwak, W., Yoon, S. H., Kang, D.-K. & Kim, H. Horizontal gene transfer of Chlamydia: Novel insights from tree reconciliation. *PLoS One* **13**, e0195139 (2018).
17. Domman, D. *et al.* Massive expansion of Ubiquitination-related gene families within the Chlamydiae. *Mol. Biol. Evol.* **31**, 2890–2904 (2014).
18. Kamneva, O. K., Knight, S. J., Liberles, D. A. & Ward, N. L. Analysis of genome content evolution in pvc bacterial super-phylum: assessment of candidate genes associated with cellular organization and lifestyle. *Genome Biol. Evol.* **4**, 1375–1390 (2012).
19. Köstlbacher, S., Collingro, A., Halter, T., Domman, D. & Horn, M. Coevolving Plasmids Drive Gene Flow and Genome Plasticity in Host-Associated Intracellular Bacteria. *Current Biology* **31**, 346–357.e3 (2021).
20. Greub, G., Collyn, F., Guy, L. & Roten, C.-A. A genomic island present along the bacterial chromosome of the Parachlamydiaceae UWE25, an obligate amoebal endosymbiont, encodes a potentially functional F-like conjugative DNA transfer system. *BMC Microbiol.* **4**, 48 (2004).
21. Guglielmini, J. *et al.* Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. *Nucleic Acids Research* **42**, 5715–5727 (2014).
22. Burstein, D. *et al.* Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. *Nat. Commun.* **7**, 10613 (2016).
23. Bertelli, C. *et al.* CRISPR System Acquisition and Evolution of an Obligate IntracellularChlamydia-Related Bacterium. *Genome Biology and Evolution* **8**, 2376–2386 (2016).
24. Benamar, S. *et al.* Developmental Cycle and Genome Analysis of Protochlamydia massiliensis sp. nov. a New Species in the Parachlamydiacae Family. *Front. Cell. Infect. Microbiol.* **7**, (2017).
25. Couvin, D. *et al.* CRISPRCasFinder, an update of CRISRFinder, includes a portable

version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res.* **46**, W246–W251 (2018).

26. Śliwa-Dominiak, J., Suszyńska, E., Pawlikowska, M. & Deptuła, W. Chlamydia bacteriophages. *Arch. Microbiol.* **195**, 765–771 (2013).
27. Corsaro, D., Müller, K.-D., Wingender, J. & Michel, R. 'Candidatus Mesochlamydia elodeae' (Chlamydiae: Parachlamydiaceae), a novel chlamydia parasite of free-living amoebae. *Parasitology Research* **112**, 829–838 (2013).
28. Konstantinidis, K. T., Rosselló-Móra, R. & Amann, R. Uncultivated microbes in need of their own taxonomy. *ISME J.* **11**, 2399–2406 (2017).

CHAPTER V

Gene gain facilitated the origin and diversification of an ancient phylum of symbionts

Authors:

Jennah E. Dharamshi‡[1], Stephan Köstlbacher‡[2,3], Max E. Schön[1], Astrid Collingro[2], Thijs J. G. Ettema‡[1,3], Matthias Horn‡[2]

‡ Equal contribution

‡ Equal contribution

Published in:

Under review (submitted to Nature Microbiology)

Gene gain facilitated the origin and diversification of an ancient phylum of symbionts

Jennah E. Dharamshi†[1], Stephan Köstlbacher†[2,3], Max E. Schön[1], Astrid Collingro[2], Thijs J. G. Ettema‡*[1,3], Matthias Horn‡*[2]

[1] Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, SE-75123 Uppsala, Sweden
[2] Centre for Microbiology and Environmental Systems Science, University of Vienna, Vienna, Austria
[3] Laboratory of Microbiology, Wageningen University, 6708 WE Wageningen, The Netherlands

† Equal contribution
‡ Equal contribution

* Correspondence to: thijs.ettema@wur.nl and matthias.horn@univie.ac.at

ABSTRACT

Members of the Planctomycetes-Verrucomicrobia-Chlamydiae (PVC) bacterial superphylum play diverse and important ecological roles in many environments. While PVC bacteria are primarily free-living, the phylum Chlamydiae is an exception. This group of highly successful intracellular symbionts of eukaryotes includes pathogens alongside diverse relatives with unknown environmental roles. Despite their conserved lifestyle, it is unclear how chlamydial symbiosis evolved. Here, we studied chlamydial evolution by performing large-scale phylogenomic analyses and gene-tree aware ancestral state reconstruction using a wide sampling of PVC genomic diversity. Unexpectedly for strict intracellular symbionts, we found that Chlamydiae evolution was characterized not just by extensive loss, but also by genome expansion from diverse horizontal gene transfer events. The Chlamydiae ancestor gained the genetic capability to infect eukaryotic hosts, indicating that the chlamydial symbiotic lifestyle has prevailed over a billion years of evolutionary history and diversification. This chlamydial ancestor was a facultative anaerobe, with the potential to transition between oxic and anoxic environments. Key differences in underlying energy metabolism and aerobiosis later emerged along a major split within Chlamydiae underpinned by complex genome dynamics. Host-associated lifestyles are widespread among bacteria, and together our analyses provide a blueprint for understanding major transitions in their evolution.

INTRODUCTION

Host-symbiont associations between eukaryotes and bacteria are ubiquitous, having evolved numerous times and spanning the mutualism-parasitism spectrum. Interactions and dependencies between eukaryotic hosts and their bacterial symbionts play essential roles, from ecosystem functioning to the evolution of biological complexity. Studying the ancient origins of host-associated symbionts is necessary for unravelling the underlying evolutionary processes and molecular mechanisms. The Planctomycetes-Verrucomicrobia-Chlamydiae (PVC) superphylum is a ubiquitous group of bacteria where host-association has evolved multiple times, thus representing an ideal case for investigating the evolution of symbioses. This superphylum shares common ancestry and includes the aforementioned phyla alongside Lentisphaerae, Kirimatiellaeota, and other potential members[1,2]. PVC bacteria have large variations in lifestyle and metabolism and include members with importance from ecological, medical, and industrial perspectives[1,3,4]. While most PVC bacteria are not host-associated, all described members of the phylum Chlamydiae are obligate intracellular symbionts of eukaryotes[3].

Chlamydiae has gained recognition for its medical importance as it includes well-known pathogens of humans (*e.g.*, *Chlamydia trachomatis*)[5]. Other members of the Chlamydiaceae family are also animal pathogens with a high health burden and zoonotic potential[5–7]. Additional members of the Chlamydiae phylum are found ubiquitously in environmental samples[8,9] and have been identified as symbionts of a wide range of both microbial and multicellular eukaryotes[2,10,11]. These so-called "environmental chlamydiae" and the Chlamydiaceae have large differences in genome size and content, with larger metabolic capacity in the former[12,13]. Despite these contrasting genomic features, described chlamydiae share a conserved biphasic lifestyle with an intracellular replicative phase as reticulate bodies (RBs), and a non-dividing extracellular phase as elementary bodies (EBs)[5]. Chlamydiae diverged from other PVC bacteria 1-2 billion years ago (Gya)[14,15], and it has been proposed that their obligate intracellular lifestyle evolved early and was accompanied by genome reduction[14,16–18]. However, since only few cultured representatives are known, these studies included minimal chlamydial genomic diversity. Little is thus known about the early evolution of symbiosis, host adaptation, and virulence in this ancient phylum of symbionts. The recent rise of culture-independent genomics has led to the retrieval of numerous uncultured chlamydial lineages from various environments[19–24]. This revealed the genetic potential for

47 motility[22,23], and anaerobic metabolism and hydrogen production[24,25], thereby fundamentally
48 changing our understanding of chlamydial physiology and biology.
49 Here, we take advantage of the recently expanded PVC genomic diversity to investigate
50 the evolution of symbiosis as exemplified by the Chlamydiae phylum. We performed in-depth
51 phylogenomic analyses to reconstruct evolutionary relationships between extant PVC lineages in
52 addition to gene-tree aware ancestral state reconstruction. We found that the Chlamydiae ancestor
53 had already obtained key genomic features associated with a symbiotic lifestyle, suggesting the
54 capability to infect eukaryotic hosts over long-term evolutionary history. Our analyses indicate
55 that this ancestor was a motile facultative anaerobe, suggestive of a lifestyle that involved frequent
56 transitions between environments with varying oxygen levels. Later, major shifts in chlamydial
57 energy metabolism and oxygen tolerance were driven by horizontal gene transfer (HGT).
58 Unexpectedly, this led to genome expansion in some groups, counter to the common perception
59 of a trajectory towards genome reduction in intracellular symbionts.

RESULTS AND DISCUSSION

Establishment of a resolved species phylogeny

60 For reconstructing chlamydial evolutionary history representative sampling of PVC genomes was
61 required to accurately resolve species relationships (Extended Data Figure 1). To this end, we
62 collected all publicly available high-quality PVC genomes (completeness ≥ 90% and redundancy
63 ≤ 2%) and selected species (Chlamydiae) and genus (other PVC bacteria) representatives with the
64 highest genome quality (Figure S1, Extended Data Figure 1, and Data S1). This selection resulted
65 in 184 PVC representatives, including 95 chlamydial species. These chlamydiae originated from
66 nine diverse environments and have a wide-range in genome size (0.77 - 3.4 Mbp) and GC content
67 (26.2 - 49.1%) (Figure 1 and Data S2). Although some of these lineages are known to be host-
68 associated, the majority are represented by metagenome-assembled genomes (MAGs) and have
69 unknown lifestyles (Figure 1).
70 A phylogenomic dataset of 74 single-copy marker genes was curated (Figure S2 and Data
71 S3) and concatenated phylogenies then inferred using both maximum-likelihood (ML) and
72 Bayesian methods, with well-known phylogenetic artefacts taken into account (*i.e.*, compositional
73 biases and long branch attraction (LBA))[26] (Figures 1 and S3-S5, and Data S4-S6). Across all

74 species phylogenies the monophyly of Chlamydiae was fully supported and chlamydial families
75 consistently resolved (Figures 1 and S3-S5). Family assignments were broadly consistent with a
76 species tree based on the 16S rRNA gene (Figure S6). However, four genomes of long-branching
77 chlamydial lineages were removed due to unstable positions (Figures S3-S4; Supplementary
78 Discussion 1). This resulted in a final dataset of 180 PVC genomes where deep evolutionary
79 relationships were consistently resolved in both Bayesian and ML analyses when compositional
80 bias was taken into account (Figures 1 and S5). Together, these phylogenomic analyses have
81 resulted in the most comprehensive and robust species tree for the Chlamydiae phylum currently
82 available.
83 Based on the Chlamydiae species tree inconsistencies were identified in regard to the
84 currently accepted chlamydial taxonomy. For example, the previously established order
85 Parachlamydiales (Simkaniaceae, Rhabdochlamydiaceae, Criblamydiaceae, Waddliaceae, and
86 Parachlamydiaceae)[27] is paraphyletic in our phylogenomic inferences (Figures 1 and S3-S5), and
87 in other recent analyses[22,24,25,28]. We thus suggest redefining taxonomic groups in keeping with
88 evolutionary relationships (Supplementary Discussion 2). There was a clear and well-supported
89 early divergence of Chlamydiae into two major groups, we refer to here as Group 1 (G1) and
90 Group 2 (G2) (Figure 1). We could further subdivide G1 into two putative orders: Simkaniales
91 (families Simkaniaceae and Rhabdochlamydiaceae) and Anoxychlamydiales
92 (Anoxychlamydiaceae, formerly Anoxychlamydiales[22], and Chlamydiae Clade III). Members of
93 G1 are primarily MAGs acquired from diverse environments, although several distinct groups are
94 likely host-associated as they were obtained from marine and terrestrial invertebrates (Figure 1).
95 The G2 subdivision includes the more classical chlamydial animal pathogens (Chlamydiaceae)
96 and protist-infecting chlamydiae. The previously established G2 Chlamydiales order[27] now
97 includes Chlamydiae Clade IV in addition to Chlamydiaceae and *Clavichlamydia*. While
98 remaining G2 families (Criblamydiaceae, Waddliaceae, and Parachlamydiaceae) and orphan
99 lineages comprise the newly defined order Amoebachlamydiales (Figure 1).

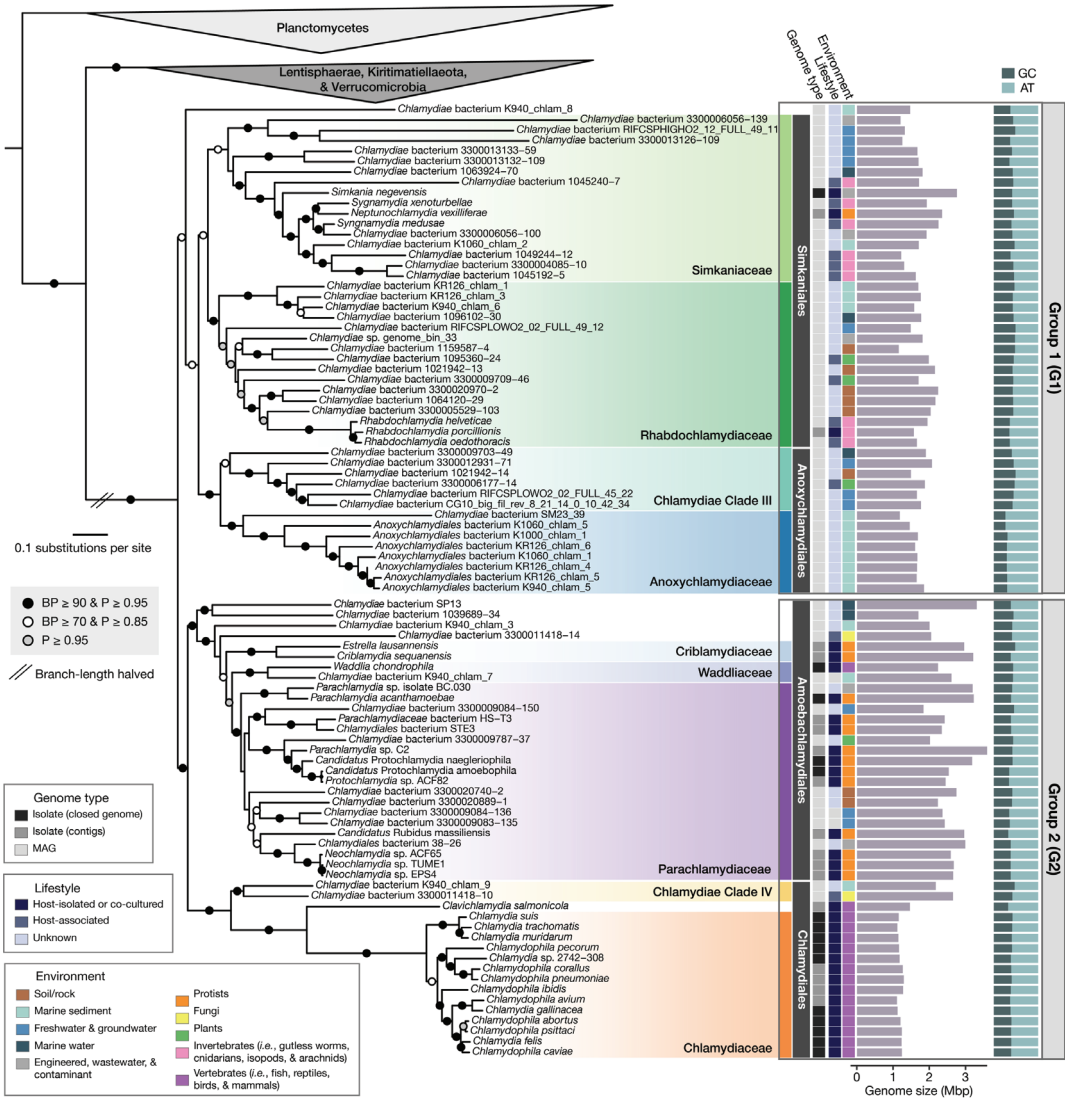


Figure 1. Robust species phylogeny of PVC bacteria. Bayesian phylogenetic tree of species relationships between PVC members (180 taxa) inferred using a concatenated supermatrix alignment of 74 single-copy marker proteins with compositionally heterogeneous sites removed (8151 amino acid sites). The consensus tree of converged chains with lowest maximum discrepancy is shown (Figure S5 and Data S6). Bipartition support is indicated by circles as noted in the legend according to posterior probability (P) (CAT+GTR+Γ4 model), and non-parametric bootstrap support values (BP) (LG+C60+F+Γ4-derived PMSF approximation). The tree is rooted by the outgroup phylum Planctomycetes. The reduced branch length leading to Chlamydiae is indicated by parallel lines. The scale bar indicates the number of substitutions per site. Genome type, lifestyle, and environmental source is indicated by coloured squares for each chlamydiae representative according to the legend. Genome size (Mbp length in purple), and GC content (%GC in dark blue and %AT in light blue) are indicated by bars. Proposed higher-level taxonomic classifications are indicated and chlamydial families outlined with coloured boxes, this colour scheme representation is continued in subsequent figures. See also Figures S1-S6, Extended Data Figure 1, and Data S1-S6.

The chlamydial symbiotic lifestyle is ancestral

To reconstruct gene content evolution in key PVC ancestors we used a gene tree aware approach, as implemented in ALE (amalgamated likelihood estimation)[29–31] (Figures S7-S9, and Data S7-S8). In short, we clustered genes from across the PVC genome dataset into protein families and inferred phylogenies for those with more than three sequences (n=11,996). Gene tree samples (*i.e.*, ML bootstraps) were then reconciled with the obtained species tree (Figure 1) to infer the likelihood of evolutionary events (speciations, originations, duplications, transfers, and losses) and gene copy numbers across species tree nodes (Extended Data Figure 1, and Figure S10). Originations can represent either *de novo* gene families (*i.e.* gene birth events) or horizontal gene transfers (HGT) from outside of the PVC bacteria genome dataset. To differentiate between these for chlamydiae originations we searched for homologs in public protein databases. If homologs could be identified, we further inferred phylogenetic trees to determine the likely donor lineage of the gene family.

Our analysis revealed that Chlamydiae evolved from a last common ancestor (LCCA) with ~1,155 protein coding genes that was already adapted to a symbiotic lifestyle (Figures 2-3, Extended Data Figure 2, and Data S9). Almost a third (n=353) of the total genes in LCCA were inferred to be gene gains (Figure 3), many associated with metabolism (n=91; Extended Data Figure 3) and obtained through HGT (Extended Data Figure 4). LCCA acquired genes representing hallmarks of an endosymbiotic lifestyle, such as genes involved in host interaction, energy parasitism, and the chlamydial biphasic lifecycle. Cultured chlamydiae use a highly conserved type III secretion system (T3SS) to facilitate interaction and entry into eukaryotic hosts through the secretion of effectors[32,33]. We found that the T3SS (*e.g.*, *sctJ*, *sctT*, *sctS*, *sctV*, and *sctW* genes) was gained in LCCA and maintained, without exception, in all chlamydial ancestors (Figure 2, Extended Data Figure 5, and Data S10). LCCA also gained genes considered Chlamydiaceae pathogenicity factors that are involved in host invasion, such as the adhesin Ctad1 and the major outer membrane protein (MOMP) (Figure 2)[5]. In their replicative RB stage cultured chlamydiae can use nucleotide transporters (NTTs) to import ATP, nucleotides, and NAD+ from the host cytosol[34–36]. Two NTTs were inferred as present in LCCA and would have allowed for energy parasitism and scavenging of metabolites (Figure 2).

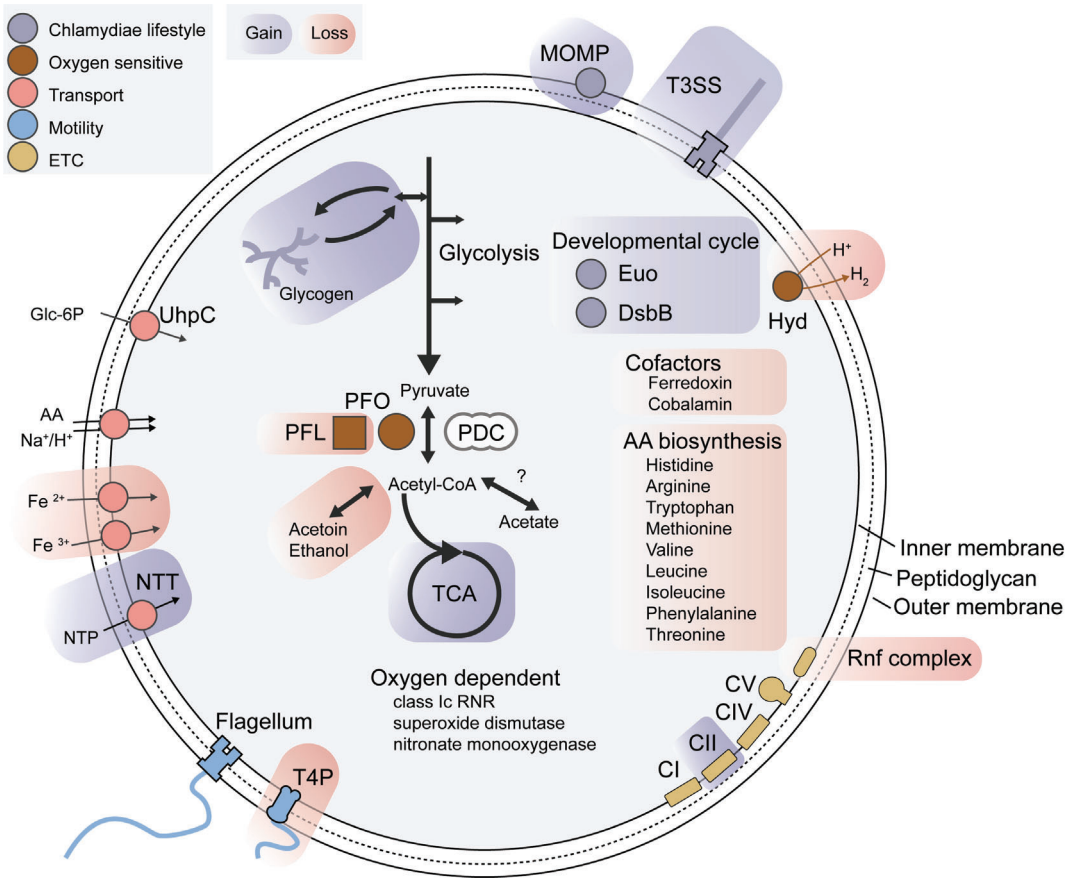


Figure 2. Schematic reconstruction of gene content in the last common Chlamydiae ancestor, LCCA. Enzyme complexes were inferred as present if at least half of the necessary genes were inferred. Gains and losses relative to the last common ancestor of Chlamydiae, Verrucomicrobia, Lentisphaerae, and Kiritimatiellaeota (LVCCA) are indicated as highlighted backgrounds in blue and red, respectively. Our analysis suggests the presence of a peptidoglycan based cell wall indicated by the dashed line between inner and outer membrane. Cytoplasm is indicated by a light grey background. ADI, arginine deiminase pathway; CI-V, electron transport chain complexes I-V; ETC, electron transport chain; Euo, early upstream ORF; Hyd, [FeFe] hydrogenase; MOMP, major outer membrane protein; NTT, nucleotide transporter; PDC, pyruvate dehydrogenase complex; PFL, pyruvate formate-lyase; PFO, pyruvate ferredoxin oxidoreductase; Rnf, a ferredoxin:NAD(+)-oxidoreductase complex; RNR, a Ribonucleotide reductase; T3SS, type III secretion system; T4P, type IV pilus; TCA, tricarboxylic acid cycle; UhpC, glucose 6-phosphate transporter. See also Extended Data Figure 2 for a detailed reconstruction of gene content present in LCCA.

When in their extracellular EB stage, chlamydiae remain metabolically active while being protected from osmotic and physical stress[37,38]. This is facilitated by a more rigid cell envelope due to disulfide crosslinking of outer membrane proteins by DsbB[39], which was reconstructed as having originated in LCCA (Figure 2). EBs are also characterized by condensed DNA, resulting from the action of histone-like proteins such as HctA (Histone H1-like). HctA was not inferred in

LCCA, but it was present in all other early chlamydial ancestors (Extended Data Figure 5). In some Chlamydiaceae, host-derived glucose is stored as glycogen in the late stages of infection before EB transition[40]. This glycogen is proposed to enhance extracellular survival by providing a carbon source while awaiting a host encounter[41]. Glycogen synthesis and degradation were inferred as ancestral features of LCCA, with key genes inferred as originating here (*e.g.*, *glgC*, *glgP*, *malQ*) (Figure 2 and Extended Data Figure 2). Intriguingly, for 28 LCCA originations homologs outside Chlamydiae were not identified and they were thus classified as *de novo* gene candidates (Extended Data Figure 4). Most notable was the early upstream open reading frame (*euo*) gene, which our results indicate was a chlamydial invention (Figure 2). Euo is considered the master regulator of the transition from RBs to EBs, as it represses genes required late in this transition such as those involved in type III secretion, DNA condensation, and cell surface modification[42,43]. Generally, gene features related to their biphasic lifecycle and endosymbiosis were highly conserved in chlamydial ancestors (Extended Data Figure 3 and Data S10).

Despite these many gains, LCCA also underwent extensive gene loss, losing 23% (n=287) of the protein-coding genes reconstructed in the last common ancestor of Chlamydiae, Verrucomicrobia, Lentisphaerae, and Kiritimatiellaeota (LVCCA) (Figure 3). These genes were likewise predominantly associated with metabolism (61%, Extended Data Figure 3). LCCA strongly reduced amino acid biosynthesis capabilities relative to LVCCA through the loss of *de novo* biosynthesis genes for histidine, arginine, tryptophan, methionine, valine, leucine, isoleucine, phenylalanine, and threonine (Figure 2, Extended Data Figure 2, and Data S9). In addition, LCCA underwent a reduction in *de novo* nucleotide biosynthesis capabilities, such as the loss of key purine biosynthesis genes (*e.g.*, *purC*, *purD*, and *purH*) (Extended Data Figure 2 and Data S9). However, LCCA maintained a suite of putative amino acid and oligopeptide transporters present in LVCCA (Data S9). The LCCA thus could have acquired amino acids and nucleotides (*i.e.*, through NTTs) from external sources. Although LCCA could synthesise NAD and NADP it likely depended on the uptake of other cofactors, such as ferredoxin and cobalamin, as biosynthesis pathways for these were lost (Figure 2 and Data S9). Surprisingly, some Chlamydiaceae pathogenicity factors associated with host metabolite degradation (*e.g.*, proteases and lipases) were inferred in LVCCA and retained throughout chlamydial evolution (Extended Data Figure 5 and Data S10). LVCCA could also import glucose-6-phosphate (Glc-6P) using UhpC, which is used by chlamydiae to scavenge host glucose and is a mechanism often employed by endosymbionts[44]

(Figure 2).

The lineage leading to Chlamydiae diverged from other PVC bacteria between one and two Gyr, coinciding with current estimates for the range in timing for eukaryotic evolution (1.2 - 2.1 Gyr)[14,15,17,18]. Our reconstruction using a phylogeny-aware approach supports earlier hypotheses[17,18] by demonstrating that LCCA already had the genetic toolkit for an endosymbiotic and biphasic lifecycle, and that this has been maintained across the phylum over long-scale evolutionary time. Furthermore, our results demonstrate that LCCA was already capable of intracellular infection of early eukaryotes. Together, this suggests a billion-year old history of symbiotic interactions during which chlamydiae infected eukaryotic hosts as they evolved.

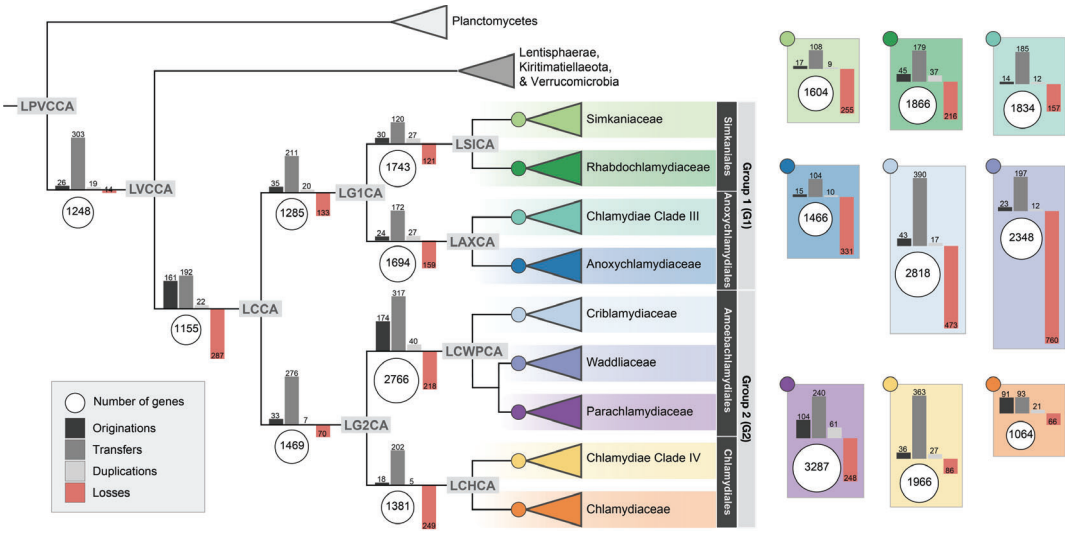


Figure 3. Ancestral reconstruction of Chlamydiae gene content evolution reveals complex genome dynamics. Schematic phylogenetic tree corresponds to the Bayesian consensus tree depicted in Figure 1. Nodes are annotated with their inferred gene copy number in the white circles, while branches are annotated with bars representing the number of originations, transfer, and duplication events in greyscale (see legend). The number of loss events are indicated negatively by a red bar. Terminal nodes represent chlamydial family ancestors, collapsed into triangles, with the corresponding node events shown to the right. Orphan lineages are excluded. See Figure S10 for events across all PVC bacteria ancestor nodes..

A facultative anaerobic origin of Chlamydiae

Recent studies have challenged whether chlamydiae are strictly aerobic with the finding of anaerobiosis-related genes in several newly identified chlamydial groups (*e.g.*, Anoxychlamydiaceae)[24,25]. With this in mind we investigated metabolism in LCCA and how it evolved from LVCCA, focusing on the evolutionary history of genes associated with aerobic and

anaerobic lifestyles. Of note, ancestral state reconstructions of pathways tend to be incomplete, as it is only possible to reconstruct gene content that is present in genomes of extant PVC members found in our dataset. Functions of genes that appear to be missing may have been represented by other gene families that have gone extinct or were not sampled.

Genes associated with both aerobic and anaerobic lifestyles were reconstructed in LCCA indicating that it was a facultative anaerobe (Figure 2). In general, core metabolic genes found in extant chlamydiae[13] were already inferred in LCCA (Figure 4, Extended Data Figures 5-6, and Data S9-S10). LCCA could generate ATP through glycolysis using glucose (glucokinase) and glucose-6-phosphate (UhpC) (Figure 2). The conversion of the resulting pyruvate to acetyl-CoA could then be achieved by two different enzymes: pyruvate dehydrogenase (PDH) under oxic conditions and conversely the highly oxygen-sensitive pyruvate:ferredoxin oxidoreductase (PFO) under anoxic conditions[45] (Figure 2). Acetyl-CoA was potentially directed into the TCA cycle, but it may also have been fermented (Extended Data Figure 6). The presence of fermentative metabolism indicates oxygen-independent energy conservation. We inferred a near-complete TCA cycle in LCCA (Extended Data Figure 2). Only citrate synthase and malate dehydrogenase were missing, although these were inferred in the majority of chlamydial order ancestors (Extended Data Figure 6). We inferred a complete electron transport chain (ETC) in LCCA for performing oxidative phosphorylation (Extended Data Figure 2). This ETC included sodium-transporting NADH dehydrogenase (Nqr; Complex I, CI), succinate dehydrogenase (Sdh; CII), the terminal oxidases (CIV) cytochrome bd ubiquinol oxidase (CydA-B) and cytochrome c oxidase cbb_3-type (CcoO/N), and sodium-driven ATP synthase (Ntp; CV) (Figure 4a-b and Data S10). The identified terminal oxidases both have high oxygen affinity and could be used to respire oxygen under microaerophilic conditions. However, they could also provide a protective effect against oxidative and nitrosative stress for oxygen-sensitive enzymes like PFO[46,47]. The LCCA ETC was likely used to generate a sodium motive force (SMF) as has been demonstrated in *C. trachomatis*[48]. This core LCCA metabolism was also reconstructed in LVCCA, with the exception of succinate dehydrogenase (Extended Data Figure 6), likewise indicating a facultative anaerobe lifestyle. LVCCA also had an Rnf complex (*i.e.*, sodium ion-translocating ferredoxin:NAD^+ oxidoreductase) as part of its ETC, which is strictly linked to sodium energetics and strongly associated with anaerobes[49] (Figure 2 and Data S9). LVCCA had more extensive fermentative capabilities with the potential to ferment pyruvate to acetate, acetoin, and potentially ethanol

(Figure 2 and Data S9). LVCCA could additionally oxidize pyruvate using the oxygen-sensitive pyruvate formate lyase (PFL)[50], and couple pyruvate oxidation to H_2 production with the highly oxygen-sensitive [FeFe]-hydrogenase (HydA) (Figure 2 and Data S9)[45,51]. PFL, HydA, the Rnf complex, and some fermentative capabilities were lost in LCCA (Figure 2 and Data S9).

Further evidence for a facultative anaerobic lifestyle in LCCA and LVCCA is demonstrated by the presence of genes encoding central enzymatic functions with varying levels of oxygen tolerance. For example, LCCA encoded complementary copies of Ribonucleotide reductase (RNR), the key metabolic enzyme for deoxyribonucleotide interconversion from corresponding ribonucleotides[52] (Figure 2). Under anoxic conditions LCCA and LVCCA could use a class III RNR, which is extremely sensitive to oxygen[52], and under oxic conditions LCCA a class Ic RNR, which is oxygen-dependent[53]. These RNRs have since been differentially retained in chlamydial groups, and an additional oxygen-independent RNR was gained in the Amoebachlamydiales (Extended Data Figure 5). Heme biosynthesis can likewise occur through both aerobic and anaerobic routes. LCCA and LVCCA encoded the oxygen-independent Coproporphyrinogen III oxidase, which was retained in all family ancestors and is part of the anaerobic pathway (Extended Data Figure 5). Conversely, most early chlamydial ancestors, though not LCCA and LVCCA, encoded the oxygen-dependent Protoporphyrinogen III oxidase, which is instead associated with the aerobic route (Extended Data Figure 5). Several additional oxygen-dependent enzymes were reconstructed in LCCA including a superoxide dismutase and nitronate monooxygenase, also inferred in LVCCA, which detoxify oxygen radicals[54] and oxidize alkyl nitronates[55], respectively (Extended Data Figure 5). An iron transport system was not reconstructed in LCCA. However, LVCCA was inferred to have been able to transport both ferrous (Fe^{2+}) and ferric (Fe^{3+}) iron (Figure 2 and Data S9), the primary species under anoxic and oxic conditions, respectively[56,57].

Collectively, our metabolic reconstructions indicate that both LCCA and LVCCA were facultative anaerobes. As outlined above, these ancestors encoded a range of oxygen-sensitive and oxygen-dependent key metabolic genes, alongside pathways associated with anaerobic and aerobic lifestyles (Figure 2, Extended Data Figure 2 and 5, and Data S9). LVCCA would have evolved around 2 Gya, briefly after the great oxidation event (2.1 - 2.4 Gya)[15], where atmospheric oxygen was far below modern levels[58]. At this time, environments with transient exposure to oxygen and oxic microclines would have been common and comparable to various transition zones found

today (*e.g.*, tidal zones, sediments, animal host tissues *etc.*). Extant facultative anaerobes living in such environments regulate the expression of aerobic and anaerobic metabolism in correspondence with oxygen levels[59]. LVCCA could have likewise lived in an environment with variable oxygen exposure and adjusted its metabolism accordingly. As LVCCA encoded genes for a flagellar apparatus and a type IV pilus (Figure 2) it was likely motile and could transit between oxic and anoxic environments as required.

As a motile facultative anaerobe LVCCA could have already employed a biphasic lifestyle that hinged on oxic-anoxic transitions, rather than host invasion as it does in modern chlamydiae. Indeed, biphasic lifecycles may have more ancient origins within the PVC superphylum, as some Planctomycetes lineages also employ two life stages, a motile stage and a sessile reproductive stage[60]. Adaptation of LCCA to the intracellular eukaryotic niche could have been facilitated by an ancestral biphasic lifecycle, the genomic background of LVCCA, and the gain of endosymbiosis-related genes. Evolution from LVCCA to LCCA may well have hinged on a transition from oxic-anoxic to extracellular-intracellular lifestages. As LCCA likewise encoded flagellar genes (Figure 2 and Data S9-S10), it could have used this motility to facilitate encounters with eukaryotic hosts. The eukaryotic intracellular environment can provide a refuge from oxygen, as some strict anaerobes have been shown to survive and divide within the vacuoles of amoeba even when exposed to high oxygen levels[61]. Thus, a possible scenario that drove the evolution of Chlamydiae was a coinciding increase in oxygen levels with the emergence of a new niche inside early eukaryotic hosts better suited to a facultative anaerobe.

Gene gain facilitated oxygen tolerance within Chlamydiae

Through additional gain and loss events Chlamydiae later diversified into two major groups, G1 and G2, which are divergent in oxygen-related gene content (Extended Data Figure 5). Genes reconstructed in the last common ancestor of G1 (LG1CA) (Figure 3) and descendents are indicative of life stages in anoxic environments. For example, LG1CA gained the arginine deiminase pathway (Extended Data Figure 6), which is also known as anaerobic substrate-level phosphorylation. Likewise, the Anoxychlamydiales and Anoxychlamydiaceae ancestors gained genes, respectively, for transporting soluble ferrous iron (Fe^{2+}), which is primarily present under anoxic conditions[56], and for producing hydrogen using HydA, which is highly oxygen-sensitive (Extended Data Figure 5). Oxygen-utilizing enzymes such as cytochrome bd ubiquinol oxidase

were also lost in Anoxychlamydiaceae (Figure 4a-b and Data S9-S10). In contrast, the last common ancestor of G2 (LG2CA) (Figure 3) lost the oxygen-sensitive PFO[45], and could exclusively use PDC to oxidize pyruvate (Extended Data Figure 6). Particularly striking within G2 was the gain of an extensive suite of genes during Amoebachlamydiales evolution indicative of adaptation to higher-oxygen environments.

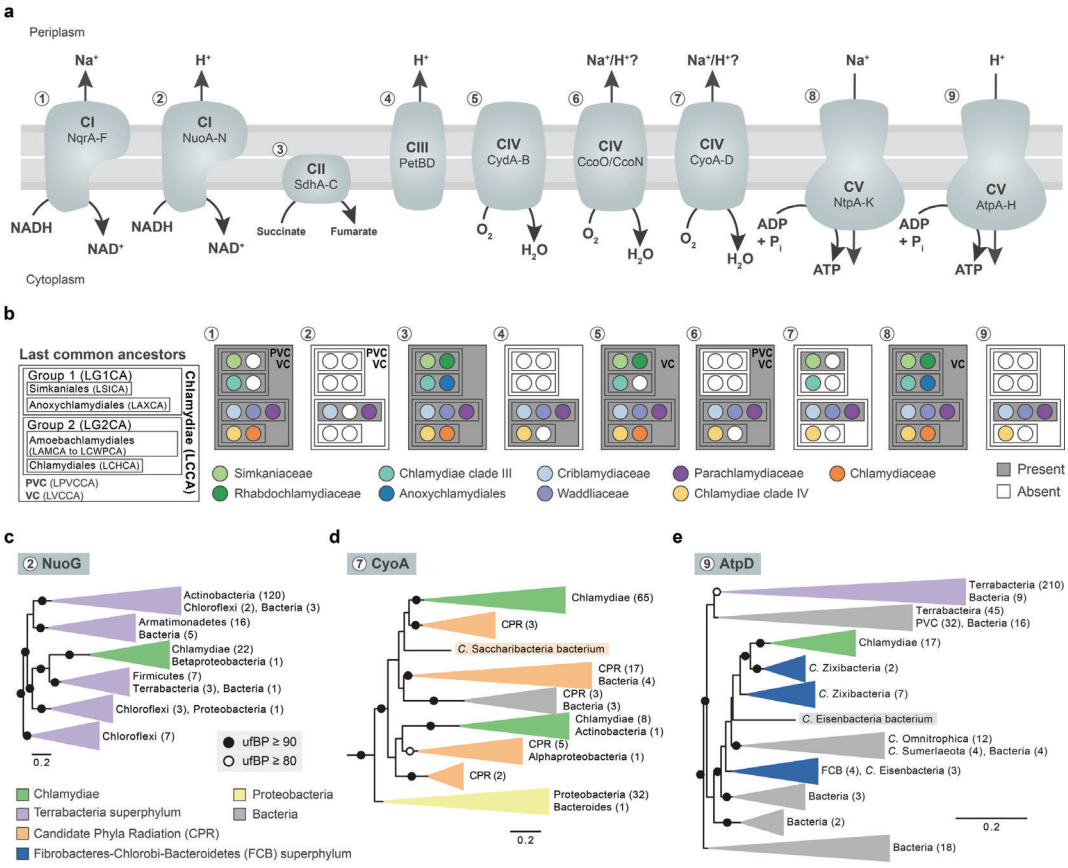


Figure 4. The ancestral chlamydial respiratory chain has undergone expansion with the gain of complexes from diverse bacteria. a Schematic representation of respiratory complexes found in Chlamydiae, alongside relevant substrates, and complex subunits. **b** Presence and absence of these respiratory complexes across key chlamydial ancestors. Presence defined as ≥ 50% of complex subunits. See Data S7 for ancestor abbreviations and Data S9 for subunit annotations and presence across all Chlamydiae nodes. Maximum-likelihood single-protein phylogenies of **c** the NuoG subunit of the proton-transporting NADH dehydrogenase (NuoA-N), **d** the CyoA subunit of cytochrome o ubiquinol oxidase (CyoA-D), and **e** the AtpD subunit of the proton-driven ATP synthase (AtpA-H). Circles summarize bipartitions supported by ultrafast bootstraps (ufBP) inferred using IQ-TREE under the LG model of evolution (see Methods). Collapsed clades are annotated with the taxonomy of members, and coloured according to the legend if most members belong to a given group. Scale bars indicate the number of substitutions per position in the alignment. See also repository files for uncollapsed phylogenies and trees of additional subunits.

Early Amoebachlamydiales ancestors gained genes for aerobiosis-associated pathways and oxygen-dependent enzymes (Extended Data Figure 5 and Data S10). This included the aerobic coproporphyrinogen III oxidase used in heme biosynthesis and iron complex transporters (*e.g.*, siderophores) for transporting ferric (Fe^{3+}) iron, the primary species under oxic conditions[56] (Extended Data Figure 5). Additional genes involved in oxidative stress response were also reconstructed in these ancestors, including catalase, several superoxide dismutases, and a bacterial globin that could be used for nitric oxide detoxification[62] (Extended Data Figure 5). Many early Amoebachlamydiales ancestors also encoded genes for the key enzymes in the glyoxylate shunt, a TCA cycle bypass that is found nearly exclusively in aerobes[63] (Extended Data Figure 5). Amoebachlamydiales ancestors also had a more extensive ETC, which included the terminal oxidase cytochrome bo_3 (CyoA-D) (Figure 4a-b). Cytochrome bo_3 has lower oxygen affinity and is hence associated with higher levels of oxygen[64] than those found across other chlamydiae. In relation, Heme O synthase (CyoE) was also found, which is essential for generating the Heme O cluster in cytochrome bo_3. Cytochrome bo_3 was also reconstructed in several other chlamydial ancestors, including Simkaniales, CC-III and CC-IV (Figure 4a-b). The evolutionary history of CyoA-D in Chlamydiae appears complex, and single-gene phylogenies of these proteins indicate that they were gained by HGT from Candidate Phyla Radiation (CPR) members, potentially in two separate events followed by later inter-chlamydial transfers (Figure 4d, and Data S10). Overall, LG2CA, and in particular Amoebachlamydiales ancestors, were better adapted to oxygen, suggesting oxygen tolerance as a major driving force in chlamydial evolution.

Genome expansion as a mode of evolution in intracellular symbionts

The gain of aerobiosis-related genes in the Amoebachlamydiales was accompanied by other prominent gene gains related to metabolic potential (Extended Data Figure 3), in line with the extended metabolic capabilities of extant members[12,13,65]. These gene gains led to genome expansion during Amoebachlamydiales evolution (Figure 5), with the ancestor of the order having had 1,783 genes, relative to 1,285 and 1,469 genes reconstructed in LG1CA and LG2CA, respectively (Figure 3). Further large gene influxes led to massive genome expansion in Amoebachlamydiales (Figures 5 and S10, and Extended Data Figure 4). The ancestor of the Criblamydiaceae, Waddliaceae, and Parachlamydiaceae (LCWPCA) nearly doubled gene content since LG1CA with 2,766 genes reconstructed (Figure 3). For example, through additive HGT of

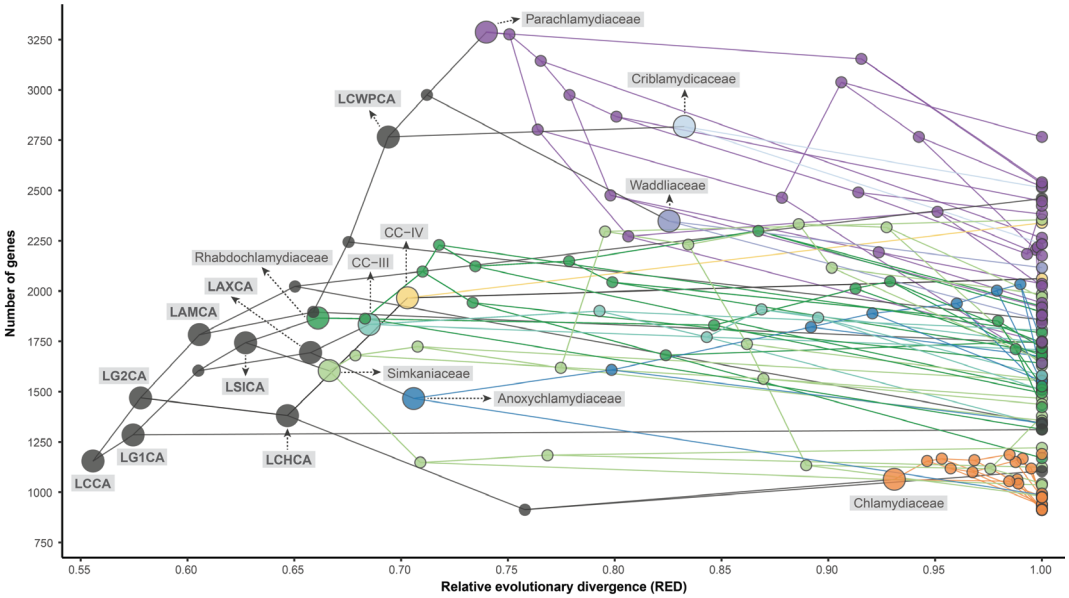


Figure 5. Genome expansion and long-term maintenance of gene content in the phylum Chamydiae. Number of genes inferred across chlamydial ancestral and extant genomes scaled to relative evolutionary divergence (RED) metric[77]. The RED metric provides an approximation of relative time from a given common ancestor (LPVCCA, RED=0) to extant taxa (RED=1). Nodes and branches are coloured by family assignment. Abbreviations of ancestors are as in Figure 3. Key ancestral nodes are labelled and enlarged.

several complexes from different bacterial groups LCWPCA gained an extended ETC with mosaic origins and clear metabolic implications (Figure 4).

LCWPCA gained complexes for generating a PMF alongside a SMF, the former of which has a larger redox gap and can thus be used to generate more ATP[66]. LCWPCA had a proton-transporting NADH dehydrogenase (NuoA-N; CI) and a proton-driven F-type ATP synthase (AtpA-H; CV) (Figure 4a-b and Data S10). These are large multi-subunit complexes and such physiologically-coupled proteins are often gained together as a functional unit[67]. In single-protein phylogenies of NuoA-H subunits chlamydial sequences branch with members of the Terrabacteria superphylum (Figure 4c). For AtpA-H subunits chlamydiae mostly affiliated with Candidatus *Zixibacteria*, a recently described phylum found in groundwater and microbial mats[68] (Figure 4e). These complexes thus appear to have been gained in single events during Amoebachlamydiales evolution. In addition, LG2CA had putative components of a cytochrome bc_1 complex (PetBD; CIII)[69] that was retained in most groups, though subsequently lost in the Chlamydiaceae (Figure 4a-b). The extended metabolic capabilities found in amoeba-infecting members of the Amoebachlamydiales relative to the Chlamydiaceae animal pathogens has been previously

noted[12,13,65]. However, it had been presumed that differences in the ETC were a result of gene loss in the Chlamydiaceae and other groups, with LCCA having retained a more flexible and branched ETC[18,21]. Our analyses instead suggest that ancestors within Amoebachlamydiales gained more comprehensive ETCs after their divergence from LCCA.

Genome evolution in obligate endosymbionts is typically thought to occur through reductive processes[70]. Here, small intracellular population sizes and genetic isolation over time lead to gene loss through genetic drift, the accumulation of slightly deleterious mutations, and a lack of recombination[71]. This stands in stark contrast to our findings of genome expansion within the Amoebachlamydiales, and more generally maintenance of gene content during evolution within the Chlamydiae phylum (Figure 5). Many members of the Parachlamydiaceae and Criblamydiaceae are well-known to infect amoebae (Figure 1), which have been proposed as 'melting-pots' of evolution[72–74]. Amoebal endosymbionts tend to experience less genome reduction due to frequent HGT from co-infecting symbionts and the availability of DNA from microbes consumed by their hosts[72]. Indeed, genome expansion has previously been found in facultative symbionts with free-living lifestages[75], and there are examples of recent gene gain through HGT in obligate symbionts[76]. However, here we have found evidence of continued genome expansion throughout evolutionary time alongside an obligately endosymbiotic lifestyle (Figure 5). Our findings challenge the overarching paradigm of long-term reductive genome evolution in obligate endosymbionts and suggest that genome complexity can be maintained, and even increase, over evolutionary time scales on the order of a billion years.

Gene exchange is common among chlamydiae

In general, PVC protein families evolved largely vertically, with a median of 70% vertical transmission events, closely mirroring reconstructions of gene content evolution in the bacterial domain[76,78]. The remaining 30% represent horizontal gene transmission (Figure S8). Based on their lifestyles it is expected that strict endosymbionts, such as chlamydiae, would be limited in gene exchange[71]. However, we found that a large number of genes originated within chlamydiae through HGT (n=1,458) (Extended Data Figure 4 and Data S10). These HGTs were primarily from diverse bacterial groups, as was the case for 94% of HGT-derived LCCA originations. Many HGT events to Chlamydiae were from groups that include well-known symbionts such as Alphaproteobacteria, Bacteroides, and CPR (Extended Data Figure 4), suggesting that they could have been acquired

during co-infection of eukaryotic hosts. We thus looked at HGTs within the PVC dataset and approximated the rate of gene transfer between ancestors. Overall, Chlamydiae were subject to significantly lower transfer rates than other PVC members (Extended Data Figure 7). However, unexpectedly, the Parachlamydiaceae, Rhabdochlamydiaceae, and CC-III families did not have significantly different transfer rates from the predominantly free-living outgroup PVC members. Indeed, these families had significantly higher transfer rates than Chlamydiaceae (Extended Data Figure 7). An exceptionally large amount of gene transfers were inferred between Amoebachlamydiales members, supporting their patterns of genome expansion (Extended Data Figure 7).

The network of inter-chlamydial HGTs was visualized by testing for statistical overrepresentation of gene transfers between chlamydial nodes where donor and acceptor lineages could be assigned (n=7,056). The resulting transfer network shows highways of gene transfer indicating likely shared environmental niches (Extended Data Figure 7). We observe these highways primarily within chlamydial families, which contribute to two thirds of significant transfers, consistent with genome sequence divergence being one of the major barriers to HGT[73]. Despite this evident barrier, we did observe 455 significant transfers between members of different chlamydial families (Extended Data Figure 7). Such elevated HGT between more distantly related chlamydiae could be explained by an overlap in ecology[79,80], whether in host or environment. Several chlamydial groups were minimal players within the larger inter-chlamydial HGT network, including Chlamydiaceae, Anoxychlamydiaceae, and *Neochlamydia* spp. (Extended Data Figure 7). For these groups we instead observe isolated transfer highways. Interestingly, these chlamydiae have undergone convergent loss of several central metabolic pathways, including TCA cycle and ETC components (Figure 4, Extended Data Figures 5-6, and Data S10), suggesting adaptation to highly specialized environments. HGT is pervasive in bacteria, and metabolic genes in particular are often subject to transfer driven by selective adaptation to changing environments[81,82]. We demonstrate that Chlamydiae appear to be no exception, with metabolic networks shaped by functional gain through HGT in key ancestors. Metabolism found in extant chlamydiae is the result of a complex interplay of gene retention, loss, and HGT events. These results suggest that HGT may play more pervasive roles in the genome evolution of endosymbionts than is recognized currently.

Concluding remarks

Here, we present a more comprehensive picture of evolution in an ancient group of endosymbionts. We found that the ancestor of the Chlamydiae phylum was already adapted to an endosymbiotic lifestyle and had the genetic capabilities to infect eukaryotic hosts. This ancestor evolved from a facultative anaerobe that was able to transition between oxic and anoxic environments, which may have facilitated co-option of a biphasic lifecycle instead based on host invasion. Surprisingly, our results show that, in addition to lineage-specific genome reduction, chlamydiae have undergone genome expansion and rampant HGT during their evolution, counter to expectations for a group of obligate endosymbionts. Our data suggests that the acquisition of genes related to energy metabolism and oxygen tolerance later shaped diversification within the phylum. Together, our results lay a foundation for further investigation of the complex, and perhaps varied, evolutionary trajectories of endosymbiotic organisms.

MATERIALS AND METHODS

See Extended Data Figure 1 for an overview of key steps in the workflow employed for the reconstruction of gene content evolution in PVC bacteria.

Selection of representative genomes

A representative dataset of PVC bacteria genomes was selected using genome quality to obtain species-level Chlamydiae representatives (ingroup), and genus-level non-Chlamydiae PVC representatives (outgroup). This dataset was selected from the genome taxonomy database (GTDB) and NCBI. GTDB is continually updated as genomes are released on NCBI and thus naming structures are non-stationary[77]. Here, Chlamydiae were initially classified as a phylum, but in the version used were classified as a class of Verrucomicrobiota (*i.e.*, Chlamydiia). More recently Chlamydiae was again classified as a phyla, but with the name “Verrucomicrobiota_A”. All genomes from GTDB[77] v86 (2018 database) classified as part of the phyla Planctomycetota and Verrucomicrobiota were selected (1,183 genomes). Non-chlamydial PVC genomes (outgroup, 773 genomes, Data S1) with a completeness ≥ 90% and contamination ≤ 2%, based on the metadata provided by GTDB, were downloaded from NCBI (182 genomes; April 3rd, 2019). For the Chlamydiae ingroup, genomes from the GTDB class "c__Chlamydiia" were downloaded from NCBI (410 genomes; April 3rd, 2019) and supplemented with newly acquired MAGs and isolates

(216 genomes) for a total of 626 chlamydial genomes. We used miComplete[83] 1.1.1 to estimate the quality of these ingroup genomes using a chlamydiae specific marker gene set[22] to select those with completeness ≥ 0.9 and redundancy ≤ 1.02 for downstream analysis (460 genomes, Data S1).

To reduce dataset redundancy, all genomes were de-replicated with dRep[84] v1.4.3 using cut-offs previously proposed for strain-level delineation[85], *i.e.* an average nucleotide identity (ANI) cutoff of 96.5% and a genome alignment fraction of at least 60% (resulting in 224 genomes; Figure S1 and Data S1). Outgroup genomes were further dereplicated on the GTDB genus level by comparing genome quality scores (GQS) per genus (Data S1). We calculated GQS for this step as described in[86], *i.e.* GQS = Completeness (%) - 5 * Contamination (%). The genome per genus with the highest GQS was selected as a representative, and when two genomes had an equal score one was manually selected (Figure S1). This resulted in a final dataset of 184 PVC genomes, with 95 species-level Chlamydiae representatives and 89 genus-level non-Chlamydiae PVC representatives (47 Planctomycetes, 34 Verrucomicrobia, 5 Lentisphaerae, and 3 Kiritimatiellaeota) (Figure S1 and Data S2). Genome characteristics were calculated using miComplete[83] 1.1.1 (Figure 1 and Data S2). Putative uncharacterized PVC-phyla were not included, alongside the phylum *Candidatus* Omnitrophica[3], due to its conflicting position in recent large-scale species trees of Bacteria[2,86].

Phylogenomic analyses

PVC species relationships were inferred using phylogenomic datasets of concatenated single-copy marker proteins (Figure S2 and Data S3) for the initial 184 selected taxa (Figure S3 and Data S2). Additional species phylogenies were inferred after the removal of genomes with unresolved phylogenetic positions resulting in datasets with 183 taxa (removal of *Chlamydiae bacterium* 1070360-7; Figure S4), and 180 taxa (further removal of Parilichlamydiaceae genomes; Figure S5). ML and Bayesian phylogenies were inferred with and without the removal of compositionally heterogeneous sites for all three datasets (184, 183, and 180 taxa) as outlined below (Data S4-S6). Species phylogenies were rooted with Planctomycetes based on its phylogenetic position in recent large-scale phylogenomic analyses of bacterial species relationships[86–88]. The 180 taxa dataset was selected for further analyses and the converged Bayesian species phylogeny (chains 1 and 3) with compositionally heterogeneous sites removed was used for ancestral state reconstruction (Figure 1).

Identification of single-copy marker proteins

Candidate single-copy marker genes were identified using non-supervised orthologous groups (NOGs) from eggNOG[89] v4.5.1. Protein sequences from all PVC bacteria genomes were mapped to NOGs at the Last Universal Common Ancestor (LUCA) level (*i.e.*, root-level "-d NOG") using emapper[90] v1.0.1. Resulting NOGs where 95% of taxa were found in a single-copy were identified as candidate markers for further investigation (116 NOGs; Data S3).

Sequences from each protein family NOG were aligned using MAFFT L-INS-i[91] v7.427 and then manually inspected, with poorly aligned and short sequences removed. Alignments were trimmed using BMGE[92] v1.12 (entropy score cutoff or "-h" of 0.6). IQ-TREE[93] v1.6.11 was used to infer phylogenetic trees, with model selection using ModelFinder[94] from empirical profile mixture models[95] combined with the LG exchangeability matrix[96] (*i.e.*, LG+C10 to LG+C60), and with 1000 ultrafast bootstraps (ufBP)[97]. Resulting trees (available in repository) were manually examined for patterns indicative of vertical inheritance and sufficient phylogenetic signal, and markers were removed that did not generally resolve PVC phyla (Data S3). Sequences were removed that could represent HGT events, distant paralogs, or contamination (Data S3). Where multiple sequences per taxon were present, if they overlapped both were removed (duplicates), and if they were partial and non-overlapping the longer sequence was retained (Data S3). A second round of sequence alignment and tree inference was performed as above, and further markers removed resulting in 79 marker proteins (Data S3).

Discordance filtering[98] was then performed to remove markers with the most anomalous phylogenetic signal relative to the majority (*i.e.*, the most discordant trees). NOGs (including COGs) were ranked by discordance score and the top scoring fraction was removed, leaving 74 single-copy marker proteins for phylogenomic analyses (Figure S2). Proteins for each selected marker were re-aligned and trimmed, as above, after the removal of taxa with unresolved phylogenetic positions (*i.e.*, datasets with 183 and 180 taxa). Trimmed protein alignments were concatenated into a supermatrix alignment for each of the three datasets.

Phylogenomic inferences

Heterogenous site removal was performed using χ2-trimming[99], with the most compositionally heterogeneous sites removed from each supermatrix alignment in incremental steps of 1%. Site removal continued until no taxa significantly heterogeneous in their amino acid composition remained (based on χ2-test score statistic; Figures S3-S5 and Data S4).

462 Using IQ-TREE[93] v1.6.10 with model selection[94], ML phylogenies were inferred for the initial
463 unrefined alignment, for alignments in 10% increments of total sites removed based on χ2-
464 trimming (up to 50%; Data S4), and for the alignment with no significantly heterogeneous taxa
465 (fully refined alignment). Following, ML trees were then reconstructed using the PMSF
466 approximation of the LG+C60+F+Γ4 model (selected in all initial trees) with 100 non-parametric
467 bootstraps (BP). Transfer bootstrap expectation (TBE)[100] bootstraps were also inferred for the
468 initial unrefined alignment and for the fully refined alignment using IQ-TREE[101] v2.0.
469 Bayesian phylogenies were also reconstructed for these two alignments for all three taxa
470 datasets. In each case, four independent MCMC chains were run using PhyloBayes-MPI v1.7b[102]
471 with the CAT+GTR+Γ4 model[95,103], for at least 10,000 iterations. CAT, a site-heterogeneous
472 model, performs more robustly against LBA artefacts[104], but is not feasible for use outside of a
473 Bayesian framework. If at least 10,000 iterations had been run but no chains had begun to converge
474 (maxdiff < 1), all chains were stopped. The number of generations, burn-in, and any chain
475 convergence (maxdiff < 0.3) can be found in Figures S3-S5 alongside a consensus tree of all 4
476 chains with posterior probability (P) indicating branch support. Posterior predictive checks were
477 also performed with PhyloBayes-MPI v1.7b[102], with configurations sampled every 10 generations
478 after burn-in. The resulting range of Z-scores for maximum heterogeneity and diversity across
479 chains can be found in Figures S3-S5. See Data S6 for all uncollapsed species phylogenies and
480 Data S5 for species phylogeny summary with the number of taxa, alignment lengths, inference
481 methods, bootstrap supports, and the model of evolution used.

16S rRNA gene species phylogeny

482 Near-full length 16S rRNA gene sequences from chlamydiae (n=233) and other PVC members
483 (n=205) were downloaded from SILVA[105] v138 SSU Ref NR 99, in addition to 79 near-full length
484 chlamydial 16S rRNA gene sequences (97% identity OTU representatives) from the study of
485 Schulz et al.[9] and 142 sequences from our reference genome dataset. The resulting 659 16S rRNA
486 gene sequences were clustered at 90% sequence identity to reduce redundancy using USEARCH[106]
487 v11.0.667 with "-cluster_smallmem". The resulting 177 family level sequence representatives
488 were aligned with SINA[107] and the alignment trimmed with trimAl[108] v1.4.1 "-gappyout" (1,533
489 aligned positions). Bayesian tree samples with four MCMC chains in parallel (n=100,000 each)
490 were inferred under the CAT+GTR+Γ4 model[95,103] in PhyloBayes v4.1c[109] (Figure S6).

Convergence was assumed once the discrepancies in bipartition frequencies dropped below 0.3 and the effective sample sizes for continuous parameters were greater than 100 (according to the bpcomp and tracecomp commands in PhyloBayes, respectively) after burn-in (n=25,000).

Generation of protein families and trees

NOG clustering

PVC protein sequences from the 180 taxa dataset (n=445,591) were mapped against eggNOG[89] v4.5.1 using emapper[90] v1.0.1 at the root-level "-d NOG" database. Of these, 326,083 (73%) protein sequences were assigned to 17,935 NOGs (Figure S7).

De-novo clustering

For the remaining 119,508 protein sequences in the 180 taxa dataset, with no homologs in eggNOG v4.5.1, we performed pairwise sequence alignment in an all-against-all fashion with DIAMOND[110] v0.9.21 using the "--more-sensitive" parameter. Subsequently, *de novo* clustering with SiLiX[111] v1.2.9 was performed with default overlap of 80% and identity thresholds ranging from 5% to 40% in 5% increments (Figure S7). To select an appropriate identity threshold, we inspected (i) the number of singleton clusters per threshold and (ii) assigned TIGRFAM[112] v15.0 domains with interproscan[113] v5.36-75.0 to protein sequences. Using the assigned TIGRFAMs the true positive rate (sensitivity) and true negative rate (specificity) were calculated for clusters, with the different clusterings evaluated using the balanced accuracy measure ((specificity + sensitivity) / 2) as suggested[111]. A 25% identity cutoff performed best, yielding 10,548 *de novo* gene families with at least two members (75,218 singletons).

Gene family phylogenetic trees

We performed phylogenetic analysis on all gene families (both NOG and *de-novo* clusters) with at least 4 members (n=11,996), except COG3119 and COG0457 due to computational infeasibility. Sequences were aligned with MAFFT[114] v4.427 using the "--localpair" strategy. Alignments were then trimmed using BMGE[92] v1.12 with default parameters and an entropy cutoff of 0.6. The allowed gap rate for alignment positions was increased to 0.5 for 94 gene families that had less than 50 informative aligned positions using the initial parameters. Gene trees were then inferred with IQ-TREE[93] v1.6.11 using the best fit model identified by modelfinder[94], with"-m TESTNEW", "-madd LG+C10, LG+C20, LG+C30, LG+C40, LG+C50, LG+C60", and 1,000

519 improved ultrafast bootstraps[97] "-bnni". We created dummy phylogenetic trees for the remaining
520 91,705 gene families with less than 4 members.
521 *Annotation of gene families*
522 We assigned protein domain annotations by using InterProScan[113] v5.36-75.0 to identify Protein
523 Families (Pfam)[115], TIGRFAM[112] and InterPro (IPR) domains[116]. We assigned KEGG orthology
524 (KO) and enzyme commision (EC) numbers using GhostKOALA[117] and inferred EggNOG[89]
525 functional annotation as described above, and also at the bacterial-level ("-d BACT").

Ancestral state reconstruction

526 *Gene-tree unaware method - Count*
527 For gene tree unaware ancestral gene content reconstruction, we ran Count[118] v10.04 with the gain-
528 loss-duplication model of evolution with a poisson distribution to model protein family size at the
529 root. We used the same gain-loss and duplication loss ratios for all lineages and inferred ancestral
530 gene content using the Wagner maximum parsimony framework with default costs.
531 *Gene-tree aware method - ALE*
532 ML trees of protein families identified in the PVC dataset were reconciled with the species tree to
533 reconstruct their gene family histories. We computed conditional clade probabilities from the
534 bootstrap samples (ALEobserve) and sampled 100 reconciliations with the species tree
535 (ALEml_undated) using ALE[29] v0.449, implemented in a nextflow[119] pipeline
536 (https://github.com/maxemil/ALE-pipeline). We added singleton clusters as originations at the
537 corresponding species node to the reconstructions.
538 *Comparison of ancestral state reconstruction methods and selection of ALE cut-off*
539 ALE improves on earlier methods by directly estimating rates of gene duplication, transfer, and
540 loss from the data as well as incorporating the uncertainty in gene trees while exploring a larger
541 gene tree space[29]. The accuracy of reconstructions can be negatively influenced by an inaccurate
542 species tree and imbalanced taxon sampling[29,30]. Here, these risks are minimized due to our
543 extensive taxon sampling and species tree reconstruction efforts (Figures S1-S6 and Data S1-S6).
544 ALE reports relative frequencies for ancestral events and protein family copy numbers that
545 express their statistical support. This support accumulates the uncertainty introduced by alignment,
546 tree reconstruction, and reconciliation and should therefore not be set at a standard level cutoff.
547 We therefore aimed to identify a suitable threshold by investigating the density distribution per

548 inferred event type and transfer ratio per protein family (Figure S8), which indicated a cutoff of
549 0.3 as a candidate with high signal to noise ratio. This identified cutoff is in accordance with recent
550 similar analyses that selected 0.3 as a frequency cutoff[120,121].
551 We further compared the tree aware reconstructions generated by ALEml with the thresholds
552 0.3 (sensitive), 0.5 (specific), and 0.7 (very specific) with the gene content only aware Count
553 reconstructions (Figure S9). In agreement with the event density distribution analysis and a, the
554 highest consensus between gene tree aware and unaware methods was reached with a threshold of
555 0.3, *i.e.* events (loss, transfer, origination or duplication) and presence/absence (copies) were
556 counted as such if they had a frequency of at least 0.3. In summary, based on this extensive
557 investigation of incremental frequency cutoffs and additional comparison with a gene tree unaware
558 method, we identified a frequency cutoff of 0.3 for inferring evolutionary events in our ASR
559 analysis (Figures S8-S9).

Inference of transfer rates and gene transfer highways

560 We approximated the rate of intra PVC HGT (*i.e.*, transfer rate) for nodes in the species tree by
561 calculating the inferred gene transfers in our reconstructions divided by the number of substitutions
562 in the species tree along the given branch. Based on the shared protein families between two extant
563 or ancestral genomes, we tested whether more HGT events occured between genomes than the
564 median of transferred protein families within chlamydiae members. We used a one-sided binomial
565 test “binom.test” with “alternative = "greater"” in the R base package[122] and false discovery rate
566 corrected p-values for multiple testing with “p.adjust” to identify enriched transfer routes (“gene
567 transfer highways”) with a p-value ≤ 0.05. Significant gene transfer highways were visualized with
568 Cytoscape[123] v3.7.0.

Identification of non-PVC gene transfer donors

569 To distinguish bona fide HGTs from outside the PVC dataset and candidate *de novo* gene families
570 for chlamydial originations we performed a homology search against the NCBI non-redundant
571 (NR) database. If no homologous proteins could be identified, gene families were referred to as *de*
572 *novo* candidates, otherwise we inferred gene trees to identify potential donor lineages of the
573 horizontally transferred gene. All gene families inferred as originations within Chlamydiae were
574 collected and analyzed using a Snakemake workflow[124] to identify the putative HGT donor group

575 of each gene (https://github.com/jennahd/HGT_trees). For each gene family a DIAMOND
576 v0.9.36.137 blastp[110] search (with "max-target-seqs 2000" and "more-sensitive") was performed
577 using all sequences against NCBI's NR database[125] (v5 accessed October 8th, 2020). Unique hits
578 per gene family were compiled and redundancy reduced using CD-HIT[126] v4.8.1 at 80% sequence
579 identity. NCBI's taxonomy database[127,128] was used for taxonomic classification. Protein
580 sequences from each gene family and any database hits were aligned with MAFFT[91] v7.471 ("--
581 auto") and trimmed with trimAl[108] v1.4.rev15 ("gappyout"). Sequences which covered less than
582 40% of the trimmed alignment were removed, followed by inference of an initial phylogenetic tree
583 using FastTree[129] 2 v2.1.11. Long-branching taxa were identified as having outlier branch lengths
584 (the 3rd quartile + 1.5 x the interquartile range) relative to others in each tree and were removed,
585 before re-inferring trees as above.
586 These initial trees were prohibitively large for performing ML analyses. Smaller subtrees were
587 therefore selected, using the above workflow, that included the chlamydial sequences of interest.
588 Here, nodes in the larger tree were identified whose descendants were composed of at least 25%
589 chlamydial sequences (> 1). To account for multiple HGT events to chlamydiae, up to three nodes
590 (with the largest number of chlamydial sequences) were identified per gene family (though in the
591 majority of cases only one was found). Larger subtrees including this node were identified by
592 finding branches at least 3 further up in the tree hierarchy that included at least 150 additional taxa
593 and up to 400 additional taxa with bipartition support ≥ 0.7. Where a subtree with these conditions
594 was identified, but with a larger number of taxa, the number of taxa was reduced to ≤ 400 by
595 removing more distant sequences (support ≥ 0.7, ≥ 5 sequences, and at least 6 steps until a common
596 ancestor with the chlamydial node). Between 20 and 50 outgroup sequences were randomly
597 selected from the clade with a position sister to the selected subtree (moving to the next subtending
598 clade with < 20 outgroup sequences). Selected subtrees were subsequently aligned, trimmed, and
599 an initial phylogenetic tree inferred as above. ML trees were then inferred for each subtree using
600 the trimmed alignment with IQ-TREE[93] v1.6.12 under the LG model of evolution[96], with 1000
601 ufBP. The clade sister to chlamydial sequences, and that subtending this clade ("nested") with
602 ufBP ≥ 80 were identified. Taxonomic labels of sister and nested taxa were each compared at
603 domain, superphylum, and phylum levels. The lowest-level shared taxonomy of sister taxa at
604 cutoffs of 50, 75, 90, and 100% of taxa was selected as the putative gene transfer donor (Data S10).
605 Originations were identified as *de-novo* in the case where no non-chlamydial hits were found or

606 when no sister group to chlamydial sequences was inferred.

Reconstruction of metabolic pathways

607 The evolutionary history of ETC components was also investigated using the above described
608 workflow. We reconstructed ancestral gene family repertoires from ALE by selecting all families
609 predicted to be present at a given node with $P \geq 0.3$. We assessed the metabolic capabilities of
610 ancestral genomes using the KEGG Module tool[130] or MetaCyc pathways[131].

Statistics and data visualization

611 Phylogenetic trees and protein domains were visualized using Figtree v1.4.4
612 (http://tree.bio.ed.ac.uk/software/figtree/), iTOL[131], and the ETE3 Toolkit[132]. Relative
613 evolutionary divergence (RED) of chlamydiae ancestors in the species tree was calculated using
614 PhyloRank[77] v0.1.10 (https://github.com/dparks1134/PhyloRank/). Plots were generated using
615 Cytoscape[123] v3.7.0 and the R v3.6.2 base package[122] alongside the packages ggplot2[133] v3.3.3,
616 ggtree[134] v2.5.0.991, and treeio[135] v1.10.0.

Data and code availability

617 Additional raw data files are hosted on the online repository figshare
618 (https://figshare.com/s/46e0aa5743b9d84fc78d). These include sequences, alignments, trimmed
619 alignments, and trees for single-copy marker proteins used for species phylogenies (both those
620 selected and not selected), the 16S rRNA gene, and concatenated supermatrix alignments and trees
621 for all three species datasets (of 184, 183, and 180 taxa). Both NOG and *de novo* protein clusters
622 used for the ancestral state reconstruction are also provided alongside alignments, trimmed
623 alignments, trees, and bootstrap trees (ufboot) provided to ALE. The raw ALE results with all
624 events are also included, alongside cluster annotations together with events, and events for each
625 cluster mapped to the species tree. Sequence datasets, alignments and trees inferred as part of the
626 analysis to determine HGT donors for chlamydiae protein originations are likewise provided. In
627 addition, pdfs of metabolic reconstructions of LVCCA, LG1CA, and LG2CA can be found in the
628 repository files. The Snakemake workflow to identify putative HGT donor groups is available on
629 GitHub (https://github.com/jennahd/HGT_trees).

ACKNOWLEDGEMENTS

We want to thank the IMG/M Data Consortium for contributing metagenomic data, and Lauren Alteo and Jeffrey Blanchard for early access to Harvard forest soil MAGs. This project has received funding from the European Research Council ERC (EVOCHLAMY, grant no. 281633, ERC Starting grant no. 310039, and Consolidator grant no. 817834), the Austrian Science Fund FWF (FunChlam, grant no. P32112, doc.funds MAINTAIN, grant no. DOC 69-B), and the Swedish Research Council (VR grant no. 2015-04959). Computational resources were provided by the Life Science Compute Cluster (LiSC; http://cube.univie.ac.at/lisc) and the Swedish National Infrastructure for Computing (SNIC) at UPPMAX with the project number SNIC 2020/15-158, and at PDC with the project numbers SNIC 2019/3-474 and SNIC 2020/5-473.

AUTHOR CONTRIBUTIONS

Conceptualization: T.J.G.E., M.H., J.E.D., and S.K. Data curation: S.K. and J.E.D. Formal analysis: J.E.D., S.K., A.C., and M.E.S. Investigation: S.K. and J.E.D. Validation: all authors. Resources: M.H. and T.J.G.E. Supervision: T.J.G.E. and M.H. Visualization: J.E.D. and S.K. Writing—original draft: S.K. and J.E.D. Writing—reviewing and editing: all authors. The contributions of J.E.D. and S.K., and T.J.G.E. and M.H. should be regarded as equal, respectively.

COMPETING INTERESTS

The authors declare that they have no competing interests.

REFERENCES

1. Rivas-Marín, E. & Devos, D. P. The Paradigms They Are a-Changin': past, present and future of PVC bacteria research. *Antonie van Leeuwenhoek* **111**, 785–799 (2018).
2. Collingro, A., Köstlbacher, S. & Horn, M. Chlamydiae in the Environment. *Trends Microbiol.* **28**, 877–888 (2020).
3. Wagner, M. & Horn, M. The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance. *Curr. Opin. Biotechnol.* **17**, 241–249 (2006).
4. van Niftrik, L. & Devos, D. P. Editorial: Planctomycetes-Verrucomicrobia-Chlamydiae Bacterial Superphylum: New Model Organisms for Evolutionary Cell Biology. *Front. Microbiol.* **8**, 1458 (2017).
5. Elwell, C., Mirrashidi, K. & Engel, J. Chlamydia cell biology and pathogenesis. *Nat. Rev. Microbiol.* **14**, 385–400 (2016).
6. Bachmann, N. L., Polkinghorne, A. & Timms, P. Chlamydia genomics: providing novel

insights into chlamydial biology. *Trends Microbiol.* **22**, 464–472 (2014).
7. Borel, N., Polkinghorne, A. & Pospischil, A. A Review on Chlamydial Diseases in Animals: Still a Challenge for Pathologists? *Vet. Pathol.* **55**, 374–390 (2018).
8. Lagkouvardos, I. *et al.* Integrating metagenomic and amplicon databases to resolve the phylogenetic and ecological diversity of the Chlamydiae. *ISME J.* **8**, 115–125 (2014).
9. Schulz, F. *et al.* Towards a balanced view of the bacterial tree of life. *Microbiome* **5**, 140 (2017).
10. Horn, M. Chlamydiae as Symbionts in Eukaryotes. *Annu. Rev. Microbiol.* **62**, 113–131 (2008).
11. Taylor-Brown, A., Vaughan, L., Greub, G., Timms, P. & Polkinghorne, A. Twenty years of research into Chlamydia-like organisms: a revolution in our understanding of the biology and pathogenicity of members of the phylum Chlamydiae. *Pathog. Dis.* **73**, 1–15 (2015).
12. Collingro, A. *et al.* Unity in Variety—The Pan-Genome of the Chlamydiae. *Mol. Biol. Evol.* **28**, 3253–3270 (2011).
13. Omsland, A., Sixt, B. S., Horn, M. & Hackstadt, T. Chlamydial metabolism revisited: interspecies metabolic variability and developmental stage-specific physiologic activities. *FEMS Microbiol. Rev.* **38**, 779–801 (2014).
14. Kamneva, O. K., Knight, S. J., Liberles, D. A. & Ward, N. L. Analysis of Genome Content Evolution in PVC Bacterial Super-Phylum: Assessment of Candidate Genes Associated with Cellular Organization and Lifestyle. *Genome Biol. Evol.* **4**, 1375–1390 (2012).
15. Betts, H. C. *et al.* Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nat Ecol Evol* **2**, 1556–1562 (2018).
16. Greub, G. & Raoult, D. History of the ADP/ATP-Translocase-Encoding Gene, a Parasitism Gene Transferred from a Chlamydiales Ancestor to Plants 1 Billion Years Ago. *Appl. Environ. Microbiol.* **69**, 5530–5535 (2003).
17. Horn, M. *et al.* Illuminating the evolutionary history of chlamydiae. *Science* **304**, 728–730 (2004).
18. Subtil, A., Collingro, A. & Horn, M. Tracing the primordial Chlamydiae: extinct parasites of plants? *Trends Plant Sci.* **19**, 36–43 (2014).
19. Taylor-Brown, A., Spang, L., Borel, N. & Polkinghorne, A. Culture-independent metagenomics supports discovery of uncultivable bacteria within the genus Chlamydia. *Sci. Rep.* **7**, 10661 (2017).
20. Taylor-Brown, A., Madden, D. & Polkinghorne, A. Culture-independent approaches to chlamydial genomics. *Microb Genom* **4**, (2018).
21. Pillonel, T., Bertelli, C. & Greub, G. Environmental Metagenomic Assemblies Reveal Seven New Highly Divergent Chlamydial Lineages and Hallmarks of a Conserved Intracellular Lifestyle. *Front. Microbiol.* **9**, 79 (2018).
22. Dharamshi, J. E. *et al.* Marine Sediments Illuminate Chlamydiae Diversity and Evolution. *Curr. Biol.* **30**, 1032–1048.e7 (2020).
23. Collingro, A. *et al.* Unexpected genomic features in widespread intracellular bacteria: evidence for motility of marine chlamydiae. *ISME J.* **11**, 2334–2344 (2017).
24. Köstlbacher, S. *et al.* Pangenomics reveals alternative environmental lifestyles among chlamydiae. *Nat. Commun.* **12**, 4021 (2021).
25. Stairs, C. W. *et al.* Chlamydial contribution to anaerobic metabolism during eukaryotic evolution. *Sci Adv* **6**, eabb7258 (2020).
26. Bergsten, J. A review of long-branch attraction. *Cladistics* **21**, 163–193 (2005).

27. Gupta, R. S., Naushad, S., Chokshi, C., Griffiths, E. & Adeolu, M. 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**, 765–781 (2015).
28. Köstlbacher, S., Collingro, A., Halter, T., Domman, D. & Horn, M. Coevolving Plasmids Drive Gene Flow and Genome Plasticity in Host-Associated Intracellular Bacteria. *Curr. Biol.* **31**, 346–357.e3 (2021).
29. Szöllősi, G. J., Davín, A. A., Tannier, E., Daubin, V. & Boussau, B. Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140335 (2015).
30. Szöllősi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E. & Daubin, V. Efficient Exploration of the Space of Reconciled Gene Trees. *Syst. Biol.* **62**, 901–912 (2013).
31. Szöllősi, G. J., Tannier, E., Lartillot, N. & Daubin, V. Lateral Gene Transfer from the Dead. *Syst. Biol.* **62**, 386–397 (2013).
32. Peters, J., Wilson, D. P., Myers, G., Timms, P. & Bavoil, P. M. Type III secretion à la Chlamydia. *Trends Microbiol.* **15**, 241–251 (2007).
33. Abby, S. S. & Rocha, E. P. C. The Non-Flagellar Type III Secretion System Evolved from the Bacterial Flagellum and Diversified into Host-Cell Adapted Systems. *PLoS Genet.* **8**, e1002983 (2012).
34. Schmitz-Esser, S. *et al.* ATP/ADP Translocases: a Common Feature of Obligate Intracellular Amoebal Symbionts Related to Chlamydiae and Rickettsiae. *J. Bacteriol.* **186**, 683–691 (2004).
35. Haferkamp, I. *et al.* Tapping the nucleotide pool of the host: novel nucleotide carrier proteins of Protochlamydia amoebophila. *Mol. Microbiol.* **60**, 1534–1545 (2006).
36. Tjaden, J. *et al.* Two Nucleotide Transport Proteins in Chlamydia trachomatis, One for Net Nucleoside Triphosphate Uptake and the Other for Transport of Energy. *J. Bacteriol.* **181**, 1196–1202 (1999).
37. Hackstadt, T., Todd, W. J. & Caldwell, H. D. Disulfide-mediated interactions of the chlamydial major outer membrane protein: role in the differentiation of chlamydiae? *J. Bacteriol.* **161**, 25–31 (1985).
38. Aistleitner, K. *et al.* Conserved features and major differences in the outer membrane protein composition of chlamydiae. *Environ. Microbiol.* **17**, 1397–1413 (2015).
39. Christensen, S. *et al.* Oxidoreductase disulfide bond proteins DsbA and DsbB form an active redox pair in Chlamydia trachomatis, a bacterium with disulfide dependent infection and development. *PLoS One* **14**, e0222595 (2019).
40. Gehre, L. *et al.* Sequestration of host metabolism by an intracellular pathogen. *Elife* **5**, e12552 (2016).
41. Colpaert, M. *et al.* Conservation of the glycogen metabolism pathway underlines a pivotal function of storage polysaccharides in Chlamydiae. *Commun Biol* **4**, 296 (2021).
42. Rosario, C. J., Hanson, B. R. & Tan, M. The transcriptional repressor EUO regulates both subsets of Chlamydia late genes. *Mol. Microbiol.* **94**, 888–897 (2014).
43. Rosario, C. J. & Tan, M. The early gene product EUO is a transcriptional repressor that selectively regulates promoters of Chlamydia late genes. *Mol. Microbiol.* **84**, 1097–1107 (2012).
44. Passalacqua, K. D., Charbonneau, M.-E. & O'Riordan, M. X. D. Bacterial Metabolism Shapes the Host-Pathogen Interface. *Microbiol Spectr* **4**, (2016).

45. Chabrière, E. *et al.* Crystal structures of the key anaerobic enzyme pyruvate:ferredoxin oxidoreductase, free and in complex with pyruvate. *Nat. Struct. Biol.* **6**, 182–190 (1999).
46. Giuffrè, A., Borisov, V. B., Arese, M., Sarti, P. & Forte, E. Cytochrome bd oxidase and bacterial tolerance to oxidative and nitrosative stress. *Biochim. Biophys. Acta* **1837**, 1178–1187 (2014).
47. Buschmann, S. *et al.* The Structure of cbb3 Cytochrome Oxidase Provides Insights into Proton Pumping. *Science* **329**, 327–330 (2010).
48. Liang, P. *et al.* Dynamic energy dependency of Chlamydia trachomatis on host cell metabolism during intracellular growth: Role of sodium-based energetics in chlamydial ATP generation. *J. Biol. Chem.* **293**, 510–522 (2018).
49. Kuhns, M., Trifunović, D., Huber, H. & Müller, V. The Rnf complex is a Na+ coupled respiratory enzyme in a fermenting bacterium, Thermotoga maritima. *Commun Biol* **3**, 431 (2020).
50. Zhang, W., Wong, K. K., Magliozzo, R. S. & Kozarich, J. W. Inactivation of Pyruvate Formate-Lyase by Dioxygen: Defining the Mechanistic Interplay of Glycine 734 and Cysteine 419 by Rapid Freeze-Quench EPR. *Biochemistry* **40**, 4123–4130 (2001).
51. Erbes, D. L., King, D. & Gibbs, M. Inactivation of Hydrogenase in Cell-free Extracts and Whole Cells of Chlamydomonas reinhardi by Oxygen. *Plant Physiol.* **63**, 1138–1142 (1979).
52. Torrents, E. Ribonucleotide reductases: essential enzymes for bacterial life. *Front. Cell. Infect. Microbiol.* **4**, 52 (2014).
53. Jiang, W. *et al.* A Manganese(IV)/Iron(III) Cofactor in Chlamydia trachomatis Ribonucleotide Reductase. *Science* **316**, 1188–1191 (2007).
54. Zhao, X. & Drlica, K. Reactive oxygen species and the bacterial response to lethal stress. *Curr. Opin. Microbiol.* **21**, 1–6 (2014).
55. Gadda, G. & Francis, K. Nitronate monooxygenase, a model for anionic flavin semiquinone intermediates in oxidative catalysis. *Arch. Biochem. Biophys.* **493**, 53–61 (2010).
56. Lau, C. K. Y., Krewulak, K. D. & Vogel, H. J. Bacterial ferrous iron transport: the Feo system. *FEMS Microbiol. Rev.* **40**, 273–298 (2016).
57. Miethke, M. Molecular strategies of microbial iron assimilation: from high-affinity complexes to cofactor assembly systems. *Metallomics* **5**, 15–28 (2013).
58. Lyons, T. W., Reinhard, C. T. & Planavsky, N. J. The rise of oxygen in Earth's early ocean and atmosphere. *Nature* **506**, 307–315 (2014).
59. Sawers, G. & Böck, A. Anaerobic regulation of pyruvate formate-lyase from Escherichia coli K-12. *J. Bacteriol.* **170**, 5330–5336 (1988).
60. Wiegand, S., Jogler, M. & Jogler, C. On the maverick Planctomycetes. *FEMS Microbiol. Rev.* **42**, 739–760 (2018).
61. Tomov, A. T., Tsvetkova, E. D., Tomova, I. A., Michailova, L. I. & Kassovski, V. K. Persistence and Multiplication of Obligate Anaerobe Bacteria in Amebae Under Aerobic Conditions. *Anaerobe* **5**, 19–23 (1999).
62. Vinogradov, S. N., Tinajero-Trejo, M., Poole, R. K. & Hoogewijs, D. Bacterial and archaeal globins — A revised perspective. *Biochim. Biophys. Acta* **1834**, 1789–1800 (2013).
63. Ahn, S., Jung, J., Jang, I.-A., Madsen, E. L. & Park, W. Role of Glyoxylate Shunt in Oxidative Stress Response. *J. Biol. Chem.* **291**, 11928–11938 (2016).
64. Degli Esposti, M., Mentel, M., Martin, W. & Sousa, F. L. Oxygen Reductases in Alphaproteobacterial Genomes: Physiological Evolution From Low to High Oxygen

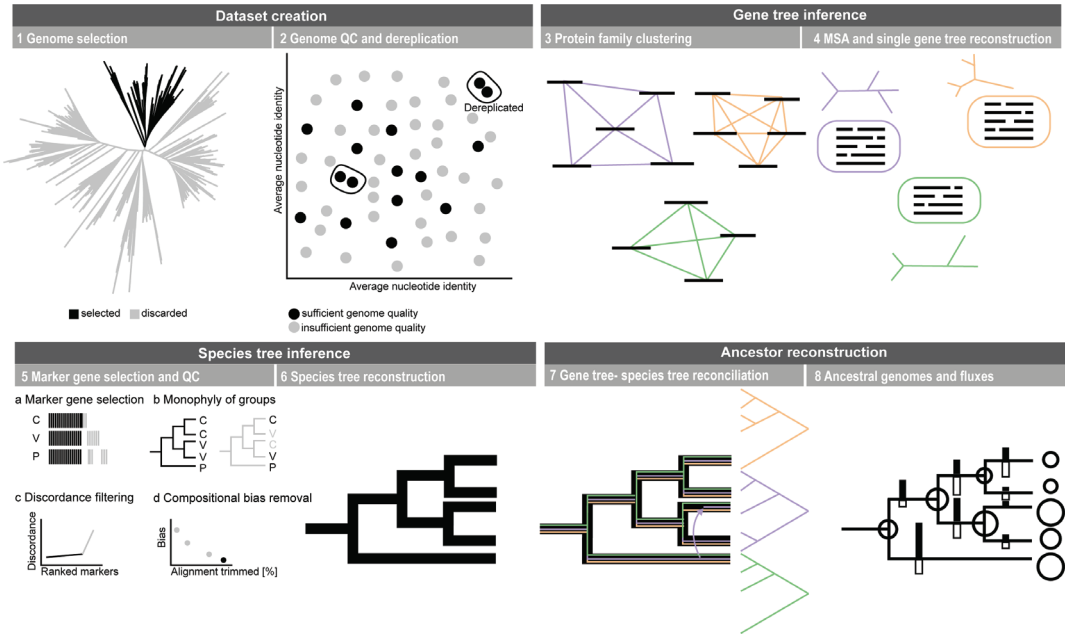
Environments. *Front. Microbiol.* **10**, 499 (2019).
65. Bertelli, C. *et al.* The Waddlia Genome: A Window into Chlamydial Biology. *PLoS One* **5**, e10890 (2010).
66. Mulkidjanian, A. Y., Dibrov, P. & Galperin, M. Y. The past and present of sodium energetics: May the sodium-motive force be with you. *Biochim. Biophys. Acta* **1777**, 985–992 (2008).
67. Pál, C., Papp, B. & Lercher, M. J. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat. Genet.* **37**, 1372–1375 (2005).
68. Castelle, C. J. *et al.* Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment. *Nat. Commun.* **4**, 2120 (2013).
69. Dibrova, D. V., Cherepanov, D. A., Galperin, M. Y., Skulachev, V. P. & Mulkidjanian, A. Y. Evolution of cytochrome bc complexes: from membrane-anchored dehydrogenases of ancient bacteria to triggers of apoptosis in vertebrates. *Biochim. Biophys. Acta* **1827**, 1407–1427 (2013).
70. Wernegreen, J. J. For better or worse: genomic consequences of intracellular mutualism and parasitism. *Curr. Opin. Genet. Dev.* **15**, 572–583 (2005).
71. Moran, N. A. Microbial Minimalism: Genome Reduction in Bacterial Pathogens. *Cell* **108**, 583–586 (2002).
72. Moliner, C., Fournier, P.-E. & Raoult, D. Genome analysis of microorganisms living in amoebae reveals a melting pot of evolution. *FEMS Microbiol. Rev.* **34**, 281–294 (2010).
73. Wang, Z. & Wu, M. Comparative Genomic Analysis of Acanthamoeba Endosymbionts Highlights the Role of Amoebae as a 'Melting Pot' Shaping the Rickettsiales Evolution. *Genome Biol. Evol.* **9**, 3214–3224 (2017).
74. Bertelli, C. & Greub, G. Lateral gene exchanges shape the genomes of amoeba-resisting microorganisms. *Front. Cell. Infect. Microbiol.* **2**, 110 (2012).
75. Medina, M. & Sachs, J. L. Symbiont genomics, our new tangled bank. *Genomics* **95**, 129–137 (2010).
76. Tsai, Y.-M., Chang, A. & Kuo, C.-H. Horizontal Gene Acquisitions Contributed to Genome Expansion in Insect-Symbiotic Spiroplasma clarkii. *Genome Biol. Evol.* **10**, 1526–1532 (2018).
77. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
78. Coleman, G. A. *et al.* A rooted phylogeny resolves early bacterial evolution. *Science* **372**, (2021).
79. Popa, O., Hazkani-Covo, E., Landan, G., Martin, W. & Dagan, T. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res.* **21**, 599–609 (2011).
80. Smillie, C. S. *et al.* Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**, 241–244 (2011).
81. Wiedenbeck, J. & Cohan, F. M. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiol. Rev.* **35**, 957–976 (2011).
82. Soucy, S. M., Huang, J. & Gogarten, J. P. Horizontal gene transfer: building the web of life. *Nat. Rev. Genet.* **16**, 472–482 (2015).
83. Hugoson, E., Lam, W. T. & Guy, L. miComplete: weighted quality evaluation of assembled microbial genomes. *Bioinformatics* **36**, 936–937 (2020).

84. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
85. Varghese, N. J. *et al.* Microbial species delineation using whole genome sequences. *Nucleic Acids Res.* **43**, 6761–6771 (2015).
86. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* **2**, 1533–1542 (2017).
87. Hug, L. A. *et al.* A new view of the tree of life. *Nat Microbiol* **1**, 16048 (2016).
88. Zhu, Q. *et al.* Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.* **10**, 5477 (2019).
89. Huerta-Cepas, J. *et al.* eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44**, D286–93 (2016).
90. Huerta-Cepas, J. *et al.* Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
91. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
92. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).
93. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
94. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
95. Quang, L. S., Gascuel, O. & Lartillot, N. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* **24**, 2317–2323 (2008).
96. Le, S. Q. & Gascuel, O. An Improved General Amino Acid Replacement Matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008).
97. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
98. Williams, K. P. *et al.* Phylogeny of Gammaproteobacteria. *J. Bacteriol.* **192**, 2305–2314 (2010).
99. Viklund, J., Ettema, T. J. G. & Andersson, S. G. E. Independent Genome Reduction and Phylogenetic Reclassification of the Oceanic SAR11 Clade. *Mol. Biol. Evol.* **29**, 599–615 (2012).
100. Lemoine, F. *et al.* Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature* **556**, 452–456 (2018).
101. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
102. Lartillot, N., Rodrigue, N., Stubbs, D. & Richer, J. PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment. *Syst. Biol.* **62**, 611–615 (2013).
103. Lartillot, N. & Philippe, H. A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. *Mol. Biol. Evol.* **21**, 1095–1109 (2004).

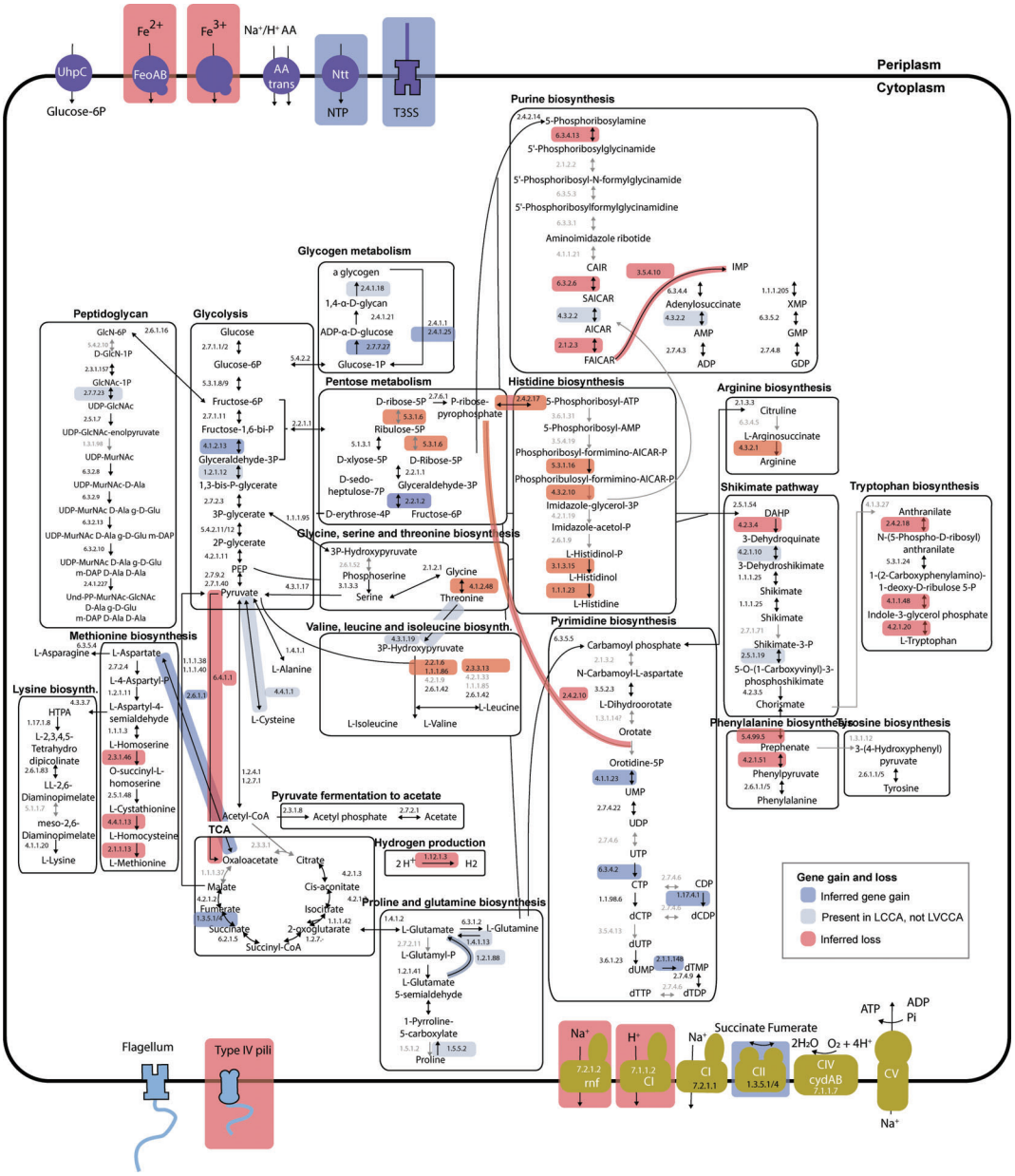
104. Lartillot, N., Brinkmann, H. & Philippe, H. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol. Biol.* **7 Suppl 1**, S4 (2007).
105. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–6 (2013).
106. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
107. Pruesse, E., Peplies, J. & Glöckner, F. O. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**, 1823–1829 (2012).
108. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
109. Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286–2288 (2009).
110. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
111. Miele, V., Penel, S. & Duret, L. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics* **12**, 116 (2011).
112. Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371–373 (2003).
113. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
114. Nakamura, T., Yamada, K. D., Tomii, K. & Katoh, K. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* **34**, 2490–2492 (2018).
115. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
116. Blum, M. *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **49**, D344–D354 (2021).
117. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J. Mol. Biol.* **428**, 726–731 (2016).
118. Csurös, M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* **26**, 1910–1912 (2010).
119. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
120. Martijn, J. *et al.* Hikarchaeia demonstrate an intermediate stage in the methanogen-to-halophile transition. *Nat. Commun.* **11**, 5490 (2020).
121. Huang, W.-C. *et al.* Comparative genomic analysis reveals metabolic flexibility of Woesearchaeota. *Nat. Commun.* **12**, 5281 (2021).
122. R Core Team. R: A Language and Environment for Statistical Computing. (2018).
123. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).
124. Köster, J. & Rahmann, S. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics* **34**, 3600 (2018).
125. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **44**, D7–19 (2016).

126. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
127. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **49**, D10–D17 (2021).
128. Schoch, C. L. *et al.* NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database* **2020**, (2020).
129. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS One* **5**, e9490 (2010).
130. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–62 (2016).
131. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic Acids Res.* **48**, D445–D453 (2020).
132. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
133. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer Science & Business Media, 2009).
134. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
135. Wang, L.-G. *et al.* Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. *Mol. Biol. Evol.* **37**, 599–603 (2020).

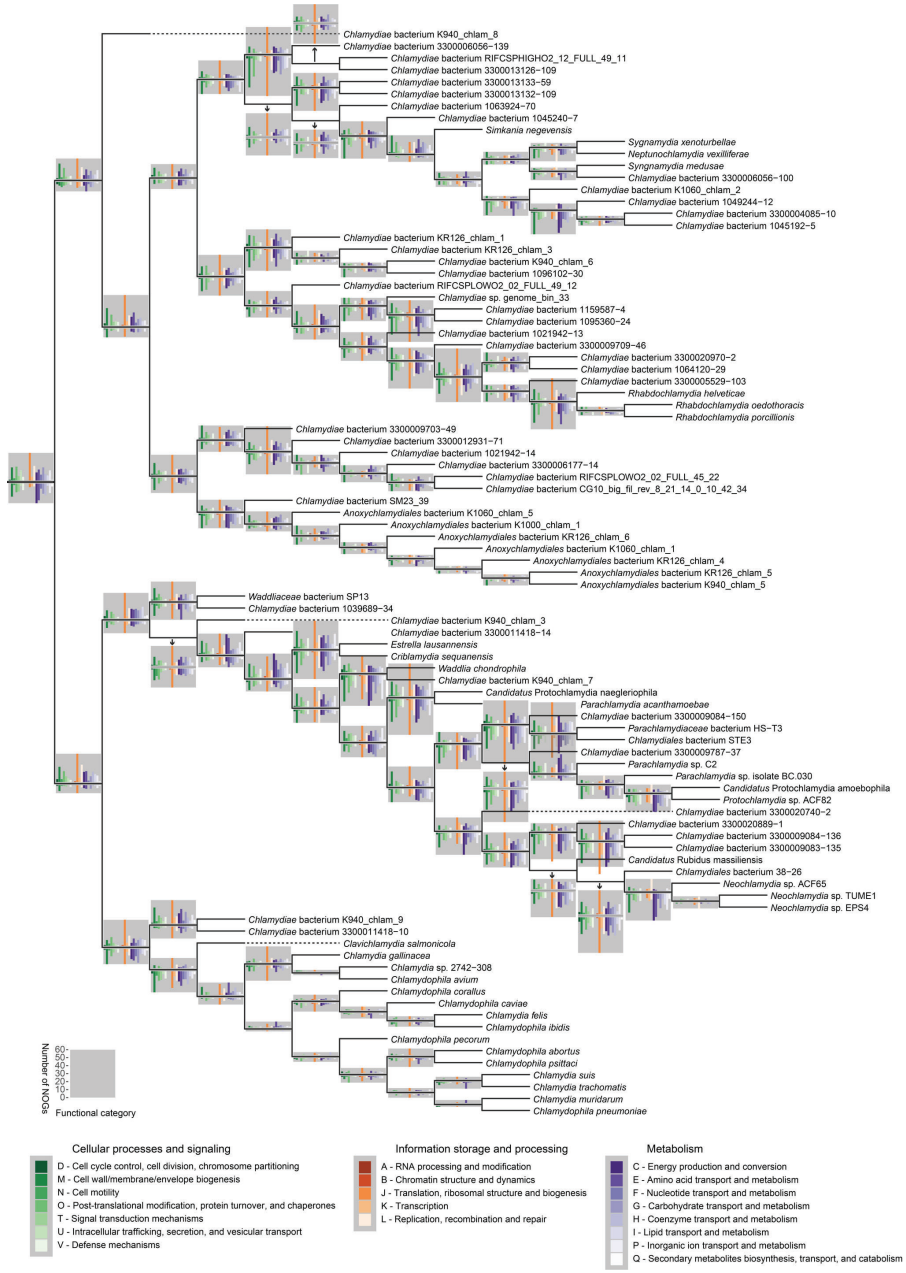
EXTENDED DATA FIGURES



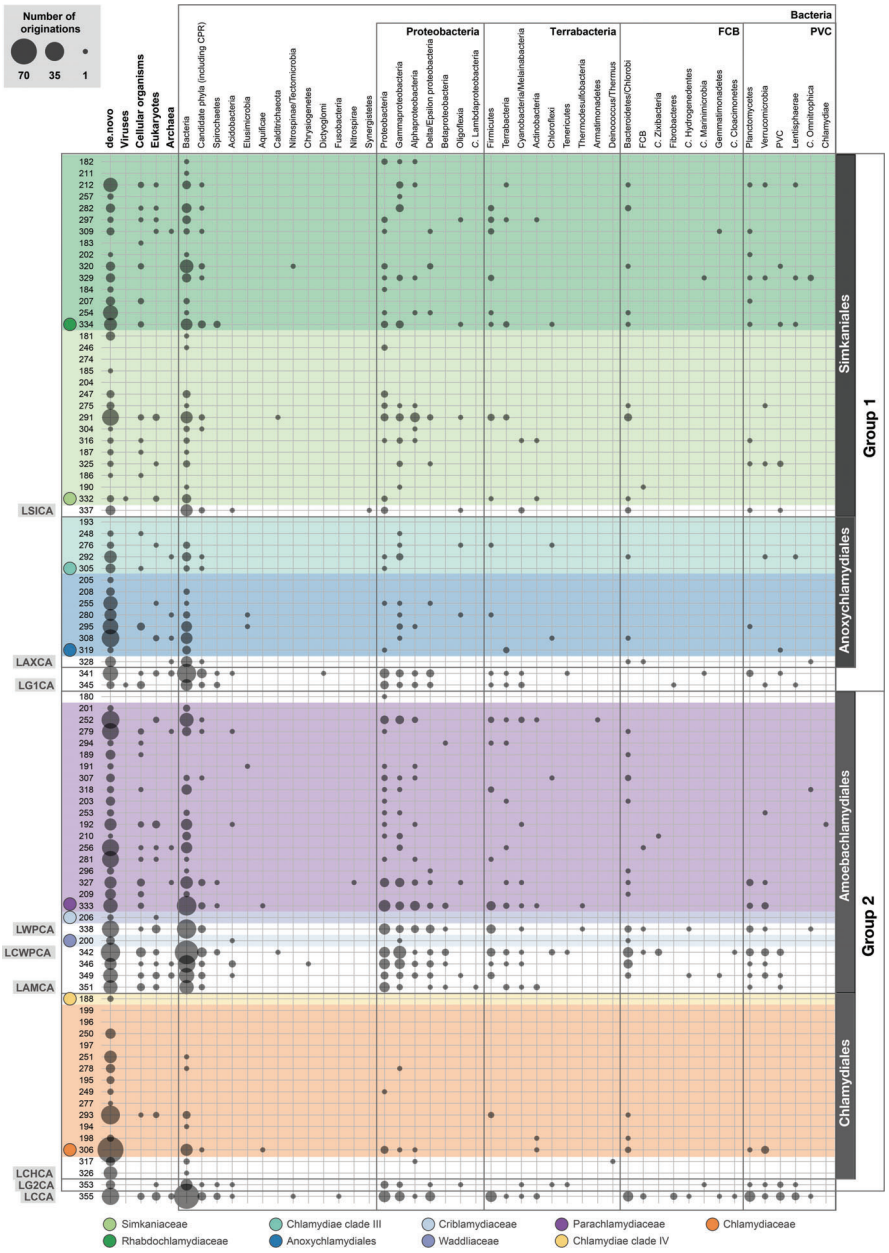
Extended Data Figure 1. Overview of workflow for ancestral state reconstruction of Chlamydiae. Summarized in short here (see Methods for details). Dataset creation: PVC representatives from public repositories were selected with completeness ≥ 90% and redundancy ≤ 2%. Species and genus-level representatives were selected for Chlamydiae and non-Chlamydiae PVC members, respectively. Gene tree inference: Proteins from the selected dataset were clustered into protein families at the last universal common ancestor level using eggNOG emapper[90]. Unmapped proteins were *de-novo* clustered using SiLiX[111]. Proteins from each resulting protein cluster were aligned into a multi-sequence alignment (MSA) and maximum-likelihood (ML) single-protein trees inferred. Species tree inference: Protein clusters found in a single-copy in at least 95% of dataset taxa were selected as potential marker genes. ML single-protein trees were inferred and manually curated, with proteins that well-resolved PVC phyla retained; further markers were removed through discordance filtering, while distant homologs, paralogs, and redundant proteins were removed for each kept marker. Individually aligned markers were then concatenated into a supermatrix alignment that was used for both ML and Bayesian phylogenetic inference, with compositionally heterogeneous sites sequentially removed to reduce bias. Ancestor reconstruction: ALE[29–31] was used to reconstruct ancestral states through gene-tree species-tree reconciliation. A reconciliation frequency cutoff of 0.3 was chosen for inferring the likely gene content of ancestors and for examining evolutionary events in Chlamydiae genome evolution.



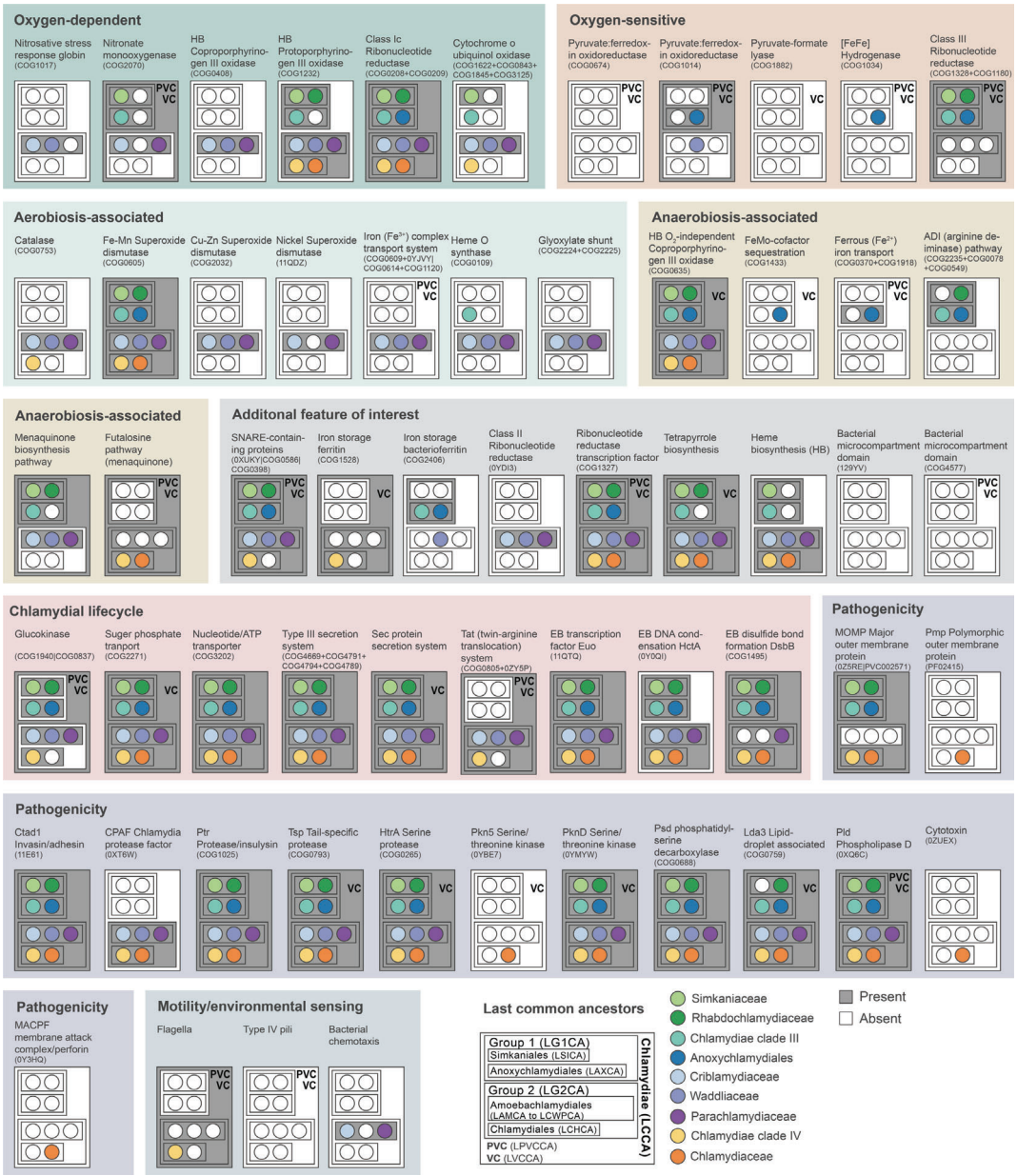
Extended Data Figure 2. Summary of Chlamydiae last common ancestor (LCCA) central metabolism. A gene was inferred to be present by ALE. Enzyme complexes were inferred as present if at least half of the necessary genes were present. Enzymes are annotated with enzyme commission numbers, arrows indicate the directionality of reactions and are colored in black or grey if inferred as present or absent, respectively. Boxes around reactions indicate metabolic pathways. Red and purple boxes indicate gain and loss of enzymes in LCCA relative to the last common ancestor of Chlamydiae, Verrucomicrobia, Lentisphaerae, and Kiritimatiellaeota (LVCCA).



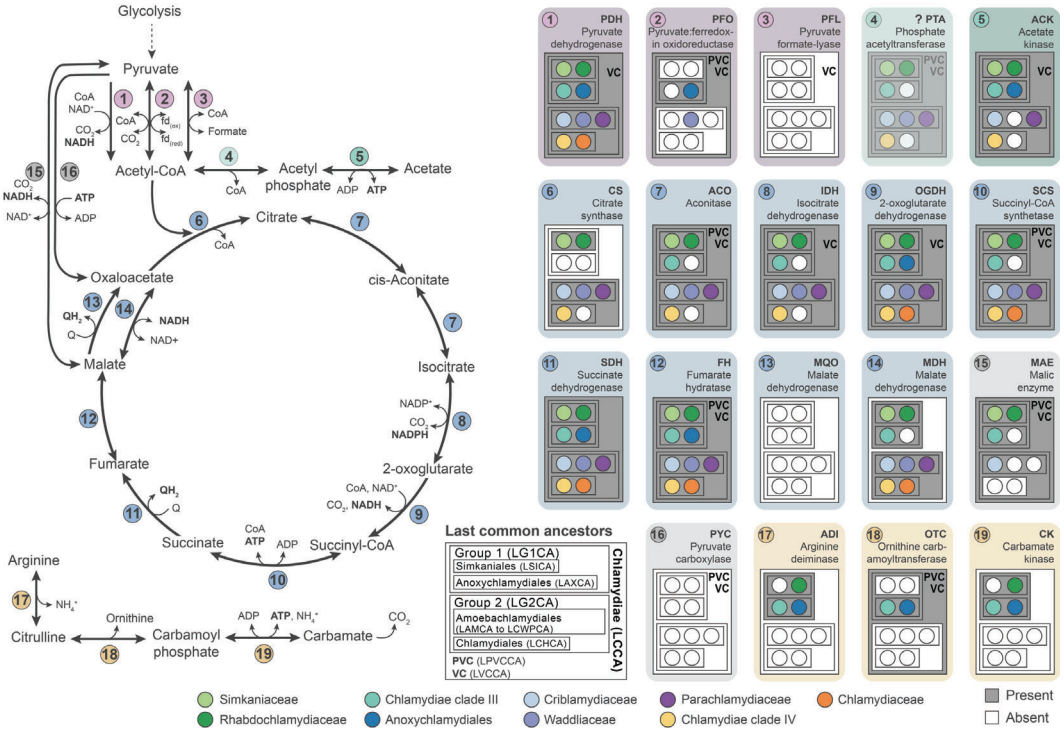
Extended Data Figure 3. Gain and loss events across the Chlamydiae tree. Cladogram of Chlamydiae phylogenetic relationships showing gain (originations and transfers) and loss events of eggNOG NOG protein family clusters across COG categories. The barplots mapped onto each branch indicate the events that have occurred leading to the ancestral node to their right. The number of NOGs gained are indicated in the positive direction, and those lost in the negative direction, with the number corresponding to the bar height (see grey box scale). Bars are sorted and coloured according to the COG category. NOGs assigned to poorly characterized COG categories (R: general function prediction only, and S: function unknown) and *de-novo* clusters were excluded. See repository files for raw data.



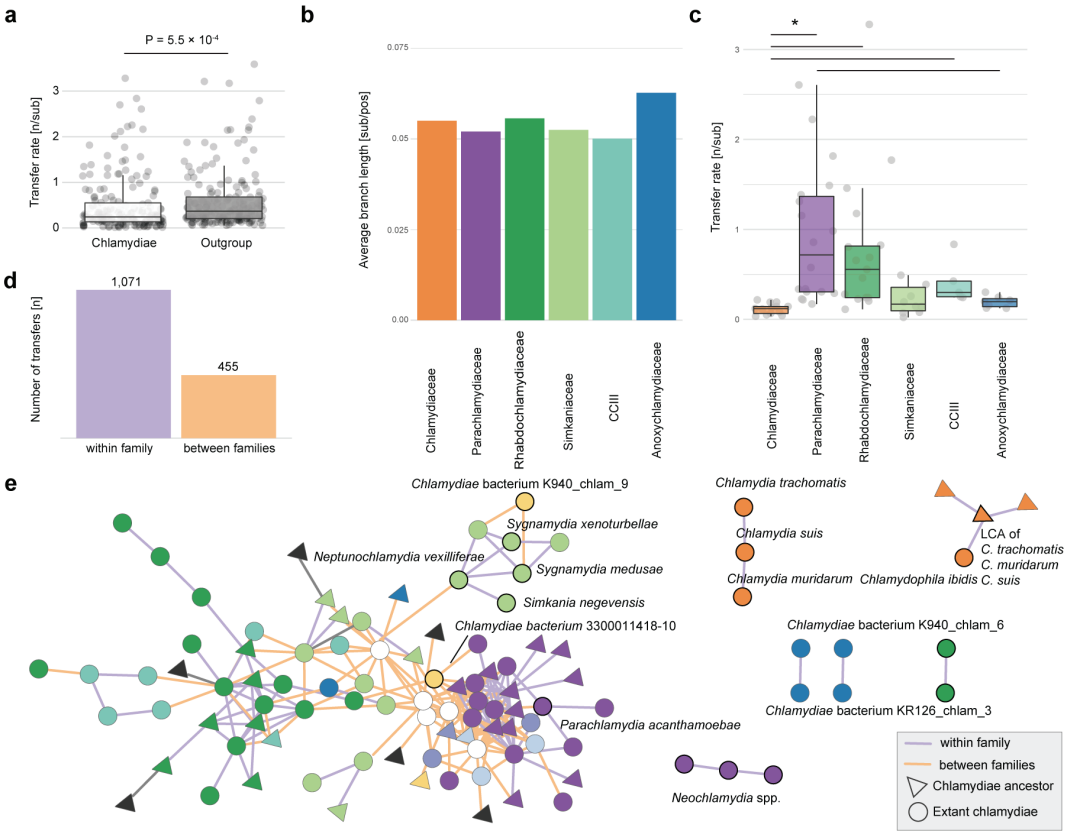
Extended Data Figure 4. Origins of gene gains during Chlamydiae evolution. Bubbleplot showing the putative donor taxonomy and number of gene originations across Chlamydiae ancestral nodes. Chlamydiae taxonomy is indicated to the right, while nodes corresponding to key ancestors are indicated to the left; see Data S7 for ancestor abbreviations and Figure S10 for node numbers mapped to chlamydial phylogeny. Single-protein trees were inferred for each protein origination alongside sequences from reference databases. The taxonomy of the putative donor lineage was determined as the highest-level taxonomy (at domain, superphylum, and phylum) that could be assigned to 75% of taxa in a well-supported monophyletic clade sister to chlamydial sequences (see Methods). See Data S10 for originations and putative donor lineages.



Extended Data Figure 5. Overview of the presence and absence of genes and pathways of interest across key Chlamydiae ancestors. The cluster ID (COG, NOG, or *de-novo* identifier) is indicated for pathways, proteins, and complexes, and where not shown can be found in Data S10, alongside cluster annotations and inferred presence across all Chlamydiae ancestors. Genes and pathways are split into functional groups of interest (see legend and ancestor abbreviations in Data S7).



Extended Data Figure 6. Overview of the evolution of energy conservation pathways involving the TCA cycle and fermentative metabolism in Chlamydiae. Presence and absence of proteins and complexes involved in converting pyruvate to acetyl-CoA (pink), pyruvate fermentation to acetate (green), the tricarboxylic acid (TCA) cycle (blue), pyruvate to TCA intermediate interconversion (grey), and the fermentative arginine deiminase pathway (yellow). A schematic overview of reactions performed by each numbered enzyme is shown to the left, alongside key metabolites (water, bicarbonate, and protons are excluded). To the right the presence and absence of each component across key ancestors is indicated according to the legend. See also Data S7 for ancestor abbreviations and Data S10 for protein annotations and presence across all Chlamydiae ancestors.



Extended Data Figure 7. Transfer rates in Chlamydiae and other PVC bacteria, and network of enriched transfer routes in Chlamydiae. a, Box plot depicting significantly lower transfer rates (Wilcoxon signed rank test, $P = 5.5 \times 10^{-4}$) in Chlamydiae (median 0.41, IQR 0.22-0.76) than other PVC bacteria (outgroup, median 0.49, IQR 0.27-1.02). **b**, Average branch length per group without terminal leaves. **c**, Transfers per substitution in the species tree per chlamydial node and family. To account for heterogeneous genome sampling in our dataset, we only evaluated transfer rates of chlamydial families with at least two ancestors reconstructed in our analysis, excluding terminal nodes. Transfer rates are significantly higher in in ancestors of the Parachlamydiaceae, Rhabdochlamydiaceae and CC-III than the Chlamydiaceae ($P = (5.1 \times 10^{-7}, 3.9 \times 10^{-5}, 1.8 \times 10^{-3})$, Wilcoxon rank sum test with Bonferroni correction) and in Parachlamydiaceae than in Anoxychlamydiaceae ($P = 2.9 \times 10^{-3}$, Wilcoxon rank sum test with Bonferroni correction), respectively. **d,** Bar plot showing the number of significant transfer events (binomial test, see Methods) in gene highways within (purple) and between (orange) chlamydial families and **e.** corresponding networks depicting chlamydiae (nodes, color coded by family; families that are only represented by one genome are white, see Figure 1) sharing significant gene highways (edges) and potentially niches. Edges are colored based on gene highways within and between families in purple and orange, respectively. See also Extended Data Figure 3 for events across all PVC nodes, and Data S7 for raw data.

SUPPLEMENTARY INFORMATION

Gene gain facilitated the origin and diversification of an ancient phylum of symbionts

Jennah E. Dharamshi†[1], Stephan Köstlbacher†[2,3], Max E. Schön[1], Astrid Collingro[2], Thijs J. G. Ettema‡*[1,3], Matthias Horn‡*[2]

[1] Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, SE-75123 Uppsala, Sweden
[2] Centre for Microbiology and Environmental Systems Science, University of Vienna, Vienna, Austria
[3] Laboratory of Microbiology, Wageningen University, 6708 WE Wageningen, The Netherlands

† Equal contribution
‡ Equal contribution

* Correspondence to: thijs.ettema@wur.nl and matthias.horn@univie.ac.at

1. SUPPLEMENTARY DISCUSSIONS

Supplementary Discussion 1 - Inconsistencies in Chlamydiae phylogenetic relationships

The Parilichlamydiaceae are fish pathogens with reduced genomes that resemble the animal pathogens Chlamydiaceae[1], and yet they have been placed sister to all other Chlamydiae in species reconstructions[1–4]. However, in our analyses of concatenated marker proteins the phylogenetic position of Parilichlamydiaceae was unstable, indicating possible long-branch attraction (LBA) artefacts (Figures S3-S4). Counter to prior suggestions of convergent evolution, Parilichlamydiaceae and Chlamydiaceae formed a monophyletic group in a Bayesian phylogeny of PVC 16S rRNA genes (Figure S6). Despite their interesting biology and likely importance in Chlamydiae evolution, we were unable to confidently resolve the position of Parilichlamydiaceae and thus chose to remove them. Likewise, the phylogenetic placement of the recently described orphan lineage[4] *Chlamydiae* bacterium 1070360-7 was inconsistent and removed from further analyses (Figure S3).

Several long-branching chlamydial lineages formed a clade sister to other chlamydiae in initial species trees, but were found to be well-supported together with Simkaniaceae in Bayesian inferences with the CAT model of evolution (Figures 1 and S3-S5), which is known to alleviate LBA artefacts caused by fast-evolving sequences[5,6], and are thus referred to as Simkaniaceae-like. In ML phylogenies the position of these Simkaniaceae-like lineages was also reconstructed as forming a monophyletic group with Simkaniaceae, but only with the removal of compositionally heterogeneous sites. The removal of such sites from sequence alignments reduces systematic error in phylogenomic analyses by alleviating the artifact of species grouping together based on shared biases in amino acid composition[7]. Based on these results, we thus classified this Simkaniaceae-like group as putatively belonging to the Simkaniaceae family and included it as such for our ancestral reconstruction (Figures 1 and S3-S5).

A similar pattern was seen for *Chlamydiae* bacterium 3300009703-49 which initially affiliated with Anoxychlamydiaceae (formerly Anoxychlamydiales[8]), but was well-supported together with Chlamydiae Clade III (CC-III) once accounting for compositional heterogeneity (Figures 1 and S3-S5). CC-III and Anoxychlamydiaceae form a well-supported clade we refer to here as the order Anoxychlamydiales. Likewise, the position of *Chlamydiae* bacterium K940_chlam_8, another long-branching lineage, is supported with methods accounting for LBA and compositional bias artefacts (Figures 1 and S3-S5).

Supplementary Discussion 2 - Taxonomic descriptions and adjustments

Previous suggestions to divide the phylum *Chlamydiae* into several orders have not been accepted by the Subcommittee on the Taxonomy of Chlamydiae of the International Committee on Systematics of Prokaryotes (ICSP) due to the lack of data[9,10], but high quality data for many diverse chlamydiae have become available within the past years[2–4]. By including these additional data and a large number of non-chlamydial PVC genome data (n=89), a stable chlamydial phylogeny could be inferred with Bayesian and maximum-likelihood methods (see Methods, Figures S1 and S3-S6, Extended Data Figure 1, Data S1). This however was only possible after removing two family-level lineages with unstable branching patterns (*Candidatus* Parilichlamydiaceae and *Candidatus* MCF-D; see Supplementary Discussion 1) from the final dataset. Based on the single copy marker gene species tree inferred in this study and lately published chlamydial phylogenies for marker genes and the 16S rRNA gene, we propose the reclassification of the order levels within the phylum Chlamydiae[3,4,11].

Reclassification of the order *Chlamydiales* ord. corrig. (Kuo, Horn and Stephens, 2011)
The order *Chlamydiales* is so far the only officially accepted one within the *Chlamydiae* and includes all previously described family-level lineages[10,12]. Based on the stable monophyletic branching in concatenated marker gene and 16S rRNA gene tree inferences, we propose to reclassify the *Chlamydiales* and to just include members of the *Chlamydiaceae*[13], *Candidatus* Clavichlamydiaceae[14], and *Candidatus* Chlamydiae Clade IV (CC-IV; *i.e.*, Sororchlamydiaceae)[3] in this order.

Description of *Amoebachlamydiales* ord. nov.
(A.moe.ba.chla.my.di.a'les. N.L. fem. n. *Amoeba* derived from Gr. amoibe; N.L. fem. *Chlamydia* taxonomic name of a bacterial genus; L. suff. -*ales* ending to denote an order; *Amoebachlamydiales* N.L. fem. pl. n. referring to Amoebozoa as the major known hosts of members of the order)

The order *Amoebachlamydiales* represents a distinct monophyletic lineage as supported by concatenated marker protein and 16S rRNA gene trees. The order includes members of the families *Criblamydiaceae*[15], *Waddliaceae*[16] and *Parachlamydiaceae*[17]. All of these families have cultivated representatives that thrive in Amoebozoa hosts. Additional families whose representatives have so far only been recovered from genomic data, including *Candidatus* MCF-F, *Candidatus* MCF-G, *Candidatus* K940_chlam_3, and *Candidatus* GCA-270938 should be included in this order[3,4]. Members of the *Amoebachlamydiales* often have extended genetic repertoires for aerobic respiration and other metabolic pathways compared to most other chlamydiae. Their genome sizes range between 2-4 Mb.

Description of *Simkaniales* ord. nov.
(Sim.ka.ni.a'les. N.L. fem. n. *Simkania* type genus of the order; L. suff. -*ales* ending to denote an order; N.L. fem. pl. n. *Simkaniales* referring to the order that includes the type genus *Simkania*)

The order *Simkaniales* represents a distinct monophyletic lineage as supported by concatenated marker protein and 16S rRNA gene trees. It includes members of the family-level lineages *Simkaniaceae* and *Rhabdochlamydiaceae*[17,18].

Description of *Candidatus* Anoxychlamydiales ord. nov.
(An.oxy.chla.my.di.a'les. Gr. pref. *An*- not; N.L. neut. n. *oxygenium* chemical element oxygen; N.L. fem. *Chlamydia* taxonomic name of a bacterial genus; L. suff. -*ales* ending to denote an order; *Anoxychlamydiales* referring to the potential anoxic lifestyle of some members of this order)

The order *Anoxychlamydiales* represents a distinct monophyletic lineage as supported by concatenated marker protein and 16S rRNA gene trees. It includes members of the family-level lineages *Candidatus* Anoxychlamydiaceae and *Candidatus* Chlamydiae Clade III (CC-III)[3].

Description of *Candidatus* Anoxychlamydiaceae fam. nov.
(An.oxy.chla.my.di.a.ce'ae. Gr. pref. *An*- not; N.L. neut. n. *oxygenium* chemical element oxygen; N.L. fem. *Chlamydia* taxonomic name of a bacterial genus; L. suff. -*aceae* ending to denote a family; *Anoxychlamydiaceae* referring to the potential anoxic lifestyle of members of this family)

The family *Candidatus* Anoxychlamydiaceae represents a distinct monophyletic lineage as supported by concatenated marker protein and 16S rRNA gene trees. Members of this family so

far are only represented by metagenome-assembled genomes[3,4]. Members of this family encode the arginine deiminase pathway, [FeFe]-hydrogenase, and a pyruvate:ferredoxin oxidoreductase indicating an obligate anoxic lifestyle for these organisms. In addition many oxygen-dependent genes are missing in the genomes of the *Candidatus* Anoxychlamydiaceae.

2. SUPPLEMENTARY FIGURES

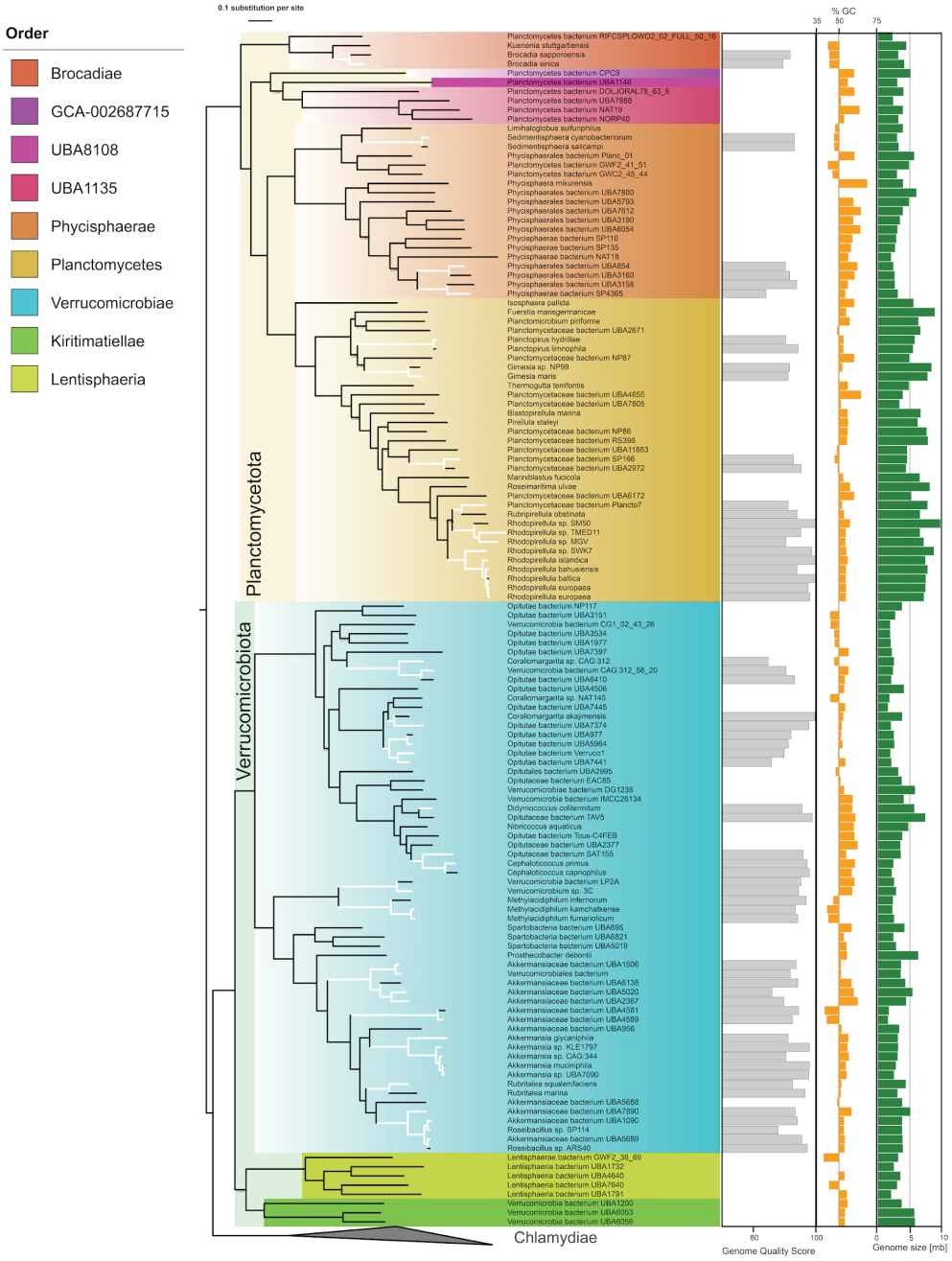


Figure S1. Genus-level dereplication of outgroup genomes based on genome quality score. The phylogenetic tree is based on 120 bacterial single copy marker proteins[19] inferred with FastTree v2.1[20]. Clades are colored by GTDB[19] assigned order rank. Black leafs in the species tree represent the selected genomes for downstream analysis, while white leafs were discarded because of higher quality genomes in the same genus. The inner bar chart depicts the genome quality score of genomes in genera with more than one member. The middle chart represents % GC deviation from 50, the outer chart depicts the genome size. See also Data S1-S2.

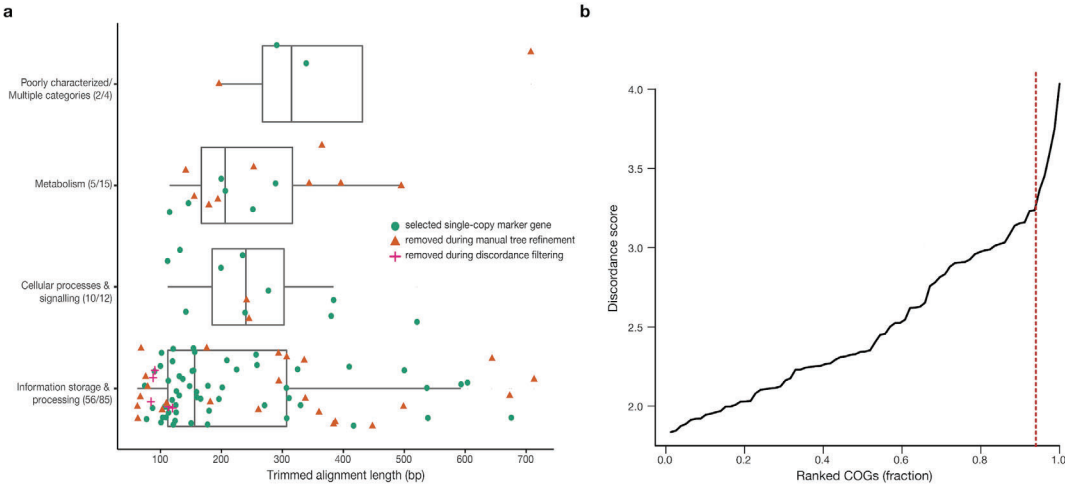


Figure S2. a. Boxplot of putative marker genes for inferring species trees and the length of the corresponding trimmed protein alignment. Proteins are split into larger COG categories with symbols indicating whether it was removed during manual tree refinement (orange triangles), removed during discordance filtering (pink pluses), or selected for use in concatenated species phylogenies (green circles). **b**. Marker protein COGs that passed manual tree refinement are ranked according to discordance score[21]. The red line indicates the fraction of proteins (n=5) that were removed based on having the largest discordance from other trees. See also Data S3.

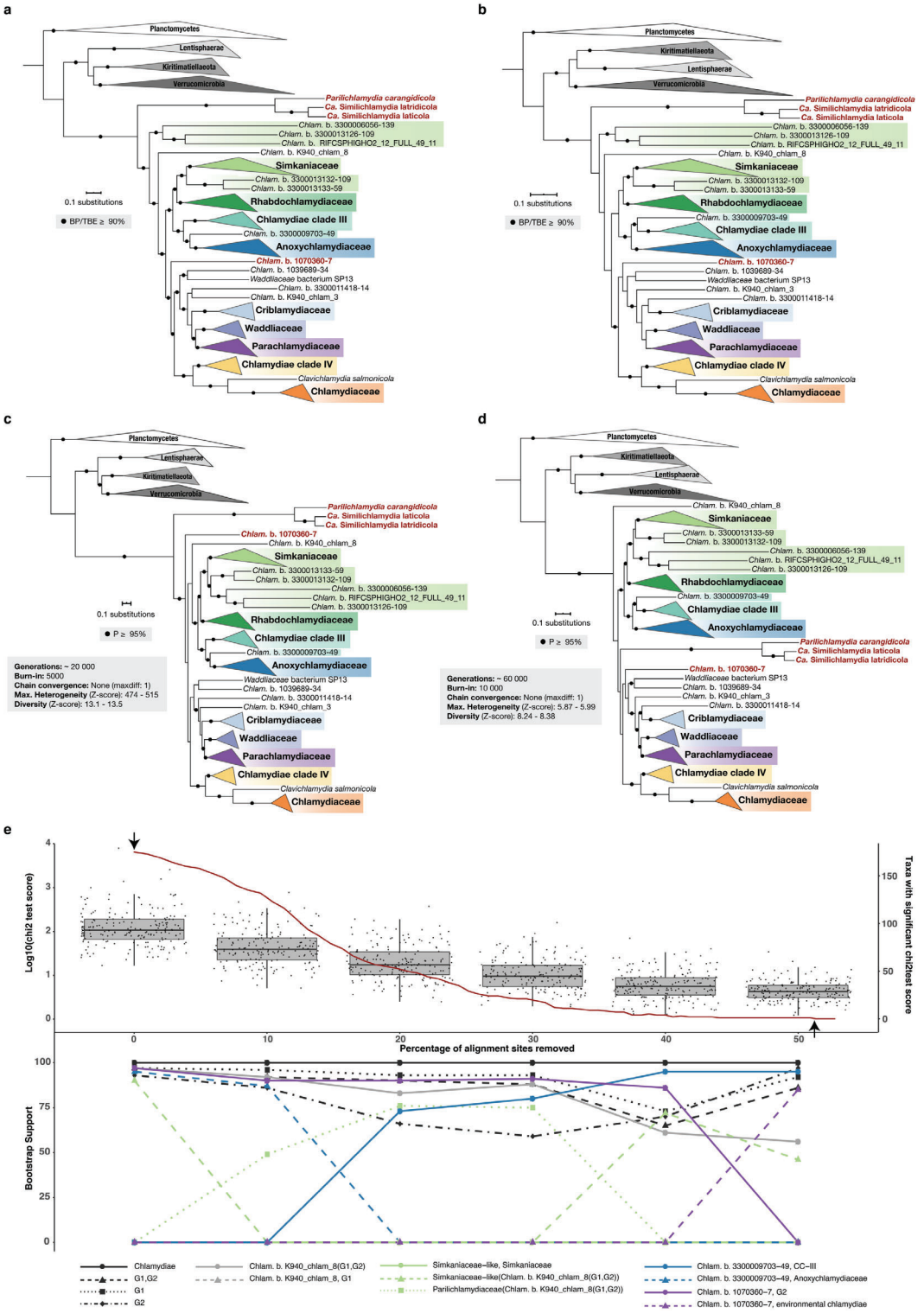


Figure S3. See legend appended.

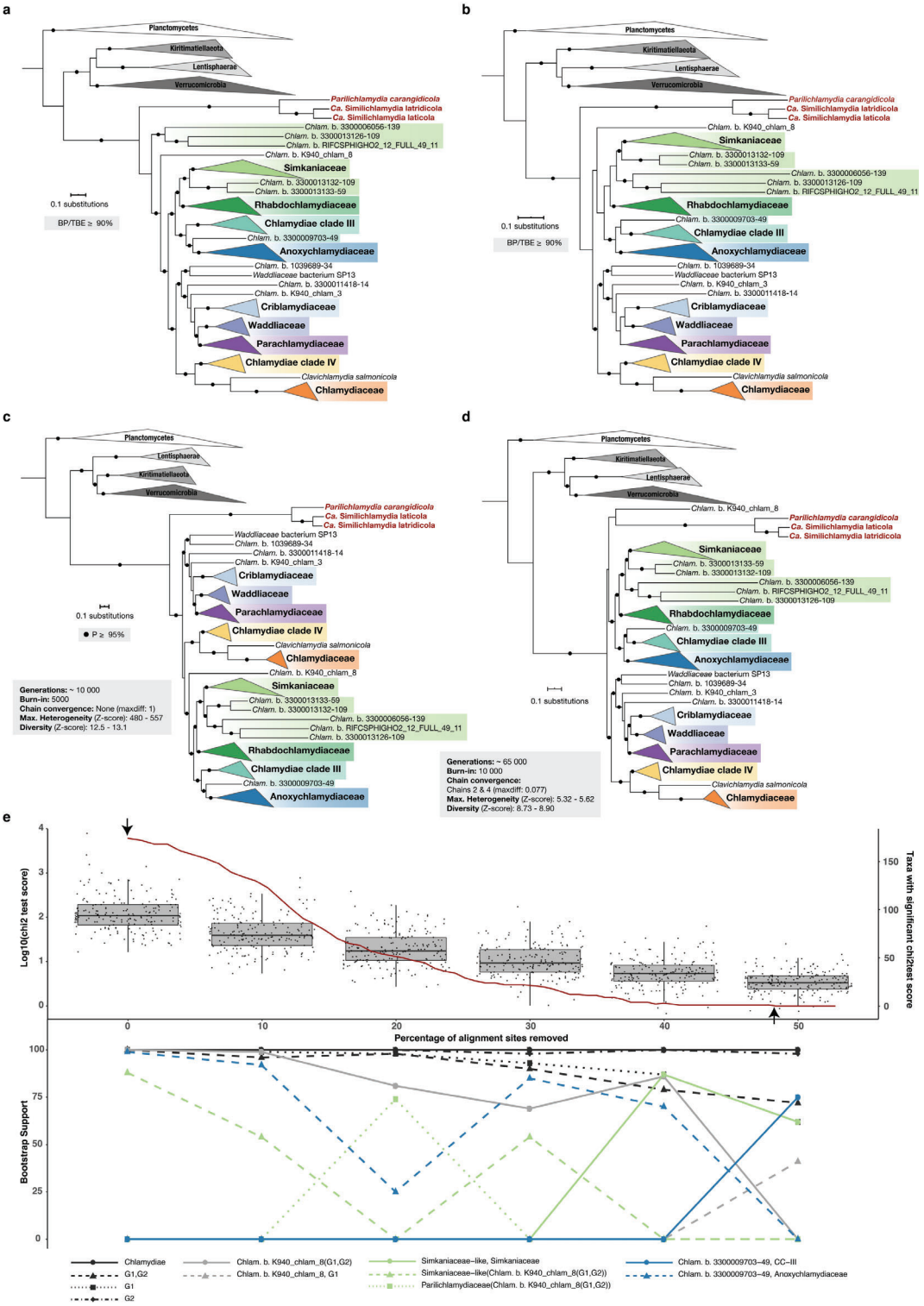


Figure S4. See legend appended.

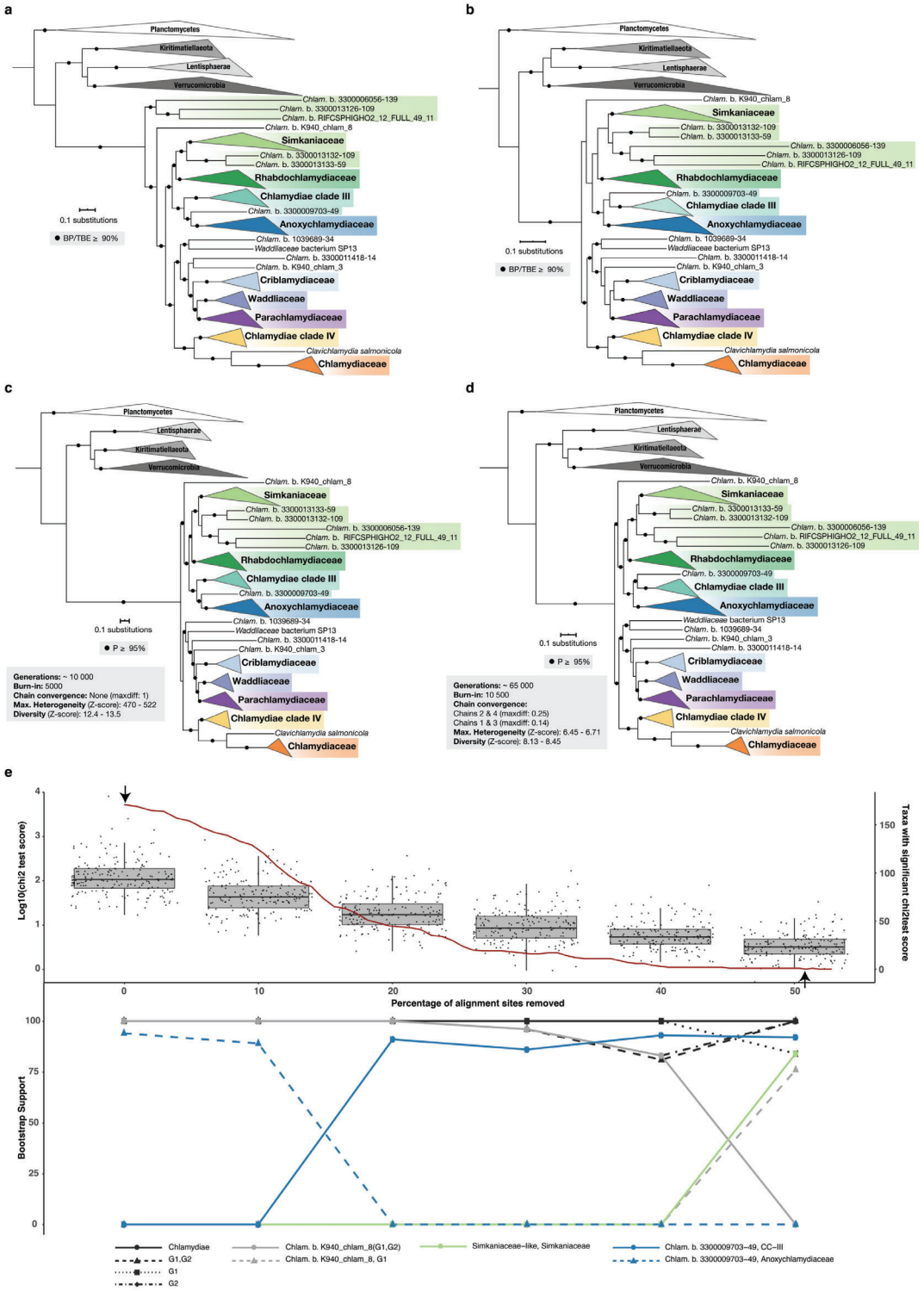


Figure S5. See legend appended.

Figure S3. Maximum-likelihood (**a,b**) and Bayesian consensus (**c,d**) species phylogenies of concatenated single-copy marker proteins from all PVC representatives (184 taxa) were inferred with the complete alignment (**a,c**) and with compositionally heterogeneous sites removed (**b,d**). High support for bipartitions is indicated by a black circle, with support measured by non-parametric bootstraps (BP) and the transfer bootstrap expectation (TBE) inferred using IQ-TREE with the LG+C60+F+G4-derived PMSF approximation, or posterior probability (P) inferred with Phylobayes using the CAT+GTR+G4 model of evolution. Consistent clades are collapsed with their taxonomy indicated and chlamydial families coloured. Taxa with unclear phylogenetic affiliation are coloured red. Run characteristics and converged chains are indicated for Bayesian phylogenies in a grey box (**c,d**). Scale bars indicate the number of substitutions per site. **e.** The 1% most compositionally heterogeneous sites were removed incrementally starting from the initial alignment. Boxplots show the log range of $\chi 2$ test scores (scale on the left) across taxa from alignments with the corresponding percentage of sites removed. The red line (scale to the right) indicates the number of taxa with significant $\chi 2$ test scores, and hence deviation from patterns of amino acid composition found in other taxa. Arrows indicate the initial alignment (**a,c**) and the percentage of alignment sites removed for no remaining taxa with significant deviations in composition (**b,d**). Changes in BP support for the monophyly of different groups of interest is shown in the bottom line plots, at different intervals of percentage sites removed. Abbreviations and short forms include: *Chlam* b. (*Chlamydiae* bacterium), G1 (Group 1), G2 (Group 2), and CC-III (Chlamydiae Clade III). See also Data S4-S6.

Figure S4. Maximum-likelihood (**a,b**) and Bayesian consensus (**c,d**) species phylogenies of concatenated single-copy marker proteins from PVC representatives with the removal of *Chlamydiae* bacterium 1070360-7 (183 taxa) were inferred with the complete alignment (**a,c**) and with compositionally heterogeneous sites removed (**b,d**). High support for bipartitions is indicated by a black circle, with support measured by non-parametric bootstraps (BP) and the transfer bootstrap expectation (TBE) inferred using IQ-TREE with the LG+C60+F+G4-derived PMSF approximation, or posterior probability (P) inferred with Phylobayes using the CAT+GTR+G4 model of evolution. Consistent clades are collapsed with their taxonomy indicated and chlamydial families coloured. Taxa with unclear phylogenetic affiliation are coloured red. Run characteristics and converged chains are indicated for Bayesian phylogenies in a grey box (**c,d**). Scale bars indicate the number of substitutions per site. **e.** The 1% most compositionally heterogeneous sites were removed incrementally starting from the initial alignment. Boxplots show the log range of $\chi 2$ test scores (scale on the left) across taxa from alignments with the corresponding percentage of sites removed. The red line (scale to the right) indicates the number of taxa with significant $\chi 2$ test scores, and hence deviation from patterns of amino acid composition found in other taxa. Arrows indicate the initial alignment (**a,c**) and the percentage of alignment sites removed for no remaining taxa with significant deviations in composition (**b,d**). Changes in BP support for the monophyly of different groups of interest is shown in the bottom line plots, at different intervals of percentage sites removed. Abbreviations and short forms include: *Chlam* b. (*Chlamydiae* bacterium), G1 (Group 1), G2 (Group 2), and CC-III (Chlamydiae Clade III). See also Data S4-S6.

Figure S5. Maximum-likelihood (**a,b**) and Bayesian consensus (**c,d**) species phylogenies of concatenated single-copy marker proteins from PVC representatives with the removal of *Chlamydiae* bacterium 1070360-7 and member of the Parilichlamydiaceae family (180 taxa) were inferred with the complete alignment (**a,c**) and with compositionally heterogeneous sites removed (**b,d**). High support for bipartitions is indicated by a black circle, with support measured by non-parametric bootstraps (BP) and the transfer bootstrap expectation (TBE) inferred using IQ-TREE with the LG+C60+F+G4-derived PMSF approximation, or posterior probability (P) inferred with Phylobayes using the CAT+GTR+G4 model of evolution. Consistent clades are collapsed with their taxonomy indicated and chlamydial families coloured. Run characteristics and converged chains are indicated for Bayesian phylogenies in a grey box (**c,d**). Scale bars indicate the number of substitutions per site. **e.** The 1% most compositionally heterogeneous sites were removed incrementally starting from the initial alignment. Boxplots show the log range of $\chi 2$ test scores (scale on the left) across taxa from alignments with the corresponding percentage of sites removed. The red line (scale to the right) indicates the number of taxa with significant $\chi 2$ test scores, and hence deviation from patterns of amino acid composition found in other taxa. Arrows indicate the initial alignment (**a,c**) and the percentage of alignment sites removed for no remaining taxa with significant deviations in composition (**b,d**). Changes in BP support for the monophyly of different groups of interest is shown in the bottom line plots, at different intervals of percentage sites removed. Abbreviations and short forms include: *Chlam* b. (*Chlamydiae* bacterium), G1 (Group 1), G2 (Group 2), and CC-III (Chlamydiae Clade III). See also Figure 1 and Data S4-S6.

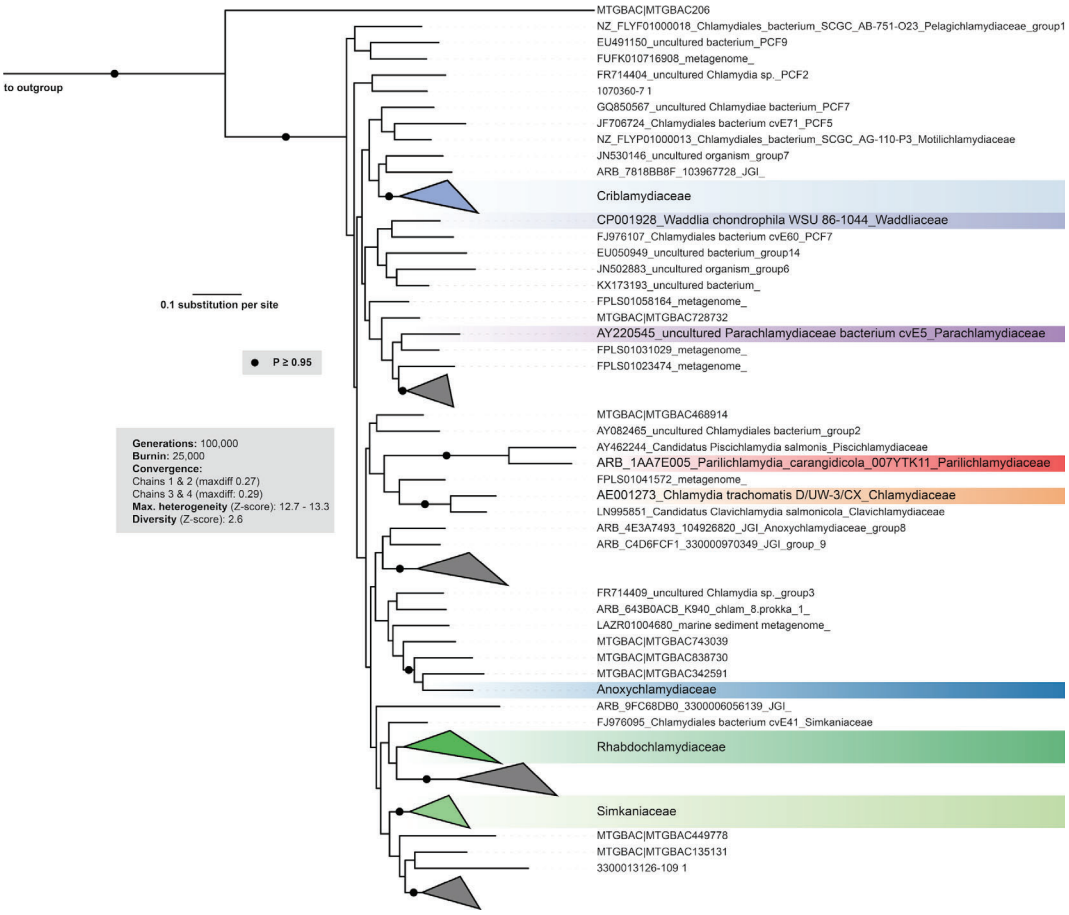


Figure S6. Bayesian 16S rRNA gene phylogenies of family level representatives. Species tree (CAT+GTR+G4 model) inferred from an alignment (1,533 aligned positions) of 177 approximately family level representative 16S rRNA gene sequences (> 1,200 nt). Well supported clades (P ≥ 0.95), indicated by filled circles at the nodes, with more than two members were collapsed. Family clades with more than one genome representative are highlighted.

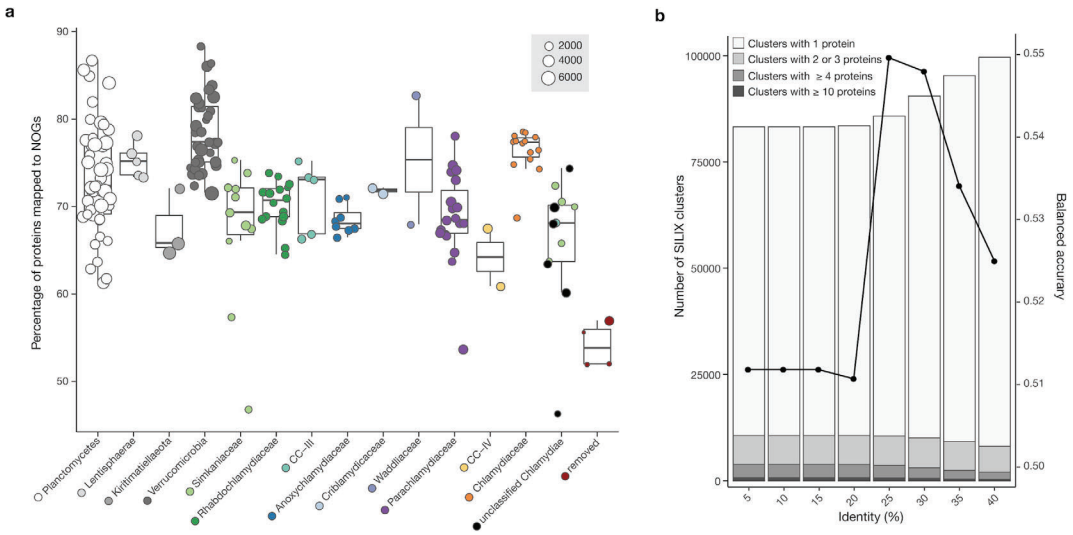


Figure S7. a. Boxplot showing the percentage of each genome mapped to eggNOG NOGs from different PVC phyla and chlamydial families (coloured accordingly). Circles represent individual genomes with size indicating the total number of proteins from that genome. Simkaniaceae-like lineages are coloured in green though they are included under unclassified Chlamydiae. Removed lineages include those excluded from further analyses based on their inconsistent positions in species trees (*i.e.*, members of the Parilichlamydiaceae family, and *Chlamydiae* bacterium 1070360-7). **b.** Proteins not mapped to NOGs were *de-novo* clustered. Barplots show the number of clusters (left axis) generated based on different % identity cutoffs, with the number of cluster members indicated by the stacked bars (see legend). Balanced accuracy at different cutoffs is shown by the line plot (right scale) (see Methods). Percentage identity of 25% maximized the balanced accuracy and was thus selected.

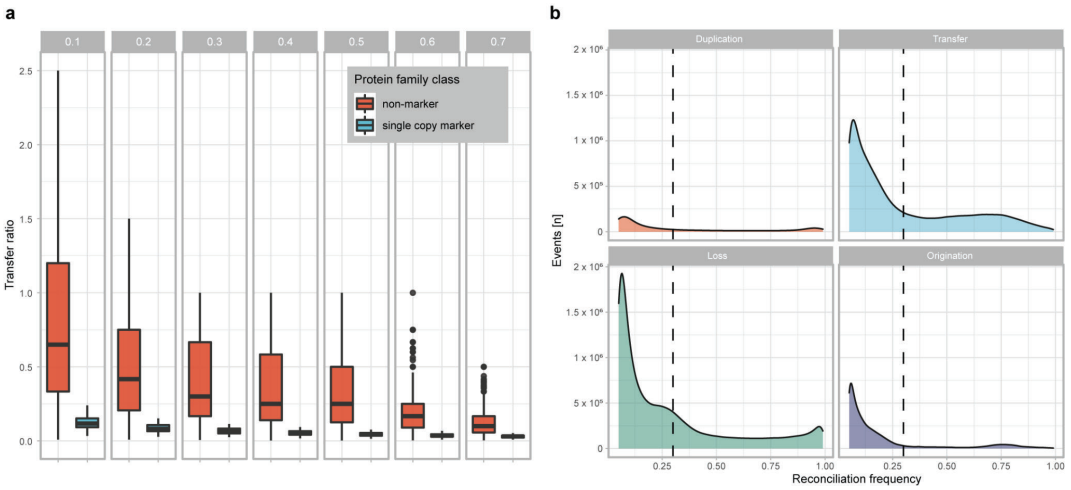


Figure S8. a. Boxplots indicating the ratio of transfers inferred for all non-marker proteins with two or more members (red; n=13,555), and single-copy marker proteins (blue; n=74; used for species phylogeny inference), and different reconciliation frequency cutoffs for transfer events. At a cutoff of 0.3 the median transfer ratio is likewise 0.3 for non-marker proteins, equalling 70% vertical transmission events. **b**. Density distributions of the number of events inferred to have occurred across reconciliation frequencies for each event type. The selected cutoff of 0.3 is indicated by a dashed line with events to the right of the line used for further analyses.

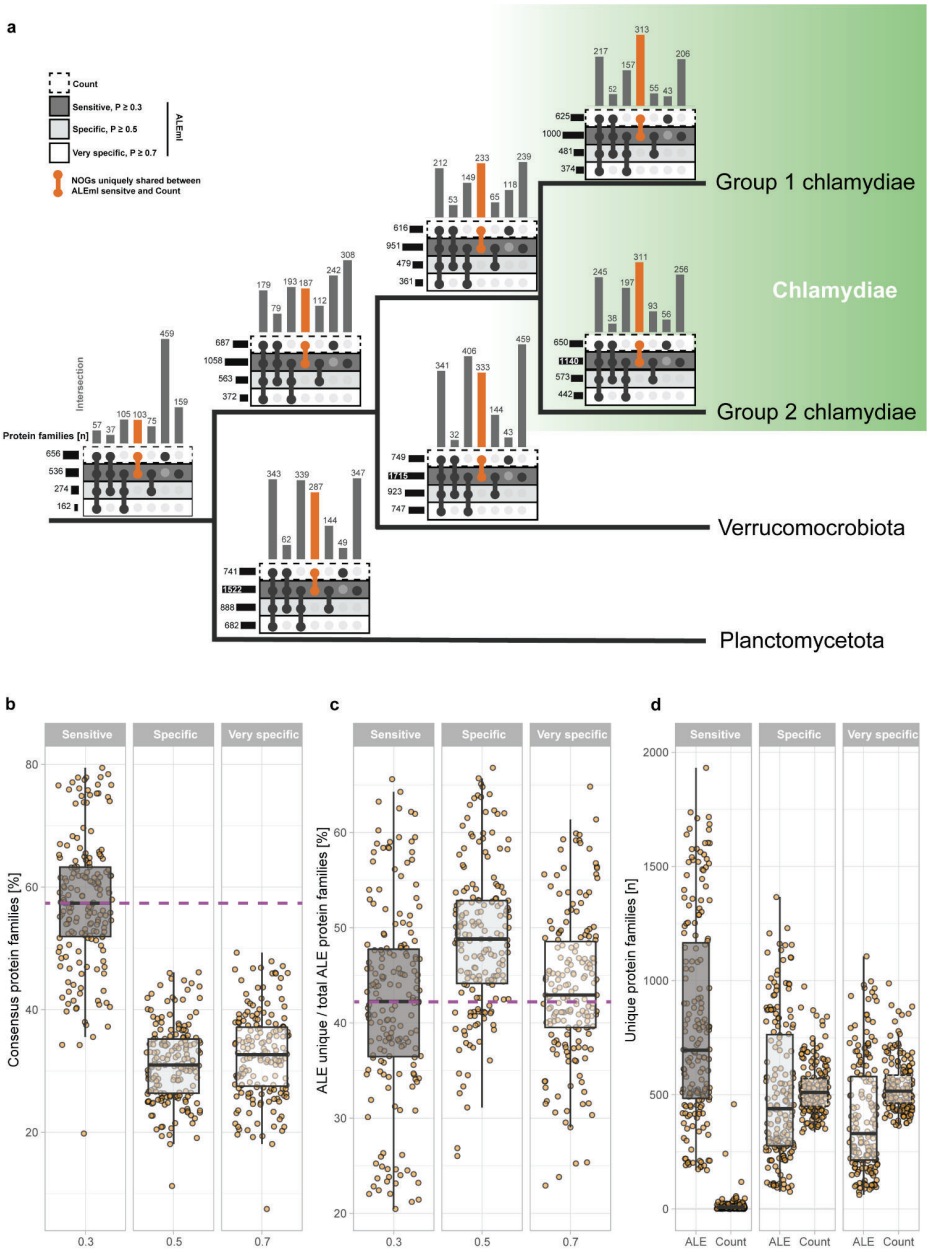


Figure S9. Comparison of the gene tree unaware ancestral gene content reconstruction software to ALE at different frequency cutoffs. For comparison we defined the ALE frequency cutoffs at $P \leq 0.3 \leq 0.5 \leq 0.7$ as sensitive, specific, and very specific, respectively. **a.** Plots showing intersections of protein families (y-axis) inferred to be present in early PVC ancestors using different reconstruction methods and mapped to a schematic phylogenetic tree based on Figure 1. The X-axis depicts the total inferred protein families per method and ancestor, respectively. **b.** Boxplot depicting percentage protein families inferred using both Count and the respective ALE cutoffs relative to all inferred protein families per PVC ancestors with both methods. Purple line indicates the median value of the sensitive cutoff. **c.** Percentage of uniquely inferred protein families of all inferred protein families in all PVC ancestors with the respective ALE cutoff in comparison to Count. Purple line indicates the median value of the sensitive cutoff. **d.** Total unique protein families inferred with ALE and Count with the respective ALE cutoffs.

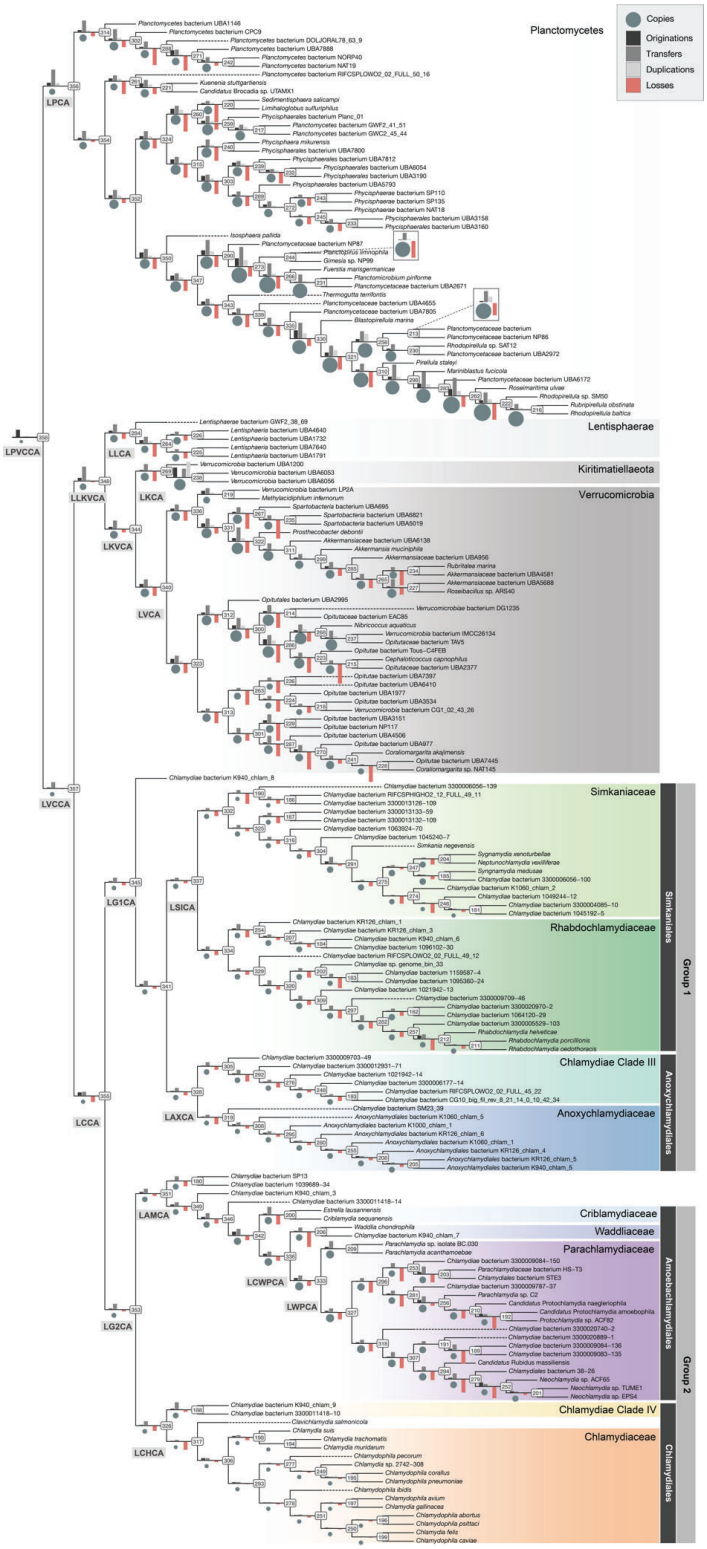

Planctomycetes
Copies
Originations
Transfers
Duplications
Losses
LPCA
LPVCCA
LLCA
LLKVCA
LKCA
LKVCA
LVCA
LVCCA
LG1CA
LSICA
LCCA
LAXCA
LG2CA
LAMCA
LCWPCA
LWPCA
LCHCA
Lentisphaerae
Kiritimatiellaeota
Verrucomicrobia
Simkaniaceae
Rhabdochlamydiaceae
Chlamydiae Clade III
Anoxychlamydiaceae
Criblamydiaceae
Waddliaceae
Parachlamydiaceae
Chlamydiae Clade IV
Chlamydiaceae
Simkaniales
Anoxychlamydiales
Amoebochlamydiales
Chlamydiales
Group 1
Group 2

Figure S10. Overview of ancestral events and genome sizes across all PVC nodes, alongside node numbers. Barplots at each branch indicate origination, transfer, and duplication events in the positive direction (grey bars; see legend), and loss events in the negative direction (red bars), while circle size represents inferred ancestral genome size (*i.e.*, number of genes or copies) for each node to the right. The maximum bar size is 1000, several cases with larger numbers of events are capped at this size (in non-Chlamydiae PVC nodes). Taxonomic groups are indicated and chlamydial families coloured. Key ancestors are indicated and abbreviations can be found in Data S7 alongside event counts for each node.

3. SUPPLEMENTARY DATA

Data S1. Overview of prefiltered PVC dataset. Includes genome quality and if genomes were selected for downstream analysis and representative genomes after dereplication of the Chlamydiae ingroup and non-Chlamydiae PVC outgroup, respectively.

Data S2. Overview of PVC genome representatives selected for use. Includes taxonomy, source information (*e.g.*, Genbank identifier), species and strain names, identifiers used in species phylogenies and for ALE analyses, and genome characteristics.

Data S3. Selection of single-copy marker proteins. NOG identifiers and annotations, alongside trimmed alignment length are given for the initial dataset of 116 markers. NOGs removed and retained from the marker protein set are indicated after each of the two rounds of tree refinement. The identifiers are listed for sequences removed based on tree refinement due to either being duplicates or partial sequences, and those that could represent HGT events, contamination, or distant paralogs. Discordance scores for each of the 79 proteins that passed tree refinement are given and those subsequently removed and the 74 selected as the final dataset are indicated.

Data S4. Summary of the stepwise removal of the most compositionally heterogeneous sites from alignments (184, 183, and 180 taxa in Figures 1 and S3-S5) in 1% increments. The corresponding alignment length, percentage of the alignment removed, and number of taxa significantly divergent in composition is given for each step, alongside the $\chi 2$ test score for each taxon. Dark green indicates alignments corresponding to trees shown in Figures 1 and S3-S5, while light green indicates alignments where ML phylogenies were also inferred and used to assess the monophyly of different groups in Figures S3-S5.

Data S5. Overview of all species phylogenies inferred, with the number of taxa, percentage of the total alignment pruned, alignment length, model of evolution, inference method and supports, and where the phylogeny can be found, with corresponding page numbers for Data S6.

Data S6. Uncollapsed trees for all species phylogenies inferred including both maximum-likelihood (both PMSF non-parametric bootstrap and TBE supports) and Bayesian trees (all chains, and relevant consensus trees with posterior probability). Information about each tree can be found in Data S5 by the corresponding tree number (*i.e.*, page number).

Data S7. Summary of key ancestor nodes and the corresponding ancestor abbreviation and included taxonomic groups. Overview of the number of events at each node inferred using different cutoffs (0.1 to 0.9, in 0.05 increments; 0.3 selected) for duplications, transfers, losses, originations, and copies (*i.e.*, ancestral gene copy number).

Data S8. Ancestral genome content reconstructions based on the gene tree unaware method Count. Includes inferred copy number per protein family per node in the species phylogeny, if the inferred copy number was larger than 0.

Data S9. Summary of annotations of gene content of selected PVC ancestors. Includes annotation from EggNOG, PFAM, TIGRFAM, and Interpro databases per protein family in addition to noted manual curation for important proteins per ancestor referred to in Figure 2 and Extended Data Figure 2.

Data S10. Summary of originations and the inferred taxonomy of their HGT donor lineage, which is present in Extended Data Figure 4. Domain, superphylum, and phylum were determined based on different cutoffs of percentage taxa (50, 75, 90, and 100%) in a supported monophyletic clade sister to chlamydial sequences with the given taxonomy, and also in the clade sister to this group (nested). An overview of protein annotations presented in Figures 2-3 and Extended Data Figures S5-S6 is also outlined. Alongside their respective annotations, and inferred copy number (hence presence) across all Chlamydiae ancestors and major PVC ancestors.

4. SUPPLEMENTARY REFERENCES

1. Taylor-Brown, A. *et al.* Metagenomic Analysis of Fish-Associated Ca. Parilichlamydiaceae Reveals Striking Metabolic Similarities to the Terrestrial Chlamydiaceae. *Genome Biol. Evol.* **10**, 2587–2595 (2018).
2. Pillonel, T., Bertelli, C. & Greub, G. Environmental Metagenomic Assemblies Reveal Seven New Highly Divergent Chlamydial Lineages and Hallmarks of a Conserved Intracellular Lifestyle. *Front. Microbiol.* **9**, 79 (2018).
3. Dharamshi, J. E. *et al.* Marine Sediments Illuminate Chlamydiae Diversity and Evolution. *Curr. Biol.* **30**, 1032–1048.e7 (2020).
4. Köstlbacher, S. *et al.* Pangenomics reveals alternative environmental lifestyles among chlamydiae. *Nat. Commun.* **12**, 4021 (2021).
5. Philippe, H. *et al.* Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol.* **9**, e1000602 (2011).
6. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095–1109 (2004).
7. Philippe, H., Delsuc, F., Brinkmann, H. & Lartillot, N. Phylogenomics. *Annual Review of Ecology, Evolution, and Systematics* vol. 36 541–562 (2005).
8. Stairs, C. W. *et al.* Chlamydial contribution to anaerobic metabolism during eukaryotic evolution. *Sci Adv* **6**, eabb7258 (2020).
9. Borel, N. & Greub, G. International Committee on Systematics of Prokaryotes (ICSP) Subcommittee on the taxonomy of Chlamydiae. Minutes of the closed meeting, 5 July 2018, Woudschoten, Zeist, The Netherlands. *International Journal of Systematic and Evolutionary Microbiology* vol. 69 2606–2608 (2019).
10. Greub, G. & Bavoil, P. International Committee on Systematics of Prokaryotes Subcommittee on the taxonomy of Chlamydiae. Minutes of the closed meeting, 7 September 2016, Oxford, UK. *International Journal of Systematic and Evolutionary Microbiology* vol. 68 3683–3684 (2018).
11. Köstlbacher, S., Collingro, A., Halter, T., Domman, D. & Horn, M. Coevolving Plasmids Drive Gene Flow and Genome Plasticity in Host-Associated Intracellular Bacteria. *Curr. Biol.* **31**, 346–357.e3 (2021).
12. Kuo, C., Horn, M. & Stephens, R. S. Order I. Chlamydiales. in *Bergey's Manual of Systematic Bacteriology* (ed. Krieg N, Staley J, Brown D, Hedlund B, Paster B, Ward N, Ludwig W, Whitman W) vol. 4, 2nd ed. 844–845 (Springer, 2011).
13. Kuo, C. & Stephens, R. Family I. Chlamydiaceae. in *Bergey's Manual of Systematic Bacteriology* (ed. Krieg NR, Staley JT, Brown DR, Hedlund BP, Paster BJ, Ward NL,

Ludwig W, and Whitman WB) vol. 4, 2nd ed. 845 (Springer, 2011).
14. Horn, M. Family II. 'Candidatus Clavichlamydiaceae'. in *Bergey's Manual of Systematic Bacteriology* (ed. Krieg NR, Staley JT, Brown DR, Hedlund BP, Paster BJ) vol. 4, 2nd ed. 865 (Springer, 2011).
15. Thomas, V., Casson, N. & Greub, G. Criblamydia sequanensis, a new intracellular Chlamydiales isolated from Seine river water using amoebal co-culture. *Environ. Microbiol.* **8**, 2125–2135 (2006).
16. Rurangirwa, F. R., Dilbeck, P. M., Crawford, T. B., McGuire, T. C. & McElwain, T. F. Analysis of the 16S rRNA gene of micro-organism WSU 86-1044 from an aborted bovine foetus reveals that it is a member of the order Chlamydiales: proposal of Waddliaceae fam. nov., Waddlia chondrophila gen. nov., sp. nov. *Int. J. Syst. Bacteriol.* **49 Pt 2**, 577–581 (1999).
17. Everett, K. D., Bush, R. M. & Andersen, A. A. Emended description of the order Chlamydiales, proposal of Parachlamydiaceae fam. nov. and Simkaniaceae fam. nov., each containing one monotypic genus, revised taxonomy of the family Chlamydiaceae, including a new genus and five new species, and standards for the identification of organisms. *Int. J. Syst. Bacteriol.* **49 Pt 2**, 415–440 (1999).
18. Horn, M. Family VI. Rhabdochlamydiaceae fam. nov. in *Bergey's Manual of Systematic Bacteriology* (ed. Krieg NR, Staley JT, Brown DR, Hedlund BP, Paster BJ, Ward NL, Ludwig W, and Whitman WB) vol. 4, 2nd ed. 873 (Springer, 2011).
19. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
20. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
21. Williams, K. P. *et al.* Phylogeny of gammaproteobacteria. *J. Bacteriol.* **192**, 2305–2314 (2010).

CHAPTER VI

Synthesis

Synthesis

Chlamydiae are a fascinating model system to study endosymbiosis. For one, all so far characterized members employ a strictly endosymbiotic lifestyle. Secondly, their endosymbiotic lifestyle can be traced back to their last common ancestor more than a billion years ago, making them one of a few groups evolving as 'professional' endosymbionts over large timeframes. Thirdly, members of the phylum can infect an array of eukaryotic hosts ranging from humans – as important sexually transmitted pathogens – to unicellular eukaryotes found in virtually every environment. In this work I studied the hidden potential of chlamydiae in the environment and reconstructed chlamydial ancestral states and evolutionary trajectories to better understand long-term evolution of endosymbiont genomes.

Chapter III was started by a naive idea: I wanted to study ancestral states of chlamydiae genomes by using methods to reconcile the chlamydiae species tree with the individual histories of all genes in the chlamydial genomes. Upon collecting the dataset, calculating trees, and performing the reconciliation analysis over a large parameter space, I needed a smaller dataset to test different parameters for their effect on inferred ancestral genome states. Previous work suggested the existence of a plasmid in the last chlamydial common ancestor (LCCA) (Collingro et al. 2011) based on phylogenetic evidence of one plasmid gene that is conserved on most extant chlamydial plasmids. In my work I indeed uncovered a gene set underpinning an ancestral plasmid hosted by LCCA more than one billion years ago and could show the extrachromosomal origin of key plasmid genes. I could further model evolutionary trajectories of chlamydiae plasmids in different families, mirroring the genome evolutionary trajectories of their host families and highlighting their importance for HGT despite rare complete plasmid transfers. My findings suggest that plasmids provide evolutionary benefits for chlamydiae that mediate degenerative evolution of endosymbiont genomes through mechanisms summarized in a very recent literature review (Rodríguez-Beltrán et al. 2021), e.g. increased genetic variability and gene flow.

For the future it would be of interest to extend the added genome diversity from metagenomics to plasmids. Plasmids are, however, often not included in metagenome-assembled genomes due to differing copy numbers and deviating nucleotide signatures from the host chromosome. In an effort to identify environmental chlamydiae plasmids, one could apply recent bioinformatic techniques like PlasFlow

(Krawczyk, Lipinski, and Dziembowski 2018) to first retrieve candidate plasmid sequences from metagenomic sequence samples known to contain chlamydiae. In a second step one could use the multiple sequence alignments inferred for the work presented in chapter lII to search for chlamydial core plasmid genes on the extracted plasmid candidates. Hidden markov models (HMMs) of highly conserved and specific chlamydial plasmid genes like the integrase Pgp8 could yield high confidence identification of candidate chlamydial plasmids. Phylogenetic analysis of plasmid core genes could then be used to further strengthen the classification and infer evolutionary relatedness to known chlamydial plasmids. While explorative in nature, such a screening of the chlamydial plasmidome could shed light on a second, so far hidden layer to chlamydial lifestyle in the environment, and might reveal additional metabolic potential, host and community interaction or resistance mechanisms.

Aside from a comparative genomic perspective, little is known about chlamydiae plasmids with respect to their in vivo function and importance for the chlamydial lifestyle outside of the small Chlamydiaceae plasmids. The much larger plasmids of environmental chlamydiae like *Simkania negevensis* or *Protochlamydia naegleriophila* encode ~18 times more genes and could be involved in a variety of functions, including infection, host range, HGT mediation, and defense against heavy metals or viruses, among others. However so far no studies have focussed on the *in vivo* effect of these plasmids.

It would be interesting to perform controlled experiments of the infection cycle of plasmid-carrying environmental chlamydiae that are well established in the lab, for example *Simkania negevensis*. Investigating transcription profiles of plasmid and chromosomal genes over the course of infection and different life stages as performed for the plasmid-free *Protochlamydia amoebophila* by König and colleagues (König et al. 2017) could yield valuable first insights into plasmid dynamics and putative functions. Down the road it would be interesting to investigate the mobility of chlamydiae plasmids. Given selection pressures such as heavy metal stress, could one initiate plasmid transfer between co-cultures of plasmid-carrying and plasmidless populations of, for example, *Simkania negevensis* and *Protochlamydia amoebophila*, respectively? Although I did not observe evidence for frequent exchanges of plasmids in our evolutionary analysis, they likely occur occasionally and are very important for endosymbionts as they represent an influx of genetic potential in a (compared to free-living bacteria) HGT-limited niche.

The research in chapter IV was only possible due to the abovementioned explorative power of metagenomics – the extraction and sequencing of environmental DNA – and even more so due to the willingness of researchers from many different countries to share their produced data. We knew from previous studies that chlamydiae are immensely more diverse than initially thought and, matching their evolutionary diversity, are present in all kinds of environments, with large reservoirs in aquatic environments. However most of chlamydiae diversity was only known through a single region of the 16S ribosomal RNA gene, lacked genomic representation, and had not been cultured in the laboratory.

With the MAGs we obtained from our collaborators at the Joint Genome Institute, I could for the first time analyze genome sequences of some of the unknown lineage in the chlamydial species tree. This analysis significantly expanded the represented chlamydial species diversity and especially increased representation of chlamydiae from aquatic ecosystems. Interestingly, analysis of the chlamydiae pangenome uncovered a highly conserved chlamydiae symbiont core genome, but a significantly more dynamic accessory genome in environmental lineages compared to the animal pathogenic chlamydiae. This finding already indicated that gene transfer in environmental chlamydiae might not be as limited as originally expected for obligate intracellular symbionts. Together with the more widespread potential of anaerobic metabolism in chlamydiae, these findings had important implications not only on the biology of extant chlamydiae, but also on the ancestral chlamydial lifestyles which were explored in chapter V. Finally, I queried the 16S rRNA gene sequences encoded on our extended dataset of chlamydiae genomes against publicly available 16S rRNA gene amplicon datasets. I learned that (1) chlamydial families have lineage-specific preferences for certain environments and (2) these families are not necessarily members of the rare biosphere, i.e. can have a relative community abundance larger than 0.1% (Lynch and Neufeld 2015). In line with our metabolic reconstruction, I could show that MCF-E, a putative anaerobic family of chlamydiae, does indeed reach highest relative abundance in anoxic marine water column habitats.

It is quite intriguing that chlamydiae can reach high relative abundance in several environments. These numbers are definitely not an exact representation of their actual abundance in the environment, as most 16S rRNA gene surveys still rely on PCR-based sequence amplification and therefore need oligonucleotide primers. However, amplicon-based approaches do not target all organisms in a community equally well. Primers are designed to cover as much diversity of a target group as possible however

this usually does not cover every target. Indeed, some organisms might be preferentially amplified, while others might be missed entirely. Chlamydiae fall into the latter fraction and are poorly covered in many of the widely used general 16S rRNA primer sets (Dharamshi et al. 2020).

We can therefore assume that our estimates of chlamydiae diversity and abundance in the environment are quite conservative. The sometimes large observed abundance of chlamydiae as obligate endosymbionts of eukaryotes in some environments opens up questions about their role in these ecosystems. While the field will benefit from the added information of environmental chlamydiae genomes, metabolic potential, and environmental distribution, the host organisms with which these chlamydiae associate remains a big unknown We can further assume that most environmental chlamydiae associate with protists (Collingro, Köstlbacher, and Horn 2020). However, the term *protist* summarizes an incredibly diverse and largely unexplored set of organisms (Singer et al. 2021) that are almost ubiquitously found in the environment and known to be key players in primary production, element cycling, and trophic networks of soil and aquatic environments (Mitra et al. 2016). Chlamydiae could therefore have important regulatory functions in ecosystems. These do not necessarily have to be negative regulatory effects expected from intracellular parasites. For example, the cultured environmental chlamydia *Protochlamydia amoebophila* has been shown to protect its protist host from infection by the parasitic *Legionella pneumophila* (Maita et al. 2018; König et al. 2019). To study the effect of this interaction with the host, and later its effect on the ecosystem, characterizing these associations under controlled conditions in the laboratory through cultivation of chlamydiae with their native host will therefore be an essential step. While cultivation tends to be a long-winded process, alternatively cultivation-free single cell sorting or microfluidic approaches coupled to genomics, transcriptomics, and imaging could be applied to quantify the effect of chlamydiae on their hosts in natural communities (Alacid and Richards 2021).

In chapter V, I could capitalize on the knowledge gained in chapters III and IV, combining ancestral-state reconstruction of chlamydiae gene content with the expanded genomic diversity gained in chapter IV. In a collaborative effort with the universities in Wageningen (the Netherlands), and Uppsala (Sweden), we could take a closer look at the chlamydial species tree, reconstruct ancestral genomes and dive into gene-transfer dynamics in chlamydiae. Due to the added sampling of anaerobic chlamydial lineages in Jennah Dharamshi and colleagues' work (Dharamshi et al. 2020)

and my work presented in chapter IV, we could shine a new light at the LCCA. While, based on gene content, we still inferred it to already be an endosymbiotic bacterium, we now predict that it contained genes indicative of a facultative anaerobic lifestyle. It was intriguing to see that the models suggested the origin of many enzymes related to higher oxygen concentrations, like cytochrome-o-ubiquinol oxidase complex, were lineage-specific gains occurring late in chlamydial evolution. This finding exemplifies a general trend of continued genome expansion we observed in several ancestors of Group-2 chlamydiae. In Chapter IV I already showed that environmental chlamydiae tend to have significantly more open pangenomes than the animal pathogenic Chlamydiaceae, suggesting higher gene transfer rates in the former. Using the species-tree-aware models of gene evolution in chlamydiae, we could now detect significantly higher gene transfer rates in several families of environmental chlamydiae than in the animal pathogenic Chlamydiaceae, corroborating and more precisely delineating the previous finding. These lineages further have gene transfer rates that are indistinguishable from the predominantly free-living relatives of chlamydiae within the PVC superphylum.

An open question remains the nature of many gene gains identified as originations, where no homologs could be identified when querying public databases. These could either represent *bona fide* gene births or, and this seems likely in most cases, proteins that originated from HGTs but their homologs cannot be detected by the search strategies applied. This could be due to their origin from yet unsampled lineages that are not represented in sequence databases, or due to their large divergence since their acquisition so that their homology signal is too weak to detect. While the first issue is hard to tackle, a strategy to identify remote homologs seems feasible. By employing more sensitive search methods based on models of sequence evolution (HMM), or even HMM-HMM searches (Söding 2005) that align such query models to subject models to identify remote homologs of these proteins and study their evolutionary history. For example, using HMM based searches I could identify a putative origin of the plasmid integrase genes *pgp7/8* in viruses, while local alignments alone were not sensitive enough. It would be intriguing to apply such remote homology detection approaches on inferred events in the Chlamydiaceae last common ancestor, where we could not infer homologs outside of Chlamydiae for 68 out of almost a hundred gene families identified as originations. Modeling the origin and evolution of such gene families could shed new light on the drivers of Chlamydiaceae diversification that made them such potent pathogens.

If I had to pick, the most exciting finding of this thesis to me is the evidence for largely unexplored expansive genome evolution in obligate endosymbionts in the environment, i.e. chlamydiae likely predominantly associated with protist hosts. Chapters III and V demonstrate expansion of plasmid and chromosomal gene content in some chlamydial lineages. In both cases, although better modeled for chromosomal gene content in Chapter V, this seems to be largely driven by HGT, which in general is thought to be limited in obligate endosymbionts. One example of recent HGT, which seems to be most memorable from Chapter IV for most people - the translated title of the Austrian Press Agency release of the published manuscript was “Some chlamydia live from sunlight instead of inflaming your genitals” - was the acquisition of genes for light-driven ATP synthesis in some antarctic lake-dwelling chlamydiae. In the future it will be interesting to better study the parameters shaping genome evolution in different chlamydial lineages with a focus on our old friends, the Parachlamydiaceae and sister lineages. We do not know yet their effective population sizes and selection working on their core and accessory genomes. From our current perspective, since their ancestral genome content expansion, they maintain larger genomes than their animal associated relatives, likely due to larger effective population sizes, higher HGT rates, and weaker genetic drift. Now, it would be of value to investigate effective population sizes, recent gene flow, and selection on chlamydial genomes to understand speciation of extant environmental chlamydiae. While these are hard issues to tackle, I believe with the improvement of long-read sequencing and cell-sorting techniques, it should be reasonable, yet by no means easy, to give such longitudinal investigations of chlamydiae populations a shot.

To end, it was an exciting and enjoyable ride to look into the hidden diversity and more than a billion years-old history of these fascinating chlamydial organisms. It seems that, based on the explosion of new chlamydial lineages that are being found these days, the future will hold a wide range of exciting new findings explaining and underpinning the importance of these great, tiny organisms.

References

Alacid, Elisabet, and Thomas A. Richards. 2021. “A Cell-Cell Atlas Approach for Understanding Symbiotic Interactions between Microbes.” *Current Opinion in Microbiology* 64 (December): 47–59.

Collingro, Astrid, Stephan Köstlbacher, and Matthias Horn. 2020. “Chlamydiae in the Environment.” *Trends in Microbiology*, June. https://doi.org/10.1016/j.tim.2020.05.020.

Collingro, Astrid, Patrick Tischler, Thomas Weinmaier, Thomas Penz, Eva Heinz, Robert C. Brunham, Timothy D. Read, et al. 2011. “Unity in Variety--the Pan-Genome of the Chlamydiae.” *Molecular Biology and Evolution* 28 (12): 3253–70.

Dharamshi, Jennah E., Daniel Tamarit, Laura Eme, Courtney W. Stairs, Joran Martijn, Felix Homa, Steffen L. Jørgensen, Anja Spang, and Thijs J. G. Ettema. 2020. “Marine Sediments Illuminate Chlamydiae Diversity and Evolution.” *Current Biology: CB* 30 (6): 1032–48.e7.

König, Lena, Alexander Siegl, Thomas Penz, Susanne Haider, Cecilia Wentrup, Julia Polzin, Evelyne Mann, Stephan Schmitz-Esser, Daryl Domman, and Matthias Horn. 2017. “Biphasic Metabolism and Host Interaction of a Chlamydial Symbiont.” *mSystems* 2 (3). https://doi.org/10.1128/mSystems.00202-16.

König, Lena, Cecilia Wentrup, Frederik Schulz, Florian Wascher, Sarah Escola, Michele S. Swanson, Carmen Buchrieser, and Matthias Horn. 2019. “Symbiont-Mediated Defense against Legionella Pneumophila in Amoebae.” *mBio* 10 (3). https://doi.org/10.1128/mBio.00333-19.

Krawczyk, Pawel S., Leszek Lipinski, and Andrzej Dziembowski. 2018. “PlasFlow: Predicting Plasmid Sequences in Metagenomic Data Using Genome Signatures.” *Nucleic Acids Research* 46 (6): e35.

Lynch, Michael D. J., and Josh D. Neufeld. 2015. “Ecology and Exploration of the Rare Biosphere.” *Nature Reviews. Microbiology* 13 (4): 217–29.

Maita, Chinatsu, Mizue Matsushita, Masahiro Miyoshi, Torahiko Okubo, Shinji Nakamura, Junji Matsuo, Masaharu Takemura, Masaki Miyake, Hiroki Nagai, and Hiroyuki Yamaguchi. 2018. “Amoebal Endosymbiont Neochlamydia Protects Host Amoebae against Legionella Pneumophila Infection by Preventing Legionella Entry.” *Microbes and Infection / Institut Pasteur* 20 (4): 236–44.

Mitra, Aditee, Kevin J. Flynn, Urban Tillmann, Ravenjohn A., David Caron, Diane K. Stoecker, Fabrice Not, et al. 2016. “Defining Planktonic Protist Functional Groups on Mechanisms for Energy and Nutrient Acquisition: Incorporation of Diverse Mixotrophic Strategies.” *Protist* 167 (2): 106–20.

Rodríguez-Beltrán, Jerónimo, Javier DelaFuente, Ricardo León-Sampedro, R. Craig MacLean, and Álvaro San Millán. 2021. “Beyond Horizontal Gene Transfer: The Role of Plasmids in Bacterial Evolution.” *Nature Reviews. Microbiology* 19 (6): 347–59.

Singer, David, Christophe V. W. Seppey, Guillaume Lentendu, Micah Dunthorn, David Bass, Lassâad Belbahri, Quentin Blandenier, et al. 2021. “Protist Taxonomic and Functional Diversity in Soil, Freshwater and Marine Ecosystems.” *Environment International* 146 (January): 106262.

Söding, Johannes. 2005. “Protein Homology Detection by HMM-HMM Comparison.” *Bioinformatics* 21 (7): 951–60.

APPENDIX

Abstract

Zusammenfassung

Abstract

Chlamydiae are a phylum of highly successful strictly intracellular bacteria that are found in diverse eukaryotic hosts. Some representatives of the family Chlamydiaceae, like *Chlamydia trachomatis,* have been known as major human pathogens for more than a century. However, chlamydiae are almost ubiquitously found in the environment where they associate with protists and a wide range of animals. These chlamydiae, collectively referred to as environmental chlamydiae, are estimated to be highly diverse and abundant. While much of the research has focussed on animal pathogenic chlamydiae, our knowledge on the hidden diversity of chlamydiae in the environment is very limited. Chlamydiae are an ancient group of bacteria that adapted an intracellular lifestyle more than a billion years ago. This work uses genomic and phylogenetic approaches to uncover the hidden genome diversity and biology of uncultured chlamydiae, and to model long evolutionary history. Firstly, I studied the evolution of chlamydiae plasmids uncovering evidence for a plasmid in the ancestor of all chlamydiae that has since coevolved with chlamydiae. However, many plasmid genes showed increased mobility between chlamydial plasmids and chromosomes, leaving a pronounced impact on chlamydial genomes. Endosymbionts tend to be isolated from horizontal gene transfer (HGT) and therefore can suffer from degenerative effects, illustrating the importance of chlamydiae plasmids that might mitigate this effect. Secondly, I used pangenomics to elucidate the hidden diversity of uncultured chlamydiae from diverse environments and infer their biology and distribution. This work almost doubled chlamydial genome-represented diversity and uncovered a surprisingly widespread potential for anaerobic metabolism in several chlamydial groups. Thirdly, I collected a genome dataset of yet unmatched chlamydial diversity to thoroughly examine chlamydial gene content evolution and reconstructed ancestral chlamydiae genomes. Using state of the art phylogenetic methods I presented evidence underpinning and more accurately delineating an endosymbiotic lifestyle already in the chlamydial last common ancestor. However, I inferred a facultative anaerobic lifestyle and lineage specific genome content expansions that are yet undescribed for obligate

endosymbionts. The analyses presented here have far reaching implications for our understanding on chlamydiae diversity and evolution.

Zusammenfassung

Chlamydien sind ein Phylum äußerst erfolgreicher, obligat intrazellulärer Bakterien, die in verschiedenen eukaryotischen Wirten auftreten. Einige Vertreter der Familie Chlamydiaceae, wie *Chlamydia trachomatis,* sind seit mehr als einem Jahrhundert als wichtige Krankheitserreger der Menschen bekannt. Chlamydien sind jedoch fast überall in der Umwelt anzutreffen, wo sie mit Protisten und einer Vielzahl von Tieren assoziiert sind. Man schätzt, dass diese Chlamydien, die als Umweltchlamydien zusammengefasst werden, sehr divers und häufig sind. Während sich ein Großteil der Forschung auf die tierische Krankheitserreger unter den Chlamydien konzentriert hat, haben wir ein nur sehr begrenztes Wissen über die verborgene Vielfalt von Chlamydien in der Umwelt. Chlamydien sind eine uralte Gruppe von Bakterien, die sich vor mehr als einer Milliarde Jahren an eine intrazelluläre Lebensweise angepasst haben. In dieser Arbeit werden genomische und phylogenetische Ansätze verwendet, um die verborgene Genomvielfalt von unkultivierten Chlamydien aufzudecken und deren lange Evolutionsgeschichte zu modellieren. Erstens untersuchte ich die Evolution der Chlamydien-Plasmide und fand Hinweise auf ein Plasmid im Vorfahren aller Chlamydien, das sich seitdem mit den Chlamydien mitentwickelt hat. Viele Plasmidgene wiesen jedoch eine erhöhte Mobilität zwischen Chlamydienplasmiden und Chromosomen auf, was deutliche Spuren auf Chlamydienchromosomen hinterlassen hat. Endosymbionten sind in der Regel von horizontalem Gentransfer (HGT) isoliert und leiden daher unter degenerativen Effekten, was die Bedeutung von Chlamydienplasmiden, welche diese Effekte mildern könnten, verdeutlicht. Zweitens haben wir metagenomische Methoden und Pangenomanalysen eingesetzt, um die verborgene Vielfalt von unkultivierten Chlamydien aus diversen Habitaten zu ergründen und Rückschlüsse auf ihre Biologie und Verbreitung zu ziehen. Die Vielfalt von Chlamydiengenomen wurde durch diese Arbeit fast verdoppelt und deckte ein überraschend weit verbreitetes Potenzial für anaeroben Stoffwechsel in mehreren Chlamydiengruppen auf. Drittens stellte ich einen Genomdatensatz

zusammen, der eine bisher unerreichte Chlamydienvielfalt repräsentiert, um die Genomevolution von Chlamydien gründlich zu untersuchen und die Genome der Vorfahren der Chlamydien zu rekonstruieren. Mit modernsten phylogenetischen Methoden lieferte ich weitere Hinweise und genauere Modelle für einen intrazellulären Lebensstil des letzten gemeinsamen Vorfahren der Chlamydien. Ich schloss jedoch auf einen fakultativ anaeroben Metabolismus in diesem Vorfahren und auf gruppenspezifische Genomexpansionen, die für obligate Endosymbionten so noch nicht beschrieben wurden. Die hier vorgestellten Analysen haben weitreichende Auswirkungen auf unser Verständnis der Diversität und Evolution der Chlamydien.