universität wien

DISSERTATION

Titel der Dissertation

„Evolutionary genomics of the *Chlamydiae*“

verfasst von

Daryl Domman

angestrebter akademischer Grad

Doctor of Philosophy (PhD)

Wien, 2015

Studienkennzahl lt. Studienblatt:	A 794 685 437
Dissertationsgebiet lt. Studienblatt:	Biology
Betreuerin / Betreuer:	Univ.-Prof. Dr. Matthias Horn

Table of Contents

Chapter I Introduction **5**

Chapter II Overview Of Publications/Manuscripts **18**

Chapter III Massive expansion of ubiquitination-related gene families Within the *Chlamydiae* **22**

Chapter IV Following the footsteps of chlamydial gene regulation **51**

Chapter V Plastid establishment did not require a chlamydial partner **97**

Chapter VI Conclusion and disccusion **125**

Chapter VI Summary and Zusammenfassung **133**

Appendix Acknowledgements and Curriculum Vitae **137**

Chapter I

Introduction

Introduction

Bacteria are truly fascinating creatures. They occupy nearly every ecological niche conceivable, from Antarctic soil, deep-sea hydrothermal vents, and the human gut. Host-associated bacteria occupy a rather intriguing niche, as their evolution and ecology is inextricably linked to that of the host. Whether these bacteria are true symbionts or more parasitic, the exploitation of the eukaryotic cell as a niche has profound implications at a basic level of understanding how organisms interact, but also on a more anthropocentric level has lent much insight into pathogen interactions. This thesis focuses on a unique group of obligate intracellular bacteria, the phylum *Chlamydiae*, which have global importance in human health. Utilizing the power of genomics we uncovered answers to key questions about the evolution and ecology of this phylum.

The amazing *Chlamydiae*

In 1903 the German radiologist Ludwig Halberstädter and the Austrian zoologist Stanislaus von Prowazek joined a research expedition to the island of Java, Indonesia to unravel the agent responsible for syphilis. While in the city of Jakarta in 1907, they took scrapings from an eye infection, which lead to the discovery of peculiar inclusions within the cytoplasm of these cells, which were called 'Halberstädter- Prowazek bodies' (Black 2013). Back in Berlin these inclusions had unusual Giemsa-staining patterns in which they observed small, condensed particles surrounding the nucleus of infected cells. These "Chlamydozoa" were so named from the ancient Greek word "chlamys" which means cloak-like mantle, in that these organisms appeared to "cloak" the nucleus of the infected cell. Although Halberstädter and von Prowazek thought these organisms were protozoa and not bacteria, they had nevertheless isolated the causative agent of "trachoma", meaning 'rough eye'. The "Chlamydozoa" then embarked on a complicated taxonomic journey over the next 50 years, as they were reclassified as not protozoan, but as a virus due to their intracellular lifestyle. It wasn't until the late 1950's and early 1960's that these organisms were correctly classified as bacteria. Despite the identity crisis, the name "chlamy" was kept throughout and thus today the causative agent of blinding trachoma is known as *Chlamydia trachomatis*.

Fascinating history aside, *Chlamydia trachomatis* remains particularly relevant today as it affects nearly 84 million people globally; a number that is greater than all other infectious

disease combined. In the developing world, *C. trachomatis* is responsible for the largest cause of preventable blindness in the world (i.e. blinding trachoma), affecting around 8 million people. Within the developed world, *C. trachomatis* is the leading cause of bacterial sexually transmitted infections, with approximately 92 million new cases a year. Additionally, *C. trachomatis* is part of a family of chlamydial organisms called the *Chlamydiaceae*, which contain other important human and animal associated pathogens. *Chlamydia pneumoniae*, for example, is a causative agent of pneumonia in humans and also infects a wide variety of other mammals, marsupials, reptiles, and amphibians (Horn 2008; Taylor-Brown et al. 2015). *Chlamydia abortus* an agent responsible for fetal abortion in a variety of animals (cattle, horse, rabbit, mice), and *Chlamydia psittaci* the agent responsible for chlamydiosis in avian hosts (Horn 2008; Taylor-Brown et al. 2015). However, for nearly the entire century since the characterization of *Chlamydia trachomatis*, the perception was that this handful of pathogens was limited to primarily to human and animal hosts. A big surprise came in 1997 when chlamydia-like organisms were discovered within amoeba cells isolated from human nasal mucosa (Amann et al. 1997), suggesting that these may be a novel reservoir for chlamydial organisms. Today, there are seven other described families of chlamydiae that are associated with a dizzying array of eukaryotic hosts (Figure 1). However, a recent study of the diversity of chlamydiae estimated that there may be upwards of 350 chlamydial families (Lagkouvardos et al. 2013). These colloquially termed “environmental chlamydia” (Figure 1) are associated with hosts ranging from a wide variety of protists, insects, arthropods, fish, enigmatic marine worms, cattle, and possibly even humans (Horn 2008; Lagkouvardos et al. 2013; Taylor-Brown et al. 2015).

With the advent of whole genome sequencing, it became possible to study the diversity between organisms on a sequence level. The first chlamydial genome sequenced was *Chlamydia trachomatis*, which revealed a reduced genome consisting of ~1 Mb and displayed limited metabolic capabilities, a reflection of the long evolutionary history of the chlamydiae exploiting host cell resources (Stephens et al. 1998). Soon after the 1.2 Mb genome of *Chlamydia pneumoniae* was sequenced and revealed over 200 genes not present in *C. trachomatis* (Kalman et al. 1999). As of writing, there are twelve species of *Chlamydiaceae* with sequenced genomes, many of which with multiple sequenced strains. The amoeba-associated *Protochlamydia amoebophila* was the first chlamydiae sequenced outside of the *Chlamydiaceae*, which revealed a genome double in size (~2.4 Mb) to that of *C. trachomatis* or *C. pneumoniae*, but still shared many of the genomic characteristics such as the presence of

nucleotide and ATP transporters to scavenge these resources from the host, host manipulation machinery (Type III secretion system), and reduced metabolic capabilities (Horn et al. 2004). An additional five genomes, adding genomes to two unrepresented families (*Waddliaceae* and *Simkaniaceae*) appeared fairly recently (Greub et al. 2009; Bertelli et al. 2010; Collingro et al. 2011) allowing much needed insights into the genomic diversity of these organisms. Comparative genomic analysis of the phylum revealed 560 genes were conserved throughout the phylum, and also that there was extensive gene content variation within the environmental chlamydia (Collingro et al. 2011). For instance, the environmental chlamydia isolated as a contaminant of cell culture, *Simkania negevensis*, harbors a genome of 2.5 Mb and contains more unique genes (i.e. those that are not found in any other member of the phylum) than the total number of genes present in *Chlamydia trachomatis* (1340 to 894, respectively) (Collingro et al. 2011).

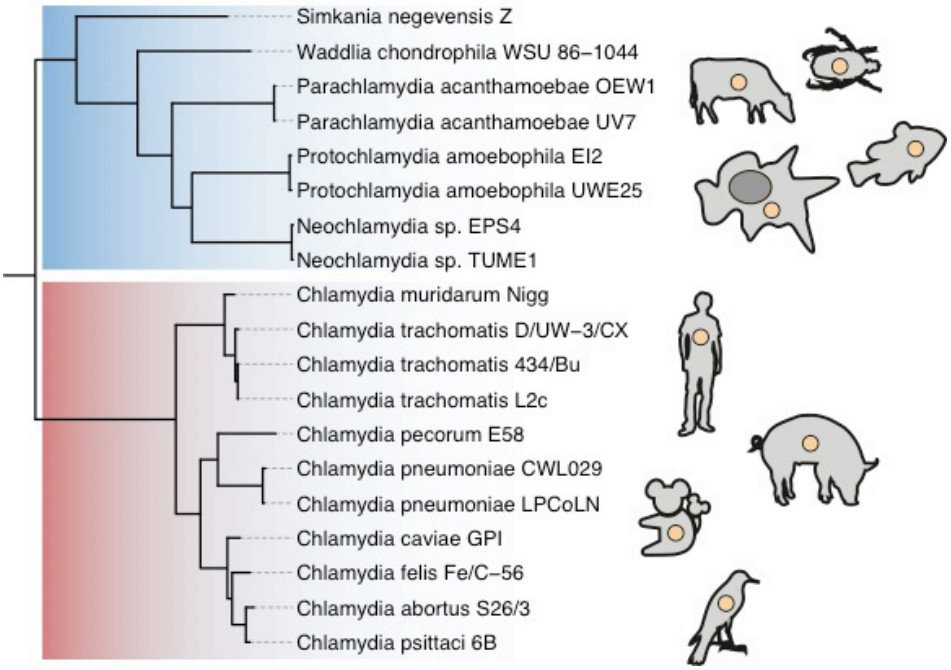


Figure 1. Phylogeny of the *Chlamydiae* and host diversity. The phylum can be dived into the family *Chlamydiaceae* (in red), which includes many human and animal pathogens, and the environmental chlamydia (in blue), which have a tremendous host range. Many of the fully

sequenced environmental chlamydiae are associated with free-living protists. The species tree phylogeny was adapted from Domman et al. (2014).

Despite a tremendous diversity in host range and phylogenetic breadth, a paramount unifying feature of all chlamydiae, aside from their obligate host-association, is that of a bi-phasic developmental cycle (Figure 2). The cycle consists of the uptake of the extracellular, non-replicative form of chlamydia, termed elementary bodies (EBs) by a eukaryotic host cell. Once within a host cell, the EBs differentiate into the fully metabolically active and replicative form, called reticulate bodies (RBs) within an inclusion membrane. RBs divide within the chlamydial inclusion until, via a still unknown mechanism, they begin to differentiate back into EBs. The EBs are then released from the host cell either as a result of host cell lysis, or via extrusion of host cell vesicles. Though this feature unites the *Chlamydiae*, a developmental cycle is certainly unique among obligate, intracellular bacteria.

As *Chlamydiae* represent some of the most successful groups of bacteria that can exploit the intracellular niche of eukaryotes they remain a fascinating case study on the evolutionary path towards the specialization of this habitat. The main theme of this thesis revolves around understanding this evolutionary path. In Chapter One we explore how changes in gene content have lead to the exploitation of different hosts. In Chapter two we examine the commonalities and differences in how chlamydial organisms regulate gene expression. In Chapter Three we explore an intriguing hypothesis that ancient chlamydia donated key genes that facilitated the endosymbiotic capture of a cyanobacteria within the proto-plant host.

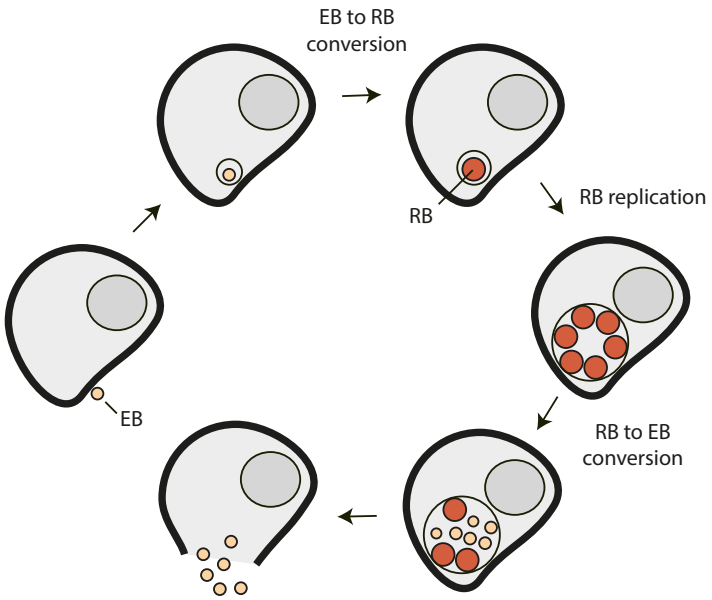


Figure 2. Chlamydial developmental cycle. The bi-phasic chlamydial developmental begins when an extracellular elementary body (EB) is taken up by a host cell. Once internalized, the EB is encapsulated by host derived edosomal membrane to form an inclusion. Within the inclusion, the EB differentiates into the fully metabolic and replicative form, termed reticulate bodies (RB). RBs divide and the inclusion increases in size. The signal for RB to EB conversion is unknown, but RBs asynchronously convert back to EBs. The cycle is completed when EBs are released from the host cell, typically as a result of lysis.

Gene family evolution

There is a great disparity of genome size among bacteria, ranging from the 13 Mb genome of the soil bacterium *Ktedonobacter racemifer* (Chang et al. 2011) to the 144 Kb genomes of sap-feeding insect symbionts (McCutcheon and Dohlen 2011). Even between organisms within the same species can differ in hundreds of genes. This was typified when the genomes of three strains of *Escherichia coli* were sequenced and though they were identical at the 16S rDNA level, they only shared 39% of their gene content (Welch et al. 2002). These genomic fluctuations are the cumulative result of mutation, gene flow, genetic drift, and selection. As mentioned above, the members of the *Chlamydiae* have reduced genomes, especially the animal and human pathogens. Genomic reduction in host-restricted bacteria is commonly attributed to the effect of small population size and the inherent deletion bias in most bacteria

(McCutcheon and Moran 2012). The power of genetic drift is greatly increased as population size and the ability to recombine is limited. As such, deleterious mutations are fixed in these genomes in a higher proportion than in large populations where purifying selection and recombination can reverse these effects. The cumulative effect of this population structure is that genes are inactivated and deleted, even if they might be even mildly advantageous (McCutcheon and Moran 2012). Thus, the cumulative effect on a genomic level for obligate host-associated microbes is to have reduced genomes.

As a corollary of genome reduction, these host-restricted microbes tend to have little redundancy in their gene content. That is to say, there are few gene duplications or gene families that have multiple members. Most of the genome, therefore, consists of single copy genes. The process of genomic reduction, as described above, is well known for these microbes, but as the genome sizes in members of the *Chlamydiae* span from 1 Mb to over 3Mb, we queried if genomic expansions, either via gene duplication or horizontal gene transfer, were also contributing to the overall genomic architecture within this phylum. Using all of the available fully sequenced chlamydial genomes, we clustered all proteins into gene families. Our analysis of gene family histories, suggest that gene family expansions, have had pronounced effects on gene content within the phylum. We discovered that the largest gene families within the phylum are largely the result of gene duplication events and appear to evolve via a unique mode of gene family evolution (rapid gene birth-and-death model). We find that variations copy number of gene between related individuals might suggest that non-adaptive processes, such as genomic drift, influence the evolution of large gene families. This mode of evolution may represent a previously unexplored mechanism by which isolated bacterial populations, such as bacterial symbionts, diversify in gene content and adapt to novel ecological niches.

Gene regulation in *Chlamydiae*

All organisms control the expression of genes in response to environmental and developmental signals. As mentioned, all *Chlamydiae* undergo a complex temporally regulated developmental cycle. These developmental transitions require massive tracts of genes to be activated or silenced in a precise manner (Belland et al. 2003; Nicholson et al. 2003; Mäurer et al. 2007; Albrecht et al. 2011). While there are a number of ways that bacteria regulate gene expression, such as RNA silencing and DNA topological properties, the primary mechanism is via DNA-binding transcription factors that either activate or repress transcription via interactions with

RNA polymerase. Transcription factors recognize and bind to specific motifs found within the promoter regions of the genes they regulate. Despite having a complex developmental cycle, the *Chlamydiae* appear to harbor a small number of transcription factors. The third chapter of this thesis explores how we can use comparative genomics to explore the diversity and evolution of regulatory networks within the *Chlamydiae*. In this chapter we provide the most comprehensive list of predicted transcription factors and their evolutionary history within the phylum. Additionally, we constructed the first predicted co-regulatory networks for all fully sequenced chlamydial genomes (n=17) and explored the similarities and differences between these networks at multiple taxonomic levels. This analysis provides the first comprehensive picture of gene regulation in these organisms and offers a unique perspective to this otherwise under-explored area of chlamydial biology.

Chlamydiae and the evolution of plants

A chapter on the evolution of plants might seem quite out of place within a thesis focusing on the evolution of the *Chlamydiae*; however, rest assured it is perfectly in line with this theme. As with all bacteria, the *Chlamydiae* are not immune to the effects of horizontal gene transfer, and they are donors and receivers both (Collingro et al. 2011; Bertelli and Greub 2012; Clarke et al. 2013). The *Chlamydiae* represent a particularly ancient group of bacteria that have been estimated to be at least 700 million years old (Horn et al. 2004) with more recent estimates of up to 1.6 billion years old (Kamneva et al. 2012). This tremendous time scale would place the last common ancestor of the *Chlamydiae* within timeframe of the origin of the plastid, which is estimated to have occurred around 1.6 -1.9 billion years ago (Yoon et al. 2004; Parfrey et al. 2011). Over the past 15 years there has been a growing body of evidence that ~ 60 of genes of chlamydial origin have been transferred into members of the Archaeplastida (plants) (Brinkman et al. 2002; Huang and Gogarten 2007; Becker et al. 2008; Moustafa et al. 2008; Ball et al. 2013). This discovery has spurred a hypothesis that the ancient eukaryotic cell was cohabitated by both a cyanobacterium and chlamydial endosymbiont and that this tripartite relationship facilitated the endosymbiotic capture and subsequent transformation of the cyanobacterial partner into the modern day plastid (Huang and Gogarten 2007; Ball et al. 2013). The driver of this relationship was that the cyanobacterial partner provided an energy rescue to the host cell that was being parasitized by the chlamydia. Key chlamydial enzymes involved in glycogen metabolism were the integral component linking the photosynthate derived from the cyanobacterial partner to an accessible form for the eukaryotic host (Ball et al. 2013). While

locked in this “ménage à trois”, the hypothesis states that these key chlamydial genes were then transferred to the host and cyanobacterial endosymbiont which subsequently led to the loss of the chlamydial partner (Ball et al. 2013).

The “ménage à trois” hypothesis makes implicit phylogenetic predictions about the directionality of gene transfer. Indeed, the aforementioned studies provide evidence from individual gene trees that imply transfer of these genes from chlamydiae to members of the Archaeplastida. In the third chapter of this thesis we applied sophisticated phylogenetic models to explicitly test the origins of the enzymes implicated in the “ménage à trois” hypothesis. Under these better fitting models we show that there is a mosaic origin for these enzymes, but do not detect a strong argument for a chlamydial origin. Thus, our analysis does not provide compelling evidence that *Chlamydiae* facilitated the plastid endosymbiosis.

References

Albrecht M, Sharma CM, Dittrich MT, Müller T, Reinhardt R, Vogel J, Rudel T. 2011. The transcriptional landscape of Chlamydia pneumoniae. Genome Biol. 12:R98.

Amann R, Springer N, Schönhuber W, Ludwig W, Schmid EN, Müller KD, Michel R. 1997. Obligate intracellular bacterial parasites of acanthamoebae related to Chlamydia spp. Appl. Environ. Microbiol. 63:115–121.

Ball SG, Subtil A, Bhattacharya D, Moustafa A, Weber APM, Gehre L, Colleoni C, Arias M-C, Cenci U, Dauvillée D. 2013. Metabolic Effectors Secreted by Bacterial Pathogens: Essential Facilitators of Plastid Endosymbiosis? Plant Cell Online 25:7–21.

Becker B, Hoef-Emden K, Melkonian M. 2008. Chlamydial genes shed light on the evolution of photoautotrophic eukaryotes. BMC Evol. Biol. 8:203.

Belland RJ, Zhong G, Crane DD, Hogan D, Sturdevant D, Sharma J, Beatty WL, Caldwell HD. 2003. Genomic transcriptional profiling of the developmental cycle of Chlamydia trachomatis. Proc. Natl. Acad. Sci. 100:8478–8483.

Bertelli C, Collyn F, Croxatto A, Rückert C, Polkinghorne A, Kebbi-Beghdadi C, Goesmann A, Vaughan L, Greub G. 2010. The Waddlia Genome: A Window into Chlamydial Biology. PLoS ONE 5:e10890.

Bertelli C, Greub G. 2012. Lateral gene exchanges shape the genomes of amoeba-resisting microorganisms. Front. Cell. Infect. Microbiol. [Internet] 2. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3423634/

Black CM. 2013. Introduction. In: Black CM, editor. Issues in Infectious Diseases. Vol. 7. Basel: S. KARGER AG. p. 1–8. Available from: http://www.karger.com?doi=10.1159/000348748

Brinkman FSL, Blanchard JL, Cherkasov A, Av-Gay Y, Brunham RC, Fernandez RC, Finlay BB, Otto SP, Ouellette BFF, Keeling PJ, et al. 2002. Evidence That Plant-Like Genes in Chlamydia Species Reflect an Ancestral Relationship between Chlamydiaceae, Cyanobacteria, and the Chloroplast. Genome Res. 12:1159–1167.

Chang Y, Land M, Hauser L, Chertkov O, Del Rio TG, Nolan M, Copeland A, Tice H, Cheng J-F, Lucas S, et al. 2011. Non-contiguous finished genome sequence and contextual data of the filamentous soil bacterium Ktedonobacter racemifer type strain (SOSP1-21T). Stand. Genomic Sci. 5:97–111.

Clarke M, Lohan AJ, Liu B, Lagkouvardos I, Roy S, Zafar N, Bertelli C, Schilde C, Kianianmomeni A, Bürglin TR, et al. 2013. Genome of Acanthamoeba castellanii highlights extensive lateral gene transfer and early evolution of tyrosine kinase signaling. Genome Biol. 14:R11.

Collingro A, Tischler P, Weinmaier T, Penz T, Heinz E, Brunham RC, Read TD, Bavoil PM, Sachse K, Kahane S, et al. 2011. Unity in Variety—The Pan-Genome of the Chlamydiae. Mol. Biol. Evol. 28:3253–3270.

Domman D, Collingro A, Lagkouvardos I, Gehre L, Weinmaier T, Rattei T, Subtil A, Horn M. 2014. Massive Expansion of Ubiquitination-Related Gene Families within the Chlamydiae. Mol. Biol. Evol. 31:2890–2904.

Greub G, Kebbi-Beghdadi C, Bertelli C, Collyn F, Riederer BM, Yersin C, Croxatto A, Raoult D. 2009. High Throughput Sequencing and Proteomics to Identify Immunogenic Proteins of a New Pathogen: The Dirty Genome Approach. PLoS ONE 4:e8423.

Horn M. 2008. *Chlamydiae* as Symbionts in Eukaryotes. Annu. Rev. Microbiol. 62:113–131.

Horn M, Collingro A, Schmitz-Esser S, Beier CL, Purkhold U, Fartmann B, Brandt P, Nyakatura GJ, Droege M, Frishman D, et al. 2004. Illuminating the Evolutionary History of Chlamydiae. Science 304:728–730.

Huang J, Gogarten J. 2007. Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? Genome Biol. 8:R99.

Kalman S, Mitchell W, Marathe R, Lammel C, Fan J, Hyman RW, Olinger L, Grimwood J, Davis RW, Stephens R. 1999. Comparative genomes of Chlamydia pneumoniae and C. trachomatis. Nat. Genet. 21:385–389.

Kamneva OK, Knight SJ, Liberles DA, Ward NL. 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.

Lagkouvardos I, Weinmaier T, Lauro FM, Cavicchioli R, Rattei T, Horn M. 2013. Integrating metagenomic and amplicon databases to resolve the phylogenetic and ecological diversity of the Chlamydiae. ISME J. [Internet]. Available from: http://www.nature.com/ismej/journal/vaop/ncurrent/full/ismej2013142a.html

Mäurer AP, Mehlitz A, Mollenkopf HJ, Meyer TF. 2007. Gene Expression Profiles of Chlamydophila pneumoniae during the Developmental Cycle and Iron Depletion–Mediated Persistence. PLoS Pathog 3:e83.

McCutcheon JP, Dohlen CD von. 2011. An Interdependent Metabolic Patchwork in the Nested Symbiosis of Mealybugs. Curr. Biol. 21:1366–1372.

McCutcheon JP, Moran NA. 2012. Extreme genome reduction in symbiotic bacteria. Nat. Rev. Microbiol. 10:13–26.

Moustafa A, Reyes-Prieto A, Bhattacharya D. 2008. Chlamydiae Has Contributed at Least 55 Genes to Plantae with Predominantly Plastid Functions. PLoS ONE 3:e2205.

Nicholson TL, Olinger L, Chong K, Schoolnik G, Stephens RS. 2003. Global Stage-Specific Gene Regulation during the Developmental Cycle of Chlamydia trachomatis. J. Bacteriol. 185:3179–3189.

Parfrey LW, Lahr DJG, Knoll AH, Katz LA. 2011. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. Proc. Natl. Acad. Sci. U. S. A. 108:13624–13629.

Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L, Mitchell W, Olinger L, Tatusov RL, Zhao Q, et al. 1998. Genome Sequence of an Obligate Intracellular Pathogen of Humans: Chlamydia trachomatis. Science 282:754–759.

Taylor-Brown A, Vaughan L, Greub G, Timms P, Polkinghorne A. 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. Pathog. Dis. 73:1–15.

Welch RA, Burland V, Plunkett G, Redford P, Roesch P, Rasko D, Buckles EL, Liou S-R, Boutin A, Hackett J, et al. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc. Natl. Acad. Sci. 99:17020–17024.

Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattacharya D. 2004. A Molecular Timeline for the Origin of Photosynthetic Eukaryotes. Mol. Biol. Evol. 21:809–818.

Chapter II

Overview of publications/manuscripts

Chapter III

Manuscript title:

Massive Expansion of Ubiquitination-Related Gene Families within the *Chlamydiae*

Author names:

Daryl Domman, Astrid Collingro, Ilias Lagkouvardos, Lena Hehre, Thomas Wienmaier, Thomas Rattei, Agathe Subtil, Matthias Horn.

Reference:

Molecular Biology and Evolution 31, no. 11 (November 1, 2014): 2890–2904. doi:10.1093/molbev/msu227.

Author contributions:

DD and MH designed all research experiments. DD, AC and IL performed all analysis except the Type III secretion assays. Type III secretion assays were performed by LG and AS. TW and TR provided bioinformatics support. DD and MH wrote the manuscript.

Chapter IV

Manuscript title:

Following the footsteps of chlamydial gene regulation

Author names:

Daryl Domman, Matthias Horn

Reference:

This manuscript was submitted to *Molecular Biology and Evolution* on 11.6.2015 and is currently under review.

Author contributions:

The project was conceived and designed by D.D. and M. H. All analyses performed by D.D. D.D. wrote the manuscript with significant input and editing by M.H.

Chapter V

Manuscript title:

Plastid establishment did not require a chlamydial partner

Author names:

Daryl Domman, Matthias Horn, T. Martin Embley, Tom A Williams.

Reference:

Nature Communications 6 (March 11, 2015). http://dx.doi.org/10.1038/ncomms7421.

Author contributions:

The project was conceived, designed and all analyses performed by D.D. and T.A.W. The manuscript was written by T.A.W. and D.D., with significant input and editing by M.H. and T.M.E.

Chapter III

Massive Expansion of Ubiquitination-Related Gene Families within the *Chlamydiae*

Massive Expansion of Ubiquitination-Related Gene Families within the *Chlamydiae*

Daryl Domman,[1] Astrid Collingro,[1] Ilias Lagkouvardos,[1] Lena Gehre,[2] Thomas Weinmaier,[1] Thomas Rattei,[1] Agathe Subtil,[2] and Matthias Horn*[,1]

[1]Department of Microbiology and Ecosystem Science, University of Vienna, Vienna, Austria

[2]Unité de Biologie des Interactions Cellulaires, Institut Pasteur, Paris, France

***Corresponding author:** E-mail: horn@microbial-ecology.net.

Associate editor: Howard Ochman

Abstract

Gene loss, gain, and transfer play an important role in shaping the genomes of all organisms; however, the interplay of these processes in isolated populations, such as in obligate intracellular bacteria, is less understood. Despite a general trend towards genome reduction in these microbes, our phylogenomic analysis of the phylum *Chlamydiae* revealed that within the family Parachlamydiaceae, gene family expansions have had pronounced effects on gene content. We discovered that the largest gene families within the phylum are the result of rapid gene birth-and-death evolution. These large gene families are comprised of members harboring eukaryotic-like ubiquitination-related domains, such as F-box and BTB-box domains, marking the largest reservoir of these proteins found among bacteria. A heterologous type III secretion system assay suggests that these proteins function as effectors manipulating the host cell. The large disparity in copy number of members in these families between closely related organisms suggests that nonadaptive processes might contribute to the evolution of these gene families. Gene birth-and-death evolution in concert with genomic drift might represent a previously undescribed mechanism by which isolated bacterial populations diversify.

Key words: **gene families, birth and death model, intracellular bacteria, effector proteins, F-box.**

Introduction

The genomes of organisms reveal complex histories of gene transfer, loss, gain, and rearrangement. The extent that these processes play in shaping gene families of both prokaryotes and eukaryotes are markedly different. Gene gain within eukaryotes is largely driven by intragenomic duplication events (Lynch and Conery 2000; Koonin et al. 2002; Yang et al. 2003; Kaessmann 2010), and although duplication certainly shapes bacterial genomes, most gains are the result of horizontal gene transfer (HGT) events (Ochman et al. 2000; Lerat et al. 2005; Treangen and Rocha 2011). Estimates of the contribution gene duplication processes play across domains of life vary from 65% to 30% in the genomes of *Arabidopsis* and *Escherichia coli*, respectively (Zhang 2003). Although genetic innovation typically arises through gene acquisition from foreign sources, gene duplication events are increasingly being recognized as an important driver of bacterial genome evolution (Goldman et al. 2006; McLeod et al. 2006; Cho et al. 2007).

Comparisons of closely related organisms have revealed a highly dynamic landscape of gene families, in which the copy number between species can vary substantially (Pushker et al. 2004; Lerat et al. 2005). Given this background, an intriguing evolutionary backdrop to study gene family evolution is within obligate, intracellular bacteria. In these populations, the fixation of mutations is strongly affected by genetic drift, with a propensity in these genomes for deletion (Kuo and Ochman 2009a), and thus gene family expansions within these genomes are generally rare (Hooper and Berg 2003; Gevers et al. 2004). Insightful analysis on gene family evolution is best approached when comparing multiple genomes from closely related species, facilitating identification of paralogs (homologous genes resulting from duplication), orthologs (homologous genes resulting from speciation), or xenologs (homologous genes derived from HGT). In this regard, the phylum *Chlamydiae* offers an ensemble of fully sequenced genomes across multiple families.

All members of the phylum *Chlamydiae* are obligate, intracellular bacteria and represent one of the most ancient and successful lineages associated with eukaryotes (Horn 2008; Subtil et al. 2014). These organisms all share a characteristic biphasic developmental cycle consisting of an infectious, extracellular state and an intracellular replicative state. The phylum can be divided into two major phylogenetic groupings: The family Chlamydiaceae, which encompass well known animal and human pathogens such as *Chlamydia trachomatis* and *C. pneumoniae*, and a group of families comprising the environmentally distributed chlamydiae such as Simkaniaceae, Waddliaceae, and Parachlamydiaceae collectively referred to as environmental chlamydiae. Recently, it was shown that the diversity of the phylum is tremendously greater with perhaps over 200 families spanning nearly every environment (Lagkouvardos et al. 2013). All members of the *Chlamydiae* show notable genomic reductions and truncated

Article

metabolic pathways including the inability to synthesize many amino acids and nucleotides (Stephens et al. 1998; Kalman et al. 1999; Horn et al. 2004; Bertelli et al. 2010; Collingro et al. 2011; Myers et al. 2012).

In this study, we set out to determine how gene families have evolved in members of the phylum *Chlamydiae*. We present four new genome sequences for members of the family Parachlamydiaceae, which include two genome sequences for the genus *Neochlamydia*. We show that organisms within the Parachlamydiaceae have unprecedented numbers of proteins harboring domains typically found in eukaryotes, the majority of which are related to eukaryotic ubiquitination pathways. We show that these genes have undergone rapid expansions and form the largest gene families within the phylum. We demonstrate that many of these large gene families are evolving under a gene-birth-death model (Nei and Rooney 2005) and that differences between closely related organisms may be explained by genomic drift.

Results

Genome Sequencing of Novel Members of the Chlamydiae

Currently, there are nine described families within the *Chlamydiae*; however, the majority of available genome sequences come from a single family, the pathogenic Chlamydiaceae. To deepen our insights into a family outside of the Chlamydiaceae, we sequenced the genomes of four members of the family Parachlamydiaceae, which include two members of the genus *Neochlamydia*, and two additional genomes of *Protochlamydia* and *Parachlamydia*. All of the newly sequenced Parachlamydiaceae members were isolated from free-living amoeba. With the exception of *Neochlamydia* sp. EPS4, the isolates have been described previously (Fritsche et al. 2000; Heinz et al. 2007; Schmitz-Esser et al. 2008). The draft genomes represent nearly complete genome sequences based on paired end read data (90–96%) and the presence of conserved single-copy marker genes (98–100%; supplementary table S1, Supplementary Material online). Using these additional genome sequences, we first aimed to construct a phylogenetic framework of the phylum *Chlamydiae* using concatenated alignments of 32 marker proteins (supplementary table S2, Supplementary Material online). Phylogenetic trees obtained with different methods confirmed the monophyly of the Chlamydiaceae and the Parachlamydiaceae with strong support (fig. 1). The Chlamydiaceae can be subdivided in two previously recognized groups, and within the Parachlamydiaceae, the genera *Protochlamydia*, *Neochlamydia*, and *Parachlamydia* were recovered with high confidence.

All members of the Chlamydiaceae show highly similar genomes in terms of gene content and synteny (Myers et al. 2012); however, between chlamydial families, rearrangements have played a major role in genome evolution (Collingro et al. 2011). Whole-genome alignments of members of the Parachlamydiaceae clearly illustrate that within the genera *Protochlamydia*, *Neochlamydia*, and *Parachlamydia*, there are few rearrangements, and the genomes are highly syntenic (fig. 1). Between these genera, however, there have been extensive genome rearrangements demonstrating the surprising dynamic nature of these reduced genomes.

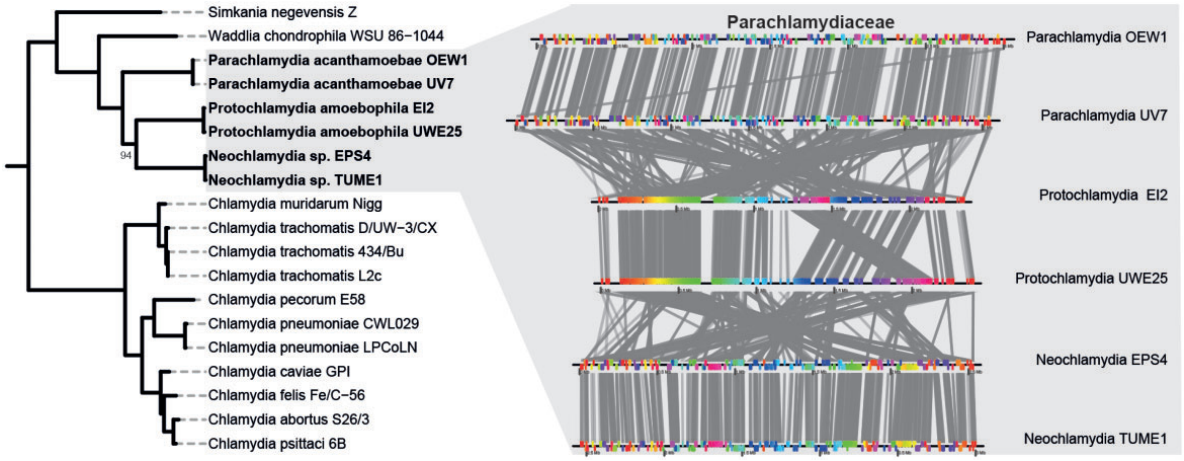


Fig. 1. Phylogeny of the *Chlamydiae* and rearrangement history of genomes within the Parachlamydiaceae. Phylogeny of the *Chlamydiae* based on 32 phylogenetic marker proteins. A Bayesian analysis using MrBayes (Ronquist and Huelsenbeck 2003) was performed on a set of 24 ribosomal proteins in addition to GyrB, RecA, RpoB, RpoC, and EF-Tu from 19 sequenced members of the phylum (supplementary table S2, Supplementary Material online). Members of the Planctomycetes (*Blastopirellula marina* DSM 3645, *Candidatus* Kuenenia stuttgartiensis, and *Gemmata obscuriglobus* UQM 2246) and the Verrucomicrobia (*Akkermansia muciniphila* MucT, *Lentisphaera araneosa* HTCC2155, *Opitutus terrae* PB90-1, and *Verrucomicrobium spinosum* DSM 4136) were used as outgroups (not shown). Colors denote family level classification. Posterior probability scores are indicated only if below 100%. To the right, conserved synteny and rearrangement history of genomes within the Parachlamydiaceae are shown. The genomes of six members of the family were aligned using MAUVE to elucidate synteny between genomes and visualized using genoPlotR. Extensive rearrangements are apparent between members of different genera, whereas within genus, comparisons show little rearrangements, with a notable exception in the *Protochlamydia* where a large block has been rearranged.

The Gene Family Landscape of the Chlamydiae

To explore gene family evolution among members of the *Chlamydiae*, we first identified gene families using clusters of orthologous groups of proteins within the predicted proteomes from 19 chlamydial genomes (supplementary table S2, Supplementary Material online). We then searched for gene families that contain expansion events, that is, those having multiple members from one organism. Previous work has demonstrated that genome size correlates with the number of paralogs, with larger genomes containing more paralogs than smaller ones (Bratlie et al. 2010). Taking into account all gene family members, regardless of whether they originate from duplication processes or transfer events, this trend is generally observed within chlamydial genomes (fig. 2*a*).

As described previously, gene family expansions are sparse within the genomes of the Chlamydiaceae (Kalman et al. 1999; Kamneva et al. 2012), with *C. pneumoniae* CWL029 harboring the largest number ($n = 24$). In line with previous observations, the largest gene families identified in our study encode the polymorphic membrane proteins (PMPs) (Grimwood and Stephens 1999; Gomes et al. 2006) including nine members from *C. pneumoniae* LPCoLN and two from the *C. trachomatis* serovars. The observed split of PMPs among several smaller gene families in our analysis is an indication that our approach is rather conservative in assigning a protein to a gene family.

The total number of expansion events ($n = 277$) detected in the genome of *Simkania negevensis* represents a 10-fold increase when compared with the Chlamydiaceae. As these group into many small gene families, the extended number of gene copies in *S. negevensis* is the result of many small-scale duplication or transfer events (fig. 2*b*). This situation is similar in *Waddlia chondrophila*. In stark contrast, roughly half of the total of genes resulting from expansion events in *Neochlamydia* and *Protochlamydia* are the contribution of only few gene families.

Large Gene Family Expansions in the Parachlamydiaceae

The detection of large gene families in *Neochlamydia* and *Protochlamydia* indicates that there have been several large-scale expansion events within the Parachlamydiaceae. Notably, different gene families are expanded in *Neochlamydia* and *Protochlamydia* (fig. 2*b*). These represent the four largest gene families (containing between 27 and 138 members) found within the phylum and include two gene families specific to *Neochlamydia* and two restricted to *Protochlamydia*.

Intrigued by these four large-scale lineage-specific expansion events between the species pairs of *Protochlamydia* and *Neochlamydia*, we sought to better characterize these gene families, as most of their members are yet unknown with respect to their functional role (i.e., they are classified as hypothetical proteins). Remarkably, despite being in different gene families from different organisms, there are several similarities between these proteins. Firstly, they all encompass protein–protein interaction domains such as leucine-rich repeats (LRRs) or tetratricopeptide repeats (TPRs). Secondly, many of them contain additional domains typically found in eukaryotes, such as F-boxes, BTB-boxes, and RING/U-boxes, which are associated with eukaryotic ubiquitination pathways (Angot et al. 2007).

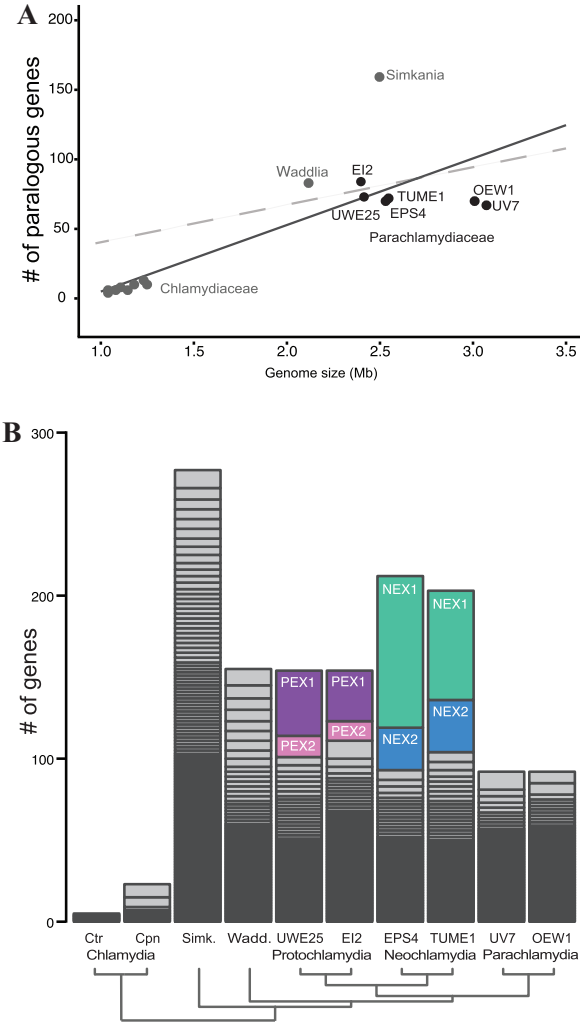


Fig. 2. The paralogous gene landscape of the *Chlamydiae*. (*A*) Number of paralogous genes within the *Chlamydiae*. The number of paralogous genes, not including multiple copies, is plotted against genome size along with a linear regression line ($y = 47.87x - 42.9$; $R^2 = 0.70$; black line). The dashed gray line is plotted as a reference from 200 prokaryotic genomes (Bratlie et al. 2010). (*B*) Distribution of chlamydial gene families per genome with two or more members. The number of genes within each family is plotted for representative genomes. The genomes of the Chlamydiaceae have relatively small gene family sizes. The polymorphic outer membrane proteins comprise the largest gene families in the Chlamydiaceae and can be seen as the two largest blocks in the *Chlamydia pneumoniae* LPCoLN (Cpn) bar. The size distribution of gene families is ordered from smallest to greatest, and the appearance of a solid "black box" at the base is merely an effect of the spacing of many small gene families. There are several extensive gene families (labeled) within members of *Neochlamydia* (NEX1, NEX2) and *Protochlamydia* (PEX1, PEX2).

Eukaryotic Ubiquitination-Associated Domains Predominate Large Gene Families

The largest gene family, termed *Neochlamydia expansion* 1 (NEX1), in the phylum comprised a total of 138 members, which are contributed by the two *Neochlamydia* genomes. The domain architecture within this large gene family is heterogeneous; however, all proteins contain various C-terminal repetitions of LRR domains (fig. 3). We have identified two subfamilies that we delineate NEX1a and NEX1b within the NEX1 family. The majority of members fall into the NEX1a subfamily, in which they have a highly conserved N-terminal F-box or F-box-like domain. The members have an average of 77% sequence similarity among each other, and the F-box-like domain is 57% and 45% similar to *Acanthamoeba castellanii* and human F-box-like domains, respectively. The smaller NEX1b family, in contrast, has a conserved RING/U-box at the N-terminus. Both the F-box and RING/U-box domains are associated with eukaryotic E3 ubiquitin ligase complexes (Willems et al. 2004). Between all members in the NEX1 family, there is a region of roughly 50 amino acids between the predicted N-terminal domains and the LRRs that is highly conserved. However, we failed to detect any known domains in this region nor was there any homology to known proteins.

The second large gene family within the *Neochlamydia* represents the third largest family of the phylum. This large gene family, termed NEX2, comprised 50 proteins (fig. 3). Similar to NEX1, the prevailing domain architecture is that of eukaryotic-like E3 ubiquitin ligase-associated domains paired with repeat domains. This family is defined by the presence of multiple TPR domains at the C-terminus, and a general conservation of an F-box domain at the N-terminus in the majority of members. In most members, there is also a conserved DUF294 domain located mid protein, which is a putative nucleotidyltransferase. In ten members, there is an ovarian tumor (OTU) (Balakirev et al. 2003) domain directly following the F-box domain followed by the DUF294 domain.

The other large gene families occur primarily in the members of the *Protochlamydia* and represent the second and fourth largest gene families of the phylum. The largest gene family in the *Protochlamydia* (PEX1) comprised 73 members in total (fig. 3). Intriguingly, another E3 ubiquitin ligase-related domain, the BTB domain, is present in all but three members, at the N-terminus. The BTB domain is then coupled to C-terminal LRR domains in all members. The PEX2 family comprised a total of 27 proteins that, despite no detectable domain at the N-terminus, share multiple TPR domains in the middle of the protein followed by a CHAT domain (Sakakibara and Hattori 2000) at the C-terminus.

In summary, the four largest gene families represent a surprisingly diverse armada of proteins, which most likely function within eukaryotic host cells where they potentially interfere with the ubiquitination pathway. The heterogeneity

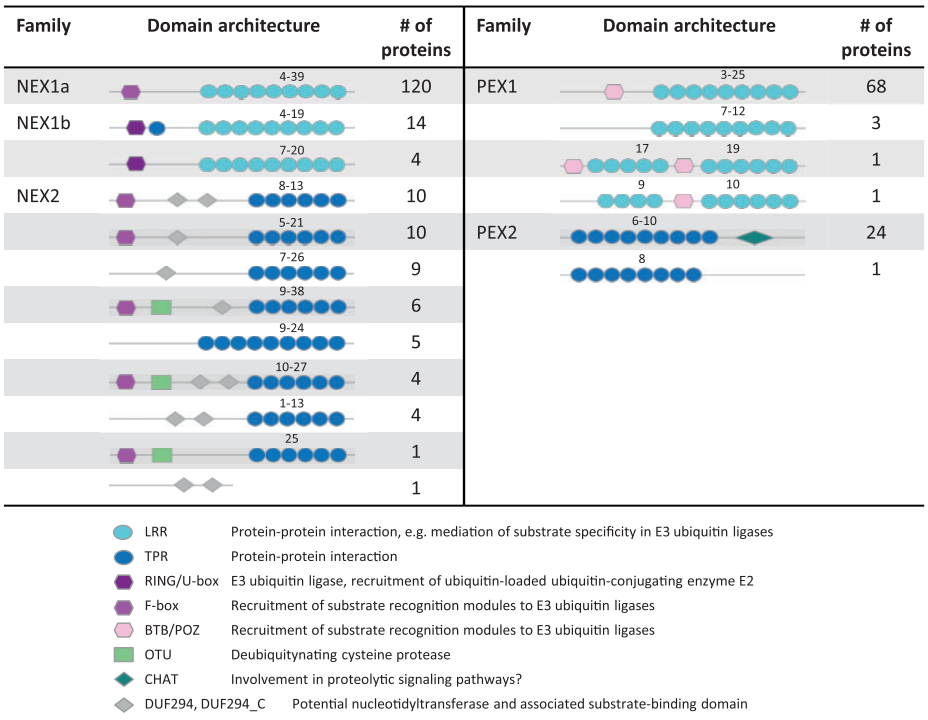


Fig. 3. Protein domain architecture of largest gene families. The domain architecture of *Neochlamydia* (NEX1, NEX2) and *Protochlamydia* (PEX1, PEX2) gene families are shown. The range of the number of domain repeats and functional assignments of the detected domains are indicated. NEX1 can be divided into two subfamilies based on phylogeny and domain presence/absence. A role of these proteins in the context of eukaryotic cells can be postulated based on the presence of domains otherwise found in eukaryotes.

in domain architecture among these proteins interestingly mirrors that of their eukaryotic counterparts (Perez-Torrado et al. 2006; Xu 2006).

A Pool of Putative Effector Proteins

If the members of the largest gene families in the phylum *Chlamydiae* serve as effector proteins for host manipulation, they would need to be secreted and transported to the host cell cytosol. This is typically achieved through a type III secretion system, a well-conserved virulence mechanism among the *Chlamydiae*, which has been shown to translocate several characterized effectors (Peters et al. 2007; Betts et al. 2009). Indeed, many of the proteins found within the expanded *Neochlamydia* and *Protochlamydia* gene families are predicted by computational analysis to be secreted by the type III secretion system and to be extracellular, host associated. Within the NEX1 family, 66 of 138 (49%) members are predicted to be secreted. A total of 37 (51%) and 10 (37%) were predicted to be secreted from within the PEX1 and PEX2 gene families, respectively. The NEX2 gene family had the fewest predicted with only two members.

As the identification of the signal for secretion via the type III secretion system is inherently difficult (Arnold et al. 2009), we tested representatives of each of the four largest gene families in vitro using a heterologous type III secretion substrate assay with *Shigella flexneri* as a host for protein expression. This assay has been used to successfully characterize type III secretion effector proteins from the chlamydiae before (Subtil et al. 2001), and because of the lack of routine genetic tools, the *Sh. flexneri* system is an attractive surrogate method for analyzing type III secretion in chlamydiae in vivo. This experiment demonstrated that the tested members of NEX1, NEX2, PEX1, and PEX2 contain a functional type III secretion recognition signal (supplementary fig. S1, Supplementary Material online). Together with the presence of eukaryotic-like domains and computational predictions, this strongly indicates that these gene families are large pools of effector proteins.

Molecular Evolution of Large Gene Families

To better understand how these large gene families may have evolved, we reconstructed their phylogenetic relationships. Gene family trees were calculated using conserved sites among the protein alignment (supplementary figs. S2–S6, Supplementary Material online). The average amino acid identity between members ranges from 45% to 64%, with the most closely related sequences belonging to PEX2. Tree topologies suggest that the members of all four gene families have rapidly diverged as indicated by their long branch lengths. Although the number of LRR and TPR domains varies dramatically between 1 and 39, this had no apparent effect on the phylogenetic placement.

We find clear cases in which the orthologs of two species group together, indicating expansions have occurred before speciation (supplementary figs. S2–S6, Supplementary Material online). Alternatively, expansions post speciation is apparent in all gene families. Reconciliation of gene family trees with the species tree indicates that, in addition to expansions, many gene losses have occurred for each gene family (supplementary table S3, Supplementary Material online). For instance, for the NEX1 family, there have been 78 expansion events, whereas 33 losses have occurred, attributed to 13 and 20 losses in *Neochlamydia* spp. EPS4 and TUME1, respectively. There were nearly equal losses between EPS4 and TUME1 (13 and 10) in NEX2 and a total of 46 expansions. Similarly, PEX1 consists of 52 expansions, and 15 and 9 losses in *Protochlamydia amoebophila* EI2 and UWE25, respectively.

A Birth-and-Death Model of Evolution

A pattern of differential gain, loss, and maintenance of gene family members is strongly indicative of these gene families evolving according to a birth-and-death model (Nei and Rooney 2005). Because of this differential maintenance of gene family members, the hallmarks of the birth-and-death model are interspecies clustering of members in the phylogenetic trees and the presence of pseudogenes from degraded members (Nei 2007). As we observe interspecies clustering for the PEX and NEX gene families (supplementary figs. S2–S6, Supplementary Material online), we also tested for pseudogenization events in the intergenic regions of the *P. amoebophila* UWE25 genome (the draft *Neochlamydia* genomes are less suitable for this analysis). By utilizing BLAST, we searched for matches to the predicted proteome using all intergenic regions as a query. We then mapped the best BLAST hits, representing 116 pseudogenes, to their respective gene families to get a picture of a given families' representation in the intergenic regions. The most represented gene family in the intergenic regions, surprisingly, was PEX1 (16 pseudogenes). For PEX2, one pseudogenized fragment was detected. The observed presence of interspecies clustering and pseudogenes, and the dynamic history of gains and losses within the gene families are indicative of a birth-and-death model of evolution.

In contrast, if these families were evolving via concerted evolution, the phylogenetic trees would depict intraspecies clustering, that is, that members of a gene family will be more homologous to the other members from the same organism than to that of other species. Intraspecies clustering occurs due to repeated recombination among gene family members within a genome, leading to an overall high sequence similarity of all members, a process known as gene conversion (Santoyo and Romero 2005). We thus tested for the possibility of recombination within the gene families using the methods implemented in the RDP4 software suite (Martin et al. 2010). Care must be taken, however, when assessing the impact of recombination among divergent proteins, as the recombination signal is quite error prone when proteins share less than 70% similarity, and these methods are heavily dependent on the alignment (Martin et al. 2010). We detect some recombination events between members within the PEX and NEX gene families; however, the majority of the predictions are only marginally significant (the consensus scores are below the confidence threshold of 0.6). In the

NEX1a family, the portion of the sequences most identified as recombinant is the F-box domain (supplementary fig. S8, Supplementary Material online). This should be judiciously interpreted, as this might represent sequence similarity due to purifying selection operating on this domain. Overall, we did not find convincing evidence that recombination has played a major role in shaping the evolution of the PEX and NEX gene families.

Gene Duplications and Purifying Selection

Gene family expansions can be the result of either gene duplication or HGT. The disentangling of these events is not trivial and, in fact, may be impossible in the case of the gene families investigated here due to lineage-specific evolution and the absence of clear homologs in other bacterial taxa (Kuo and Ochman 2009b). However, several lines of evidence suggest that, regardless of the initial origin of these genes, gene duplication processes have played a clear role in the evolution the large Parachlamydiaceae gene families. First, a hallmark of gene duplication is the presence of tandem arrays of gene copies. In this regard, we find several large tandem arrays with members of the large gene families in *Protochlamydia* and *Neochlamydia*, including examples for recent duplications (fig. 4, supplementary fig. S7, Supplementary Material online). In the *P. amoebophila* UWE25 genome, we detected 47 tandem duplication events (distance within 10 genes) represented by only 13 gene families. Nearly half of these are the contribution of PEX1 ($n = 20$); however, they appear in several clusters spread throughout the genome. PEX2 demonstrates the most dramatic case of a tandem array consisting of 13 members. After the PEX gene families, the third largest tandem array only comprised three members, intriguingly also F-box domain containing proteins. NEX1 has 50 instances (36%) where members are found within five genes of each other. Second, the majority of the large gene family members meet the generally used criterion for identification of paralogs, that is, they show at least 30% amino acid sequence identity over at least 60% of the protein length. Thirdly, phylogenetic trees show several species-specific expansion events (including those genes still arranged in tandem arrays), which are best explained by gene duplication.

Effector proteins have been shown to be among the fastest evolving proteins in a number of pathogen genomes (Nogueira et al. 2012) and are often shown to be under positive selection. To assess the mode of selection acting on members of the expanded gene families, we calculated the ratio of nonsynonymous (d*N*) versus synonymous (d*S*) substitution rates for each of the four gene families. A d*N*/d*S* ratio equal to 1 indicates a neutral state of selection, whereas values higher or lower than 1 indicate positive or purifying selection, respectively. Global estimates of d*N*/d*S* under the Model M0 from CodeML (Yang 2007) ranged from 0.29 (NEX2) to 0.63 (NEX1a). Additionally, using pairwise sequence comparisons for d*N*/d*S* calculation, we did not find support that these gene families are currently evolving under positive selection. However, the probability of these gene families evolving under purifying selection was highly significant ($P < 0.01$). Therefore, purifying selection is currently the dominant force driving the evolution of these gene families and thus facilitates their maintenance.

Massive Expansion of Ubiquitination-Associated Proteins

Driven by the discovery that members of the four largest Parachlamydiaceae gene families showed rapid divergence and are kept in chlamydial genomes despite apparent functional redundancy, we asked whether there are additional genes not included in these gene families but encoding similar functional domains. The common theme of the large Parachlamydiaceae gene families is the presence of domains that serve in the recruitment of target proteins to the eukaryotic ubiquitination machinery. We therefore extracted all proteins containing F-box/F-box-like, BTB/POZ, and RING/U-box domains by scanning all chlamydial proteomes with each respective HMM profile. We found no RING/U-box containing proteins apart from those identified earlier as members of NEX1b, and we detected only few additional proteins in *Neochlamydia* and *Parachlamydia* harboring a BTB/POZ domain similar to those of PEX1. However, our search

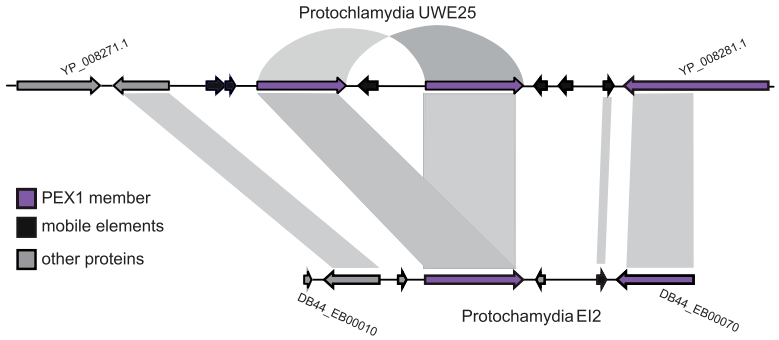


FIG. 4. Example of duplication of BTB-box proteins in the *Protochlamydia*. A new duplicate has arisen in *Protochlamydia amoebophila* UWE25 (shown via the arrow) after the split from *P. amoebophila* EI2. BTB-box proteins are indicated in purple. The phylogenetic placement of these proteins (supplementary fig. S5, Supplementary Material online) supports this scenario. Orthologous proteins between the *Protochlamydia* are indicated by connecting blocks.

unveiled an astonishing number of proteins harboring F-box/F-box-like domains, with over 370 proteins within the phylum. Nearly 300 of the F-box proteins are the contribution of the two *Neochlamydia* species (129 in TUME1 and 158 in EPS4). To characterize the relationships among this F-box superfamily, we constructed a phylogenetic tree based on a domain alignment. This analysis shows that many of the additional F-box proteins found in *Neochlamydia* cluster with either NEX1a or NEX2 (fig. 5*a*). We also see several lineage-specific expansions of F-box proteins within *Protochlamydia* and *Parachlamydia* species. Reconciliation of the F-box superfamily tree with the chlamydial species tree confirms an extremely dynamic history of large-scale gene birth and death events (fig. 5*b*), mirroring the evolutionary pattern seen for the large Parachlamydiaceae gene families (supplementary figs. S2–S6 and table S3, Supplementary Material online).

Furthermore, we searched for additional chlamydial proteins containing domains for protein–protein interaction, such as LRR, TPR, or ankyrin repeats, which are often associated with F-box/F-box-like, BTB/POZ, and RING/U-box domains in the large gene families. This search identified a vast number of proteins for each domain. For instance, chlamydial genomes encode 409 proteins with LRR domains, nearly doubling the amount of LRR proteins contributed by the large Parachlamydiaceae gene families. Taken together, in addition to the four large gene families, there is an even greater pool of chlamydial proteins with a putative role in host interaction.

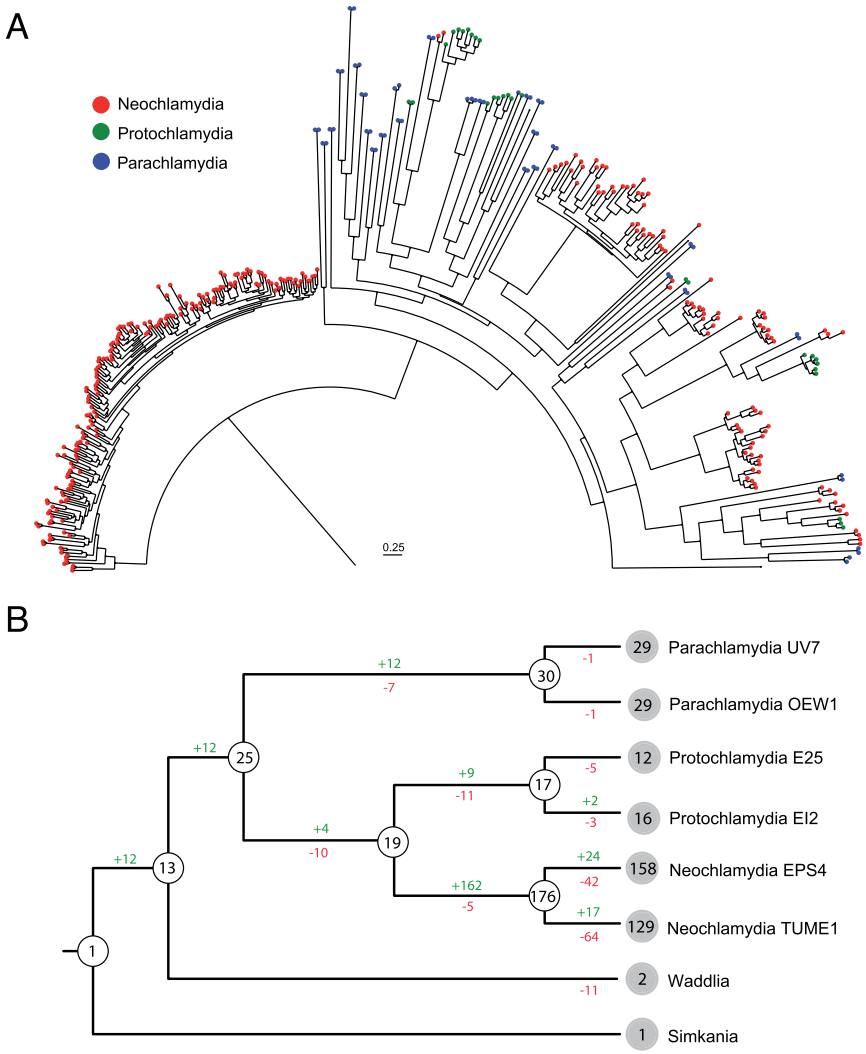


Fig. 5. The phylogeny and evolutionary history of the F-box superfamily within the *Chlamydiae*. (*A*) The phylogeny of 376 proteins within the *Chlamydiae* that harbor an F-box/F-box-like domain. This domain was extracted from each protein and aligned using MAFFT. Maximum-likelihood reconstruction of the phylogeny of the superfamily was performed with FastTree2. (*B*) The F-box domain superfamily gene tree was reconciled with the chlamydial species tree to reconstruct the evolutionary history of this group for members of the *Chlamydiae*. The nodes in blue indicate the predicted number of F-box proteins, and numbers on the branches depict the gains and losses. The extant species are indicated with their respective counts for F-box proteins. The *Neochlamydia* have undergone massive gains and losses after the divergence from *Protochlamydia*.

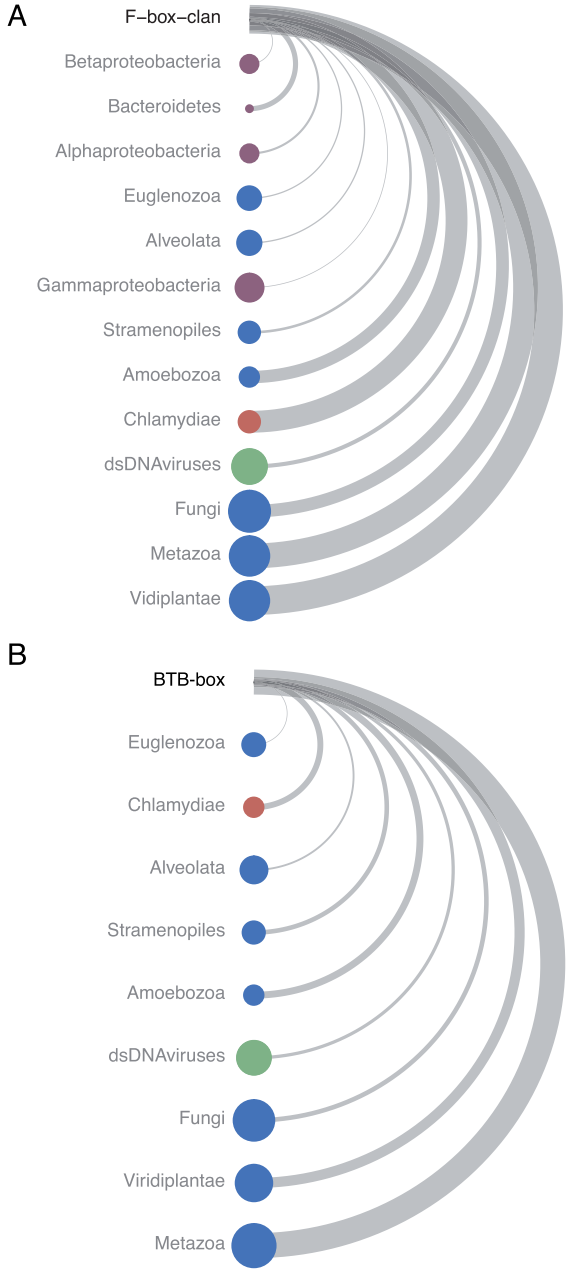


Fig. 6. Taxonomic profile of F-box and BTB domains. The distribution of (*A*) the F-box clan, and (*B*) the BTB-box throughout sequenced organisms. The size of the node indicates the number of species harboring proteins with the domain. Thus, larger node size indicates a larger number of species in which a domain is found within a taxon. Nodes are ordered from least to greatest by the total number of proteins that contain the domain within the taxon. This is different than the number of species as one species can have many proteins harboring a given domain. To reflect this disparity and to facilitate comparisons, we computed a normalized value for each taxon that represents the number of total proteins divided by the number of species. This normalization value is represented by the width of the arc in the diagram. For instance, the chlamydiae are represented by few species (small node size) but are among the taxa containing the largest numbers of proteins with F-Box and BTB domains (position on vertical axis) and show a high number of proteins with these domains per species (arc width). All bacterial taxa are plotted in purple and selected major eukaryotic taxa in blue. The *Chlamydiae* are labeled in red, and double-stranded DNA viruses are shown in green. The data were obtained from the Pfam database for each domain, and counts were updated to reflect findings in this study.

To gain a broader overview of the occurrence of F-box/F-box-like and BTB/POZ domains among other prokaryotes and eukaryotes, we extracted domain abundance data from Pfam (Finn et al. 2013) and included the current counts for the *Chlamydiae* genomes present in this study (fig. 6). This revealed a striking pattern that the few bacterial groups encoding proteins with F-box and BTB domains are almost exclusively amoeba-associated organisms. These include members of the Legionellales (Gammaproteobacteria), the Rickettsiales (Alphaproteobacteria), and the amoeba symbiont *Amoebophilus asiaticus* (Bacteroidetes). When the number of F-box proteins is normalized against the total number of species in a given taxon, the *Chlamydiae* lead in the number of F-box proteins found in bacteria and even harbor more than several lineages of eukaryotes including the Amoebozoa. For the BTB proteins, the *Chlamydiae* appear to be the only bacterial lineage that harbors this domain. It is intriguing that many of the large double-stranded DNA viruses, namely the amoeba-infecting giant viruses, contain many proteins with an F-box or BTB domain.

Discussion

Large Gene Families Are Rare within Reduced Bacterial Genomes

Chlamydial genomes are among the smallest known for prokaryotes due to genome degradation consistent with long-term, obligate associations with eukaryotic organisms (McCutcheon and Moran 2011). The genomes of host-associated bacteria such as *Coxiella*, *Mycoplasma*, *Rickettsia* species, and members of the Chlamydiaceae, tend to have small or single copy, gene families (Gevers et al. 2004), and gene family expansions, either by gene duplication or HGT appear to have less effect on shaping the genomes of these obligate host-associated bacteria (Bordenstein and Reznikoff 2005). However, there is evidence that HGT may be more prevalent than once thought in these organisms (Blanc et al. 2007). A notable exception to this paradigm is the genome of the obligate intracellular pathogen *Orientia tsutsugamushi* (Rickettsiales), in which there has been massive expansion of type IV secretion system (over 350 *tra*-related genes) and host–microbe interaction genes (Cho et al. 2007). However, these expansions are thought to be the result of copious plasmid integration, and the genome is also littered with mobile elements, a scenario not shared within chlamydial genomes. Here, we have shown that several members of the *Chlamydiae* harbor gene families that have expanded at immense magnitudes, especially when compared with other intracellular bacteria (fig. 2). We find strong support for duplication processes contributing to the expansion of these families, thus highlighting that innovation through gene

duplication has had pronounced effects in shaping these chlamydial genomes.

Chlamydial Proteins Putatively Involved in Interference with Eukaryotic Ubiquitination Pathways

We demonstrated that the largest chlamydial gene families harbor proteins with domains associated with eukaryotic ubiquitination pathways and that the chlamydial F-box superfamily, in particular, is tremendous in size (figs. 3 and 5). A recent survey of prokaryotes found a total of 74 F-box proteins distributed in 22 species (Price and Kwaik 2010), which means the number of these proteins present in a single *Neochlamydia* genome is twice that of all previously known bacterial F-box proteins combined. The 76-member BTB superfamily is also remarkable in that the Parachlamydiaceae are the only prokaryotes known to harbor this domain (fig. 6).

In eukaryotes, ubiquitin plays a pivotal regulatory role as a posttranslational modification that includes targeting proteins for degradation. The process of adding ubiquitin to a protein occurs when an assembled E3 ubiquitin ligase complex carrying a target protein is bound to the ubiquitin conjugating enzyme E2. The ubiquitin ligase is a multiprotein complex termed the Skp1-Cullin-F-box protein complex, by which F-box proteins recruit target proteins and subsequently bind to Skp1, which is linked to a Cullin protein (Zheng et al. 2002). A RING/U-box protein then serves as a linker of E2 to the newly formed ubiquitin ligase, and the transfer of the ubiquitin moiety to the target occurs. BTB-box proteins can have multiple functions, but chief among them is a functional equivalent to the F-box protein in recruiting targets by binding to the E3 ubiquitin ligase complex (Perez-Torrado et al. 2006). Protein–protein interaction domains, such as ankyrin, kelch, WD-40, LRR, or TPR repeat domains, are coupled to F-box and BTB-box domains and confer the specificity for target proteins.

Given the conservation and essentiality of the ubiquitination pathway throughout eukaryotes, it should come as no surprise that bacterial pathogens have engineered ways to manipulate this pathway. The use of F-box proteins appears to be a common feature among plant pathogens, such as *Agrobacterium tumefaciens* and *Ralstonia solanacearum*, whose genomes encode one and four F-box proteins, respectively (Magori and Citovsky 2011). Intriguingly, F-box proteins seem to be a common feature of amoeba-associated bacteria and viruses. The amoeba symbiont *A. asiaticus*, a Bacteroidetes, is predicted to harbor 15 F-box proteins and until now was the largest known pool of these proteins among sequenced genomes (Schmitz-Esser et al. 2010). Additionally, *Legionella pneumophila* exports an F-box protein coupled to ankyrin repeats, termed AnkB, that is essential for infection of both human cell lines and *Acanthamoeba* (Price et al. 2009; Lomma et al. 2010). AnkB blocks host proteosomal degradation and thus generates increased levels of required amino acids (Price et al. 2009; Price et al. 2011). Several other F-box proteins secreted by *L. pneumophila* have been shown to interact with host E3 ligase complexes (Ensminger and Isberg 2010). Among the members of the Chlamydiaceae, we could not detect an F-box or BTB-box domain. However, there are several proteins within the Chlamydiaceae that function as deubiquinating proteases, such as the *C. trachomatis* ChlaDub1 (Misaghi et al. 2006) and the recently described ChlaOTU characterized in *Chlamydia caviae* (Furtado et al. 2013).

We have shown experimentally that representative members of the investigated chlamydial gene families contain functional type III secretion signals (supplementary fig. S1, Supplementary Material online). Thus, they likely represent an extensive pool of effector proteins with a putative role in hijacking the host ubiquitination machinery, perhaps in a manner similar to AnkB from *Legionella*, by increasing nutrient availability. In the absence of direct protein–protein interaction data, however, we can only speculate as to what the interaction partner(s) are for the members of the PEX and NEX gene families.

Birth-and-Death Evolution Has Shaped Large Parachlamydiaceae Gene Families

Large gene families are thought to generally either evolve via concerted evolution or according to a birth-and-death model (Nei and Rooney 2005). When gene families are evolving concertedly, all members experience the same evolutionary pressure and evolve as a unit. The gene family is marked by recombination between members that leads to a homogenization of all members, and thus in the phylogenetic analysis, one observes intraspecies clustering of gene family members. Although we do find minor evidence that recombination has occurred between members in the PEX and NEX gene families (supplementary fig. S8, Supplementary Material online), the effect does not appear to be that of homogenization. In contrast, long branch lengths in the trees and moderate overall sequence similarity indicate these proteins have diverged quite extensively (supplementary figs. S2–S6, Supplementary Material online). As we do not observe a dominance of intraspecies clustering in the phylogenetic analysis, these gene families are not evolving in a fashion as would be predicted via concerted evolution.

We provide clear evidence supporting birth-and-death evolution of the PEX and NEX gene families, which is marked by independent gains and losses of members (Nei and Rooney 2005). We detected frequent lineage-specific duplication and loss events, leading to high rates of variation in copy number between closely related organisms (supplementary table S3, Supplementary Material online). In the phylogenetic trees of the PEX and NEX gene families, we find interspecies clustering of members, which is a hallmark of the birth-and-death model (supplementary figs. S2–S6, Supplementary Material online). Additionally, we detected pseudogenized gene fragments of members related to the large gene families, which is another hallmark of this mode of evolution. To our knowledge, this mode of evolution has so far only been described once for a bacterial gene family (Rooney and Ward 2008).

It has been proposed that gene families that control phenotypic characters are generally subject to birth-and-death

evolution (Nei and Rooney 2005; Nei 2007). In the case of the expansions seen in the Parachlamydiaceae, this character is likely the ability to effectively infect and replicate in protist hosts. The advantage of the birth-and-death scenario, as opposed to concerted evolution, is that individual members of the gene family are able to functionally differentiate from each other and thus might facilitate the adaptation to new ecological niches (Nei 2007). Therefore, a possible driver for the birth-and-death model for the PEX and NEX families is the exploitation of new ecological niches, which in this case is most likely new host(s) species or a novel way to subvert host cell machinery.

A Confluence of Drift and Selection

Birth-and-death evolution of gene families occurs by both adaptive processes and chance events, such as genetic drift (Nei et al. 2008). This is due to the fact that, although gene duplications are intrinsically stochastic events, their fixation is influenced by both selection and genetic drift. If a duplicate is fixed, functional divergence can occur due to relaxed selection or diversifying selection in one of the gene copies, a process known as neofunctionalization (Lynch and Conery 2000). Once these genes have diverged, the new functions are then maintained in the genome via purifying selection. We envision an evolutionary scenario where the PEX and NEX gene family members rapidly diverged either due to positive diversifying or relaxed selection after duplication, likely leading to functional diversification. The window for detecting these early diversification processes is very small, but our analysis indicates that the gene family members are now being maintained via purifying selection.

The retention of such large gene families in organisms that typically undergo extreme genome reduction is perplexing, especially when considering the variation in copy number between organisms. The large number of gains and losses for F-box domain containing proteins across chlamydial lineages indicate substantial fluctuations in gene content, sometimes over short evolutionary time (fig. 5*b*). One described corollary of gene families evolving via a birth-and-death model is that the copy number variation of members within a gene family may vary due to chance duplication/loss events both between and within species, a process coined "genomic drift" (Nei et al. 2008). Genomic drift has been invoked to explain the large copy number variations in several large gene families in eukaryotes, including animal chemosensory receptors (Nozawa et al. 2007), homeobox genes (Nam and Nei 2005), and fatty-acid reductases (Eirín-López et al. 2012).

Remarkably, eukaryotic F-box and BTB-box gene families demonstrate the same dramatic evolution following a rapid birth-and-death model, and extensive copy number variation is seen both between and within eukaryotic species (Xu 2006; Navarro-Quezada et al. 2013). This is especially true in higher plant species (Stogios et al. 2005; Xu et al. 2009), where *Arabidopsis thaliana* and *Oryza sativa* harbor upwards of 800 F-box proteins (Xu et al. 2009), and large expansions have also been described in some nematode lineages (Thomas 2006). A case has been made that genomic drift also influences the evolution of these eukaryotic F-box/BTB-box protein families (Xu et al. 2009). Genomic drift therefore seems a plausible mechanism describing some aspects of evolution of the PEX and NEX gene families in the Parachlamydiaceae. Genome sequences from other chlamydial genomes and data from within populations would allow further testing of this hypothesis.

Another possibility is that the expansion of these families is the result of selection pressure due to a coevolutionary arms race with host counterpart protein(s). Under this scenario, the expansion and diversification of these families are in direct response to changes in target proteins. These evolutionary dynamics, also referred to as the Red Queen hypothesis (Valen 1974), have been most exemplified in bacteria within plant–pathogen relationships, most notably between the pathogen *Pseudomonas syringae* and its host *Ar. thaliana* (Ma et al. 2006; Baltrus et al. 2011). Copy number variation of effector proteins has been shown for *Ps. syringae*, and it is speculated that these differences confer differential host ranges (Baltrus et al. 2011). It is plausible, therefore, that variations in copy number between the PEX and NEX gene families is influenced by the host range of particular chlamydial lineages. This would imply that these proteins interact with many targets from a narrow host range or that there are a limited number of targets for a large number of possible hosts. It seems most parsimonious that the latter was the case and that these genes serve as accessory virulence factors allowing these chlamydiae to expand their host range. However, the failure to detect positive selection as the major force acting on the PEX and NEX gene family members casts doubts on a coevolutionary arms race as the sole mechanism for the evolution of these gene families.

Another conceivable scenario might be that selection pressure for increased gene dosage has contributed to the expansion of the large chlamydial gene families. In reduced genomes where the pool of regulatory proteins is limited, a path for increased protein expression might be gene duplication. This mode of "gene amplification" has been described, for instance, for the increase of antibiotic resistance genes in *E. coli*, the cholera toxin gene in *Vibrio cholerae*, and capsule biosynthesis genes in *Haemophilus influenza* (reviewed in Andersson and Hughes 2009). Although a gene dosage scenario cannot be fully ruled out for the PEX and NEX gene families, it seems unlikely as the high divergence of their members makes a completely overlapping function highly improbable. Additionally, gene amplifications are in response to strong selection pressure and would therefore be under strong positive selection. It is also known that once these selection pressures are removed, the amplified gene copies are rapidly lost within the population, often within several generations (Andersson and Hughes 2009).

We favor a hypothesis that incorporates both selection and chance into the equation. We envision that the expansion of these gene families is reflective of their role in host–microbe interaction, where they are interfering with the host ubiquitination pathway. The expansions may reflect either a change in the environmental niche, perhaps the ability to

infect a new host organism, or they may be in response to a growing number of targets within a current host. Because the PEX and NEX gene families have, respectively, expanded in the *Protochlamydia* and *Neochlamydia*, they appear to provide lineage-specific functions related to particular host interactions. The PEX and NEX variation in copy number between closely related organisms may be reflective of genomic drift, in which the independent gain and loss of members within a family has been determined, to some extent, by chance events. The fact that members of the *Chlamydiae*, including pathogenic *Chlamydia*, also harbor various proteins to subvert the host ubiquitination pathway indicates a case of convergent evolution toward exploitation of this system within this phylum.

To summarize, the *Chlamydiae* harbor several lineage-specific gene families, which are the largest among intracellular microbes with small genomes. Experimental evidence and computational analysis strongly indicate that members of these large gene families function as effector proteins involved in manipulating the ubiquitination machinery of their eukaryotic host cells. The large chlamydial gene families follow a birth-and-death model of evolution, where genomic drift may influence copy number variation. This might represent a previously undescribed mechanism by which organisms with limited exposure to larger gene pools generate genetic diversity.

Materials and Methods

Genome Sequencing

We sequenced the genomes of *P. amoebophila* EI2 (Schmitz-Esser et al. 2008), *Parachlamydia acanthamoebae* OEW1 (Heinz et al. 2007), *Neochlamydia* sp. TUME1 (Fritsche et al. 2000), and *Neochlamydia* sp. EPS4. The *Acanthamoeba* sp. harboring the latter was isolated from pond sediment from Elba, Italy. Cells and DNA were prepared as described previously (Schmitz-Esser et al. 2010). All four genomes were sequenced using 454 technology, and assemblies were performed using Newbler 2.6. We implemented an in-house pipeline for genome annotation that combines multiple approaches for gene calling and function prediction. Gene calling was performed by combining ab initio predictions from GeneMark (Besemer et al. 2001), Glimmer v3.02 (Delcher et al. 2007), Prodigal (Hyatt et al. 2010), and Critica v1.05 (Badger and Olsen 1999) with homology information derived from a BLAST search against National Center for Biotechnology Information nonredundant protein database (NCBI Resource Coordinators 2014). RNA genes were called by tRNAscan-SE (Lowe and Eddy 1997), RNAmmer (Lagesen et al. 2007), and Rfam (Griffiths-Jones et al. 2005). Function prediction was performed via BLAST against the UniProt database (UniProt Consortium 2014), and domain prediction was performed via InterProScan 5 (Jones et al. 2014). Completeness estimates were performed based on the presence of single-copy marker genes ($n = 54$) found in 99% of all bacterial genomes. Sequences have been deposited at Genbank/EMBL/DDBJ under accession numbers PRJNA242498, PRJNA242497, PRJNA242499, and PRJNA242500.

Comparative Genome Analysis

Orthologous protein groups were calculated using OrthoMCL (Li et al. 2003) with default parameters using the predicted proteomes from 19 chlamydial organisms (supplementary table S2, Supplementary Material online). For those gene families under analysis, membership to a given family was further evaluated by alignment and assessment of the phylogenetic trees. Members that had obvious major differences in the alignment and trees were dubbed spurious, and likely to have been grouped due to homology in the repeat region, and thus were removed from the gene family. Whole-genome alignments were performed using the MAUVE progressive-Mauve algorithm (Darling et al. 2010). The MAUVE alignments and local synteny plots were visualized in R (v 3.0.1) with the genoPlotR package (Guy et al. 2010).

Species Tree Construction

Phylogeny of the *Chlamydiae* was reconstructed using 32 phylogenetic marker proteins (supplementary table S2, Supplementary Material online). Multiple sequence alignments were performed with MAFFT (Katoh and Standley 2013) using the settings "–maxiterate 1000" and all alignments subsequently concatenated using FASconCAT v1.0 (Kück and Meusemann 2010). Maximum-likelihood analysis was performed with RAxML (Stamatakis 2006) with 1,000 bootstrap iterations under the PROTGAMMAGTR model, and Bayesian inference was performed with MrBayes v3.2 (Ronquist and Huelsenbeck 2003) using the mixed amino acid model and standard settings via the CIPRES science gateway (Miller et al. 2010). Alignment and tree files are available as supplementary material, Supplementary Material online.

Gene Family Analysis

As multiple sequence alignments of repeat containing proteins are not trivial, we employed several methods to obtain reliable alignments. We compared the alignments produced by MUSCLE (Edgar 2004), MAFFT using the "genapairs" option (Katoh and Standley 2013), and DIALIGN-PFAM (Al Ait et al. 2013), all with and without trimming by Gblocks using relaxed parameters "-b5=h" (Castresana 2000). In nearly all cases, the MAFFT alignment combined with Gblocks yielded the best alignment as judged by manual inspection. The exceptions were that DIALIGN-PFAM with Gblocks was the best method for the NEX1b and PEX1 alignments. To ensure robustness, we calculated neighbor joining, maximum likelihood, and Bayesian trees for each of the alignment data sets and selected the most supported tree for the final analysis. In all cases, the Bayesian trees yielded the most support, which were calculated on the CIPRES gateway (Miller et al. 2010) with MrBayes as mentioned above for the species tree. For protein domain phylogenies, we extracted and aligned the domains using the hmmalign program from HMMER3 package (Eddy 2011), and phylogenetic trees for the domains were calculated using FastTree2 (Price et al. 2010). Reconciliations of gene trees to the species tree to infer gene gain and loss were performed in Notung v2.6 (Stolzer et al. 2012), and the root for the gene trees was assigned within the program to

achieve the lowest duplication-to-loss ratio. Phylogenetic trees were visualized using iTOL (Letunic and Bork 2007) or the ETE2 toolkit (Huerta-Cepas et al. 2010). Alignment and tree files are available as supplementary material, Supplementary Material online.

Detection of Selection and Recombination

Whole-protein alignments were converted to codon alignments via the Pal2Nal program (Suyama et al. 2006). To detect the mode of selection acting on these gene families, we used CodeML from the PAML package (Yang 2007). Only sequences that were nearly full length and not contig fusions were used for the CodeML analysis. We additionally used the modified Nei–Gojobori method for codon selection from the MEGA v5.1 (Tamura et al. 2011) program with 1,000 bootstraps and treating missing data with pairwise deletions. We used the RDP4 software suite (Martin et al. 2010) to detect recombination events, which is an amalgamation of many individual recombination programs linked into one software architecture. We employed seven recombination detection programs that include RDP (Martin and Rybicki 2000), GENECONV (Sawyer 1989), MaxChi (Smith 1992), BootScan (Martin et al. 2005), Chimaera (Posada and Crandall 2001), SiScan (Gibbs et al. 2000) and 3Seq (Boni et al. 2007). Only events that were predicted with more than four programs were considered.

Intergenic Sequences Analysis

The proteome of *Protochlamydia* was used as query (tBLASTn) against the intergenic regions. The hits were conservatively filtered based on size ($>$ 50 aa), bitscore ($>$ 50), and *E* value ($<10^{-10}$). For each intergenic region when two query proteins overlap, only the best hit was considered. In addition, if different parts of the same protein had hits in one intergenic space only, the top scoring (bitscore) was considered. Putative domains in the intergenic regions were detected using InterProScan 5 (Jones et al. 2014).

Domain Distribution

For each domain of interest, we downloaded the Pfam HMM model (Finn et al. 2013) and scanned all chlamydial proteomes using the hmmscan program from the HMMER3 suite (Eddy 2011). In the case of the F-box and F-box-like domains, the results were combined into a nonredundant list. The data for the other taxa were obtained through the Pfam website under the species distribution tab for each domain or clan as in the case for the F-box. The networks were created using the "arcdiagram" package in R.

Type III Secretion Analysis

Fusion proteins containing the 5′-part of the genes of interest (including the first 20 codons) and the adenylate cyclase Cya were expressed in *Sh. Flexneri* SF401 and SF620, derivatives of the wild-type strain M90T, in which the *mxiD* and *ipaB* genes have been inactivated (Allaoui et al. 1993). The 5′-part of the target genes were amplified by polymerase chain reaction and cloned in the puc19cya vector as described (Subtil et al. 2001). Secretion assays were performed on 30 ml of exponentially grown cultures as described previously (Subtil et al. 2001). Antibodies against CRP, a cytosolic marker, were used to estimate the contamination of supernatant fractions with bacterial proteins as a result of bacterial lysis. Antibodies against IpaD, a type-III-secreted protein of *Shigella*, were used to verify that type III secretion occurred normally in the transformed strains. A monoclonal antibody against Cya and polyclonal antibodies against CRP and IpaD were generously provided by Drs N. Guiso, A. Ullmann, and C. Parsot, respectively (Institut Pasteur, Paris). Prediction of type III secretion was performed via the Effective database (Jehl et al. 2011) and the PSORTb webserver (Yu et al. 2010).

Supplementary Material

Supplementary tables S1–S3 and figures S1–S8 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

Thomas Penz and Gabriele Schwammel are acknowledged for isolating genomic DNA. This work was supported by Austrian Science Fund FWF grants Y277-B06 and I1628-B22; European Research Council ERC StG EVOCHLAMY (281633); and Marie Curie Initial Training Network Symbiomics. Genome sequencing was performed at the Norwegian Sequencing Center.

References

Al Ait L, Yamak Z, Morgenstern B. 2013. DIALIGN at GOBICS—multiple sequence alignment using various sources of external information. *Nucleic Acids Res.* 41:W3–W7.

Allaoui A, Sansonetti PJ, Parsot C. 1993. MxiD, an outer membrane protein necessary for the secretion of the *Shigella flexneri* Ipa invasins. *Mol Microbiol.* 7:59–68.

Andersson DI, Hughes D. 2009. Gene amplification and adaptive evolution in bacteria. *Annu Rev Genet.* 43:167–195.

Angot A, Vergunst A, Genin S, Peeters N. 2007. Exploitation of eukaryotic ubiquitin signaling pathways by effectors translocated by bacterial type III and type IV Secretion Systems. *PLoS Pathog.* 3:e3.

Arnold R, Brandmaier S, Kleine F, Tischler P, Heinz E, Behrens S, Niinikoski A, Mewes HW, Horn M, Rattei T. 2009. Sequence-based prediction of type III secreted proteins. *PLoS Pathog.* 5: e1000376.

Badger JH, Olsen GJ. 1999. CRITICA: coding region identification tool invoking comparative analysis. *Mol Biol Evol.* 16:512–524.

Balakirev MY, Tcherniuk SO, Jaquinod M, Chroboczek J. 2003. Otubains: a new family of cysteine proteases in the ubiquitin pathway. *EMBO Rep.* 4:517–522.

Baltrus DA, Nishimura MT, Romanchuk A, Chang JH, Mukhtar MS, Cherkis K, Roach J, Grant SR, Jones CD, Dangl JL. 2011. Dynamic evolution of pathogenicity revealed by sequencing and comparative genomics of 19 *Pseudomonas syringae* isolates. *PLoS Pathog.* 7: e1002132.

Bertelli C, Collyn F, Croxatto A, Rückert C, Polkinghorne A, Kebbi-Beghdadi C, Goesmann A, Vaughan L, Greub G. 2010. The *Waddlia* genome: a window into chlamydial biology. *PLoS One* 5: e10890.

Besemer J, Lomsadze A, Borodovsky M. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res. 29:2607–2618.
Betts HJ, Wolf K, Fields KA. 2009. Effector protein modulation of host cells: examples in the *Chlamydia* spp. arsenal. *Curr Opin Microbiol.* 12:81–87.
Blanc G, Ogata H, Robert C, Audic S, Claverie J-M, Raoult D. 2007. Lateral gene transfer between obligate intracellular bacteria: evidence from the *Rickettsia massiliae* genome. *Genome Res.* 17: 1657–1664.
Boni MF, Posada D, Feldman MW. 2007. An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics* 176:1035–1047.
Bordenstein SR, Reznikoff WS. 2005. Mobile DNA in obligate intracellular bacteria. *Nat Rev Microbiol.* 3:688–699.
Bratlie MS, Johansen J, Sherman BT, Huang DW, Lempicki RA, Drabløs F. 2010. Gene duplications in prokaryotes can be associated with environmental adaptation. *BMC Genomics* 11:588.
Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 17: 540–552.
Cho N-H, Kim H-R, Lee J-H, Kim SY, Kim J, Cha S, Kim SY, Darby AC, Fuxelius HH, Yin J, et al. 2007. The *Orientia tsutsugamushi* genome reveals massive proliferation of conjugative type IV secretion system and host-cell interaction genes. *Proc Natl Acad Sci U S A.* 104: 7981–7986.
Collingro A, Tischler P, Weinmaier T, Penz T, Heinz E, Brunham RC, Read TD, Bavoil PM, Sachse K, Kahane S, et al. 2011. Unity in variety—the pan-genome of the *Chlamydiae*. *Mol Biol Evol.* 28:3253–3270.
Darling AE, Mau B, Perna NT. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 5: e11147.
Delcher AL, Bratke KA, Powers EC, Salzberg SL. 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 23: 673–679.
Eddy SR. 2011. Accelerated profile hmm searches. *PLoS Comput Biol.* 7: e1002195.
Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32: 1792–1797.
Eirín-López JM, Rebordinos L, Rooney AP, Rozas J. 2012. The birth-and-death evolution of multigene families revisited. *Genome Dyn.* 7:170–196.
Ensminger AW, Isberg RR. 2010. E3 ubiquitin ligase activity and targeting of Bat3 by multiple *Legionella pneumophila* translocated substrates. *Infect Immun.* 78:3905–3919.
Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, et al. 2013. Pfam: the protein families database. *Nucleic Acids Res.* 42:D222–D230.
Fritsche TR, Horn M, Wagner M, Herwig RP, Schleifer K-H, Gautom RK. 2000. Phylogenetic diversity among geographically dispersed *Chlamydiales* endosymbionts recovered from clinical and environmental isolates of *Acanthamoeba* spp. *Appl Environ Microbiol.* 66: 2613–2619.
Furtado AR, Essid M, Perrinet S, Balañá ME, Yoder N, Dehoux P, Subtil A. 2013. The chlamydial OTU domain-containing protein ChlaOTU is an early type III secretion effector targeting ubiquitin and NDP52. *Cell Microbiol.* 15:2064–2079.
Gevers D, Vandepoele K, Simillon C, Van de Peer Y. 2004. Gene duplication and biased functional retention of paralogs in bacterial genomes. *Trends Microbiol.* 12:148–154.
Gibbs MJ, Armstrong JS, Gibbs AJ. 2000. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* 16:573–582.
Goldman BS, Nierman WC, Kaiser D, Slater SC, Durkin AS, Eisen JA, Ronning CM, Barbazuk WB, Blanchard M, Field C, et al. 2006. Evolution of sensory complexity recorded in a myxobacterial genome. *Proc Natl Acad Sci U S A.* 103:15200–15205.
Gomes JP, Nunes A, Bruno WJ, Borrego MJ, Florindo C, Dean D. 2006. Polymorphisms in the nine polymorphic membrane proteins of *Chlamydia trachomatis* across all serovars: evidence for serovar Da recombination and correlation with tissue tropism. *J Bacteriol.* 188: 275–286.
Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. 2005. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* 33:D121–D124.
Grimwood J, Stephens RS. 1999. Computational analysis of the polymorphic membrane protein superfamily of *Chlamydia trachomatis* and *Chlamydia pneumoniae*. *Microb Comp Genomics.* 4: 187–201.
Guy L, Kultima JR, Andersson SGE. 2010. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* 26:2334–2335.
Heinz E, Kolarov I, Kàstner C, Toenshoff ER, Wagner M, Horn M. 2007. An *Acanthamoeba* sp. containing two phylogenetically different bacterial endosymbionts. *Environ Microbiol.* 9:1604–1609.
Hooper SD, Berg OG. 2003. On the nature of gene innovation: duplication patterns in microbial genomes. *Mol Biol Evol.* 20:945–954.
Horn M. 2008. *Chlamydiae* as symbionts in eukaryotes. *Annu Rev Microbiol.* 62:113–131.
Horn M, Collingro A, Schmitz-Esser S, Beier CL, Purkhold U, Fartmann B, Brandt P, Nyakatura GJ, Droege M, Frishman D, et al. 2004. Illuminating the evolutionary history of chlamydiae. *Science* 304: 728–730.
Huerta-Cepas J, Dopazo J, Gabaldón T. 2010. ETE: a python environment for tree exploration. *BMC Bioinformatics* 11:24.
Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119.
Jehl M-A, Arnold R, Rattei T. 2011. Effective—a database of predicted secreted bacterial proteins. *Nucleic Acids Res.* 39: D591–D595.
Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30:1236–1240.
Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* 20:1313–1326.
Kalman S, Mitchell W, Marathe R, Lammel C, Fan J, Hyman RW, Olinger L, Grimwood J, Davis RW, Stephens RS. 1999. Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. *Nat Genet.* 21: 385–389.
Kamneva OK, Knight SJ, Liberles DA, Ward NL. 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.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 30:772–780.
Koonin EV, Wolf YI, Karev GP. 2002. The structure of the protein universe and genome evolution. *Nature* 420:218–223.
Kück P, Meusemann K. 2010. FASconCAT: convenient handling of data matrices. *Mol Phylogenet Evol.* 56:1115–1118.
Kuo C-H, Ochman H. 2009a. Deletional bias across the three domains of life. *Genome Biol Evol.* 1:145–152.
Kuo C-H, Ochman H. 2009b. The fate of new bacterial genes. *FEMS Microbiol Rev.* 33:38–43.
Lagesen K, Hallin P, Rødland EA, Staefeldt H-H, Rognes T, Ussery DW. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* 35:3100–3108.
Lagkouvardos I, Weinmaier T, Lauro FM, Cavicchioli R, Rattei T, Horn M. 2013. Integrating metagenomic and amplicon databases to resolve the phylogenetic and ecological diversity of the *Chlamydiae*. *ISME J.* 8:115–125.
Lerat E, Daubin V, Ochman H, Moran NA. 2005. Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol.* 3:e130.
Letunic I, Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 23: 127–128.

Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13: 2178–2189.
Lomma M, Dervins-Ravault D, Rolando M, Nora T, Newton HJ, Sansom FM, Sahr T, Gomez-Valero L, Jules M, Hartland EL, et al. 2010. The *Legionella pneumophila* F-box protein Lpp2082 (AnkB) modulates ubiquitination of the host protein parvin B and promotes intracellular replication. *Cell Microbiol.* 12:1272–1291.
Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25:955–964.
Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290:1151–1155.
Ma W, Dong FFT, Stavrinides J, Guttman DS. 2006. Type III effector diversification via both pathoadaptation and horizontal transfer in response to a coevolutionary arms race. *PLoS Genet.* 2: e209.
Magori S, Citovsky V. 2011. Hijacking of the host SCF ubiquitin ligase machinery by plant pathogens. *Front Plant Sci.* 2:87.
Martin D, Rybicki E. 2000. RDP: detection of recombination amongst aligned sequences. *Bioinformatics* 16:562–563.
Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. 2010. RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* 26:2462–2463.
Martin DP, Posada D, Crandall KA, Williamson C. 2005. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Res Hum Retroviruses.* 21:98–102.
McCutcheon JP, Moran NA. 2011. Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol.* 10:13–26.
McLeod MP, Warren RL, Hsiao WWL, Araki N, Myhre M, Fernandes C, Miyazawa D, Wong W, Lillquist AL, Wang D, et al. 2006. The complete genome of *Rhodococcus* sp. RHA1 provides insights into a catabolic powerhouse. *Proc Natl Acad Sci U S A.* 103: 15582–15587.
Miller MA, Pfeiffer W, Schwartz T. 2010. Creating the CIPRES science gateway for inference of large phylogenetic trees. In: Proceedings of the Gateway Computing Environments Workshop (GCE), 2010. p. 1–8.
Misaghi S, Balsara ZR, Catic A, Spooner E, Ploegh HL, Starnbach MN. 2006. *Chlamydia trachomatis*-derived deubiquitinating enzymes in mammalian cells during infection. *Mol Microbiol.* 61:142–150.
Myers GSA, Crabtree J, Huot CH. 2012. Deep and wide: comparative genomics of *Chlamydia*. In: Tan M, Bavoil PM, editors. Intracellular pathogens: Chlamydiales. Washington (DC): ASM Press. p. 27–50.
Nam J, Nei M. 2005. Evolutionary change of the numbers of homeobox genes in bilateral animals. *Mol Biol Evol.* 22:2386–2394.
Navarro-Quezada A, Schumann N, Quint M. 2013. Plant F-box protein evolution is determined by lineage-specific timing of major gene family expansion waves. *PLoS One* 8:e68672.
NCBI Resource Coordinators. 2014. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 42:D7–D17.
Nei M. 2007. The new mutation theory of phenotypic evolution. *Proc Natl Acad Sci U S A.* 104:12235–12242.
Nei M, Niimura Y, Nozawa M. 2008. The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. *Nat Rev Genet.* 9:951–963.
Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet.* 39:121–152.
Nogueira T, Touchon M, Rocha EPC. 2012. Rapid evolution of the sequences and gene repertoires of secreted proteins in bacteria. *PLoS One* 7:e49403.
Nozawa M, Kawahara Y, Nei M. 2007. Genomic drift and copy number variation of sensory receptor genes in humans. *Proc Natl Acad Sci U S A.* 104:20421–20426.
Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405:299–304.
Perez-Torrado R, Yamada D, Defossez P-A. 2006. Born to bind: the BTB protein–protein interaction domain. *BioEssays* 28:1194–1202.
Peters J, Wilson DP, Myers G, Timms P, Bavoil PM. 2007. Type III secretion à la *Chlamydia*. *Trends Microbiol.* 15:241–251.
Posada D, Crandall KA. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc Natl Acad Sci U S A.* 98:13757–13762.
Price CT, Kwaik YA. 2010. Exploitation of host polyubiquitination machinery through molecular mimicry by eukaryotic-like bacterial F-box effectors. *Cell Infect Microbiol.* 1:122.
Price CT, Al-Khodor S, Al-Quadan T, Santic M, Habyarimana F, Kalia A, Kwaik YA. 2009. Molecular mimicry by an F-box effector of *Legionella pneumophila* hijacks a conserved polyubiquitination machinery within macrophages and protozoa. *PLoS Pathog.* 5: e1000704.
Price CTD, Al-Quadan T, Santic M, Rosenshine I, Kwaik YA. 2011. Host proteasomal degradation generates amino acids essential for intracellular bacterial growth. *Science* 334:1553–1557.
Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490.
Pushker R, Mira A, Rodriguez-Valera F. 2004. Comparative genomics of gene-family size in closely related bacteria. *Genome Biol.* 5: R27.
Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.
Rooney AP, Ward TJ. 2008. Birth-and-death evolution of the internalin multigene family in *Listeria*. *Gene* 427:124–128.
Sakakibara A, Hattori S. 2000. Chat, a Cas/HEF1-associated adaptor protein that integrates multiple signaling pathways. *J Biol Chem.* 275: 6404–6410.
Santoyo G, Romero D. 2005. Gene conversion and concerted evolution in bacterial genomes. *FEMS Microbiol Rev.* 29:169–183.
Sawyer S. 1989. Statistical tests for detecting gene conversion. *Mol Biol Evol.* 6:526–538.
Schmitz-Esser S, Tischler P, Arnold R, Montanaro J, Wagner M, Rattei T, Horn M. 2010. The genome of the amoeba symbiont "*Candidatus* Amoebophilus asiaticus" reveals common mechanisms for host cell interaction among amoeba-associated bacteria. *J Bacteriol.* 192: 1045–1057.
Schmitz-Esser S, Toenshoff ER, Haider S, Heinz E, Hoenninger VM, Wagner M, Horn M. 2008. Diversity of bacterial endosymbionts of environmental *Acanthamoeba* isolates. *Appl Environ Microbiol.* 74: 5822–5831.
Smith JM. 1992. Analyzing the mosaic structure of genes. *J Mol Evol.* 34: 126–129.
Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L, Mitchell W, Olinger L, Tatusov RL, Zhao Q, et al. 1998. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* 282:754–759.
Stogios PJ, Downs GS, Jauhal JJ, Nandra SK, Prive GG. 2005. Sequence and structural analysis of BTB domain proteins. *Genome Biol.* 6:R82.
Stolzer M, Lai H, Xu M, Sathaye D, Vernot B, Durand D. 2012. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics* 28: i409–i415.
Subtil A, Collingro A, Horn M. 2014. Tracing the primordial *Chlamydiae*: extinct parasites of plants? *Trends Plant Sci.* 19:36–43.
Subtil A, Parsot C, Dautry-Varsat A. 2001. Secretion of predicted Inc proteins of *Chlamydia pneumoniae* by a heterologous type III machinery. *Mol Microbiol.* 39:792–800.
Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34:W609–W612.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol.* 28:2731–2739.

Thomas JH. 2006. Adaptive evolution in two large families of ubiquitin-ligase adapters in nematodes and plants. *Genome Res.* 16:1017–1030.

Treangen TJ, Rocha EPC. 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet.* 7:e1001284.

UniProt Consortium. 2014. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* 42:D191–D198.

Valen LV. 1974. Molecular evolution as predicted by natural selection. *J Mol Evol.* 3:89–101.

Willems AR, Schwab M, Tyers M. 2004. A hitchhiker's guide to the cullin ubiquitin ligases: SCF and its kin. *Biochim Biophys Acta.* 1695: 133–170.

Xu G, Ma H, Nei M, Kong H. 2009. Evolution of F-box genes in plants: different modes of sequence divergence and their relationships with functional diversification. *Proc Natl Acad Sci U S A.* 106:835–840.

Xu J. 2006. Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. *Mol Ecol.* 15: 1713–1731.

Yang J, Lusk R, Li W-H. 2003. Organismal complexity, protein complexity, and gene duplicability. *Proc Natl Acad Sci U S A.* 100: 15661–15665.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.

Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ, et al. 2010. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 26:1608–1615.

Zhang J. 2003. Evolution by gene duplication: an update. *Trends Ecol Evol.* 18:292–298.

Zheng N, Schulman BA, Song L, Miller JJ, Jeffrey PD, Wang P, Chu C, Koepp DM, Elledge SJ, Pagano M, et al. 2002. Structure of the Cul1–Rbx1–Skp1–F boxSkp2 SCF ubiquitin ligase complex. *Nature* 416: 703–709.

Massive expansion of ubiquitination-related gene families within the *Chlamydiae*

Domman *et al*.

Supplementary Information

Table S1: Genome features of members of the *Parachlamydiaceae*

	***Protochlamydia amoebophila* UWE25**	***Protochlamydia amoebophila* EI2**	***Neochlamydia* sp. TUME1**	***Neochlamydia* sp. EPS4**	***Parachlamydia acanthamoeabae* OEW1**	***Parachlamydia acanthamoeabae* UV-7**
Sequencing approach	Sanger	454	454	454	454	Sanger
Assembly		Newbler 2.6	Newbler 2.6	Newbler 2.6	Newbler 2.6	
Sequence length (nt)	2,417,793	2,397,675	2,546,323	2,530,677	3,008,885	3,072,383
Contigs >1 kb	1	178	254	112	162	1
Predicted CDSs	2,031	2,150	2,345	2,174	2,756	2,788
G+C content (%)	35	35	38	38	39	39
Coding regions (%)	82	81	80	80	88	90
Average CDSs length (nt)	1,003	900	867	934	956	988
tRNAs	37	36	36	36	38	40

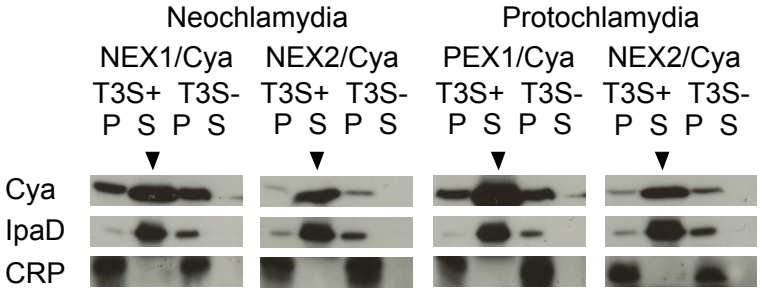
Table S2: Organisms used in this study and proteins used for species tree construction

Organism	Accession number	Genome size	Largest single expansion	Total expansions
***Chlamydia muridarum* Nigg**	NC_002620.2	1080453	4	10
***Chlamydia trachomatis* 434/Bu**	NC_010287.1	1038843	2	4
***Chlamydia trachomatis* D/UW-3/CX**	NC_000117.1	1042519	2	5
***Chlamydia trachomatis* L2c**	NC_015744.1	1038814	2	6
***Chlamydophila abortus* S26/3**	NC_004552.2	1144378	3	7
***Chlamydophila caviae* GPIC**	NC_003361.3	1181358	4	13
***Chlamydophila felis* Fe/C-56**	NC_007899.1	1173793	6	17
***Chlamydophila pecorum* E58**	NC_015408.1	1106198	6	13
***Chlamydophila pneumoniae* CWL029**	NC_000922.1	1230231	6	24
***Chlamydophila pneumoniae* LPCoLN**	NC_017285.1	1248552	9	23
***Chlamydophila psittaci* 6BC**	NC_015470.1	1179222	6	14
***Neochlamydia sp.* EPS4**	this study	2530677	94	212
***Neochlamydia sp.* TUME1**	this study	2546323	68	203
***Parachlamydia sp.* OEW-1**	this study	3008885	8	92
***Parachlamydia acanthamoebae* UV-7**	NC_015702.1	3072383	12	92
***Protochlamydia amoebophila* UWE25**	NC_005861.1	2414465	41	154
***Protochlamydia amoebophila* El2**	this study	2397675	32	154
***Simkania negevensis* Z**	NC_015713.1	2496337	12	277
***Waddlia chondrophila* WSU 86-1044**	NC_014225.1	2116312	11	155

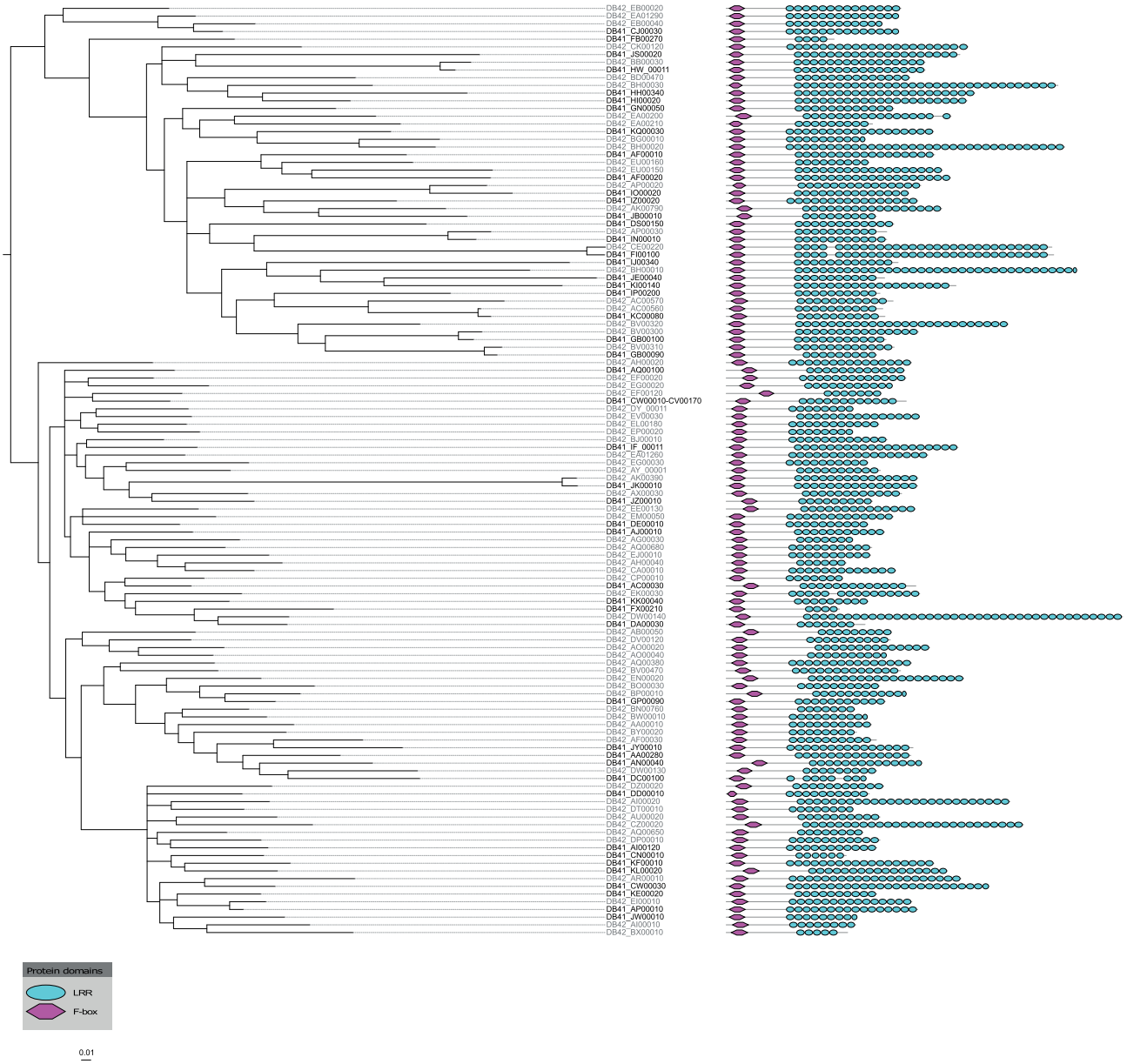
Ribosomal proteins		Other marker proteins
Large subunit	**Small subunit**	
rl1	rs3	RpoB
rl2	rs4	RpoC
rl3	rs5	GyrB
rl4	rs8	RecA
rl5	rs9	EfTu
rl9	rs15	
rl10	rs17	
rl11	rs18	
rl13	rs19	
rl14		
rl16		
rl20		
rl22		
rl23		
rl24		
rl27		
rl31		

Table S3. Gains and losses for large *Parachlamydiaceae* gene families. Events for each gene family after Notung reconciliation of the gene trees with the species tree are given as 'duplications/losses' for each organism.

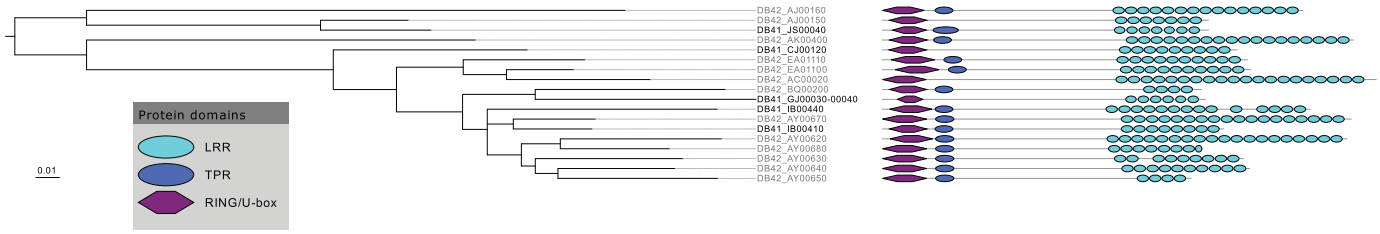
	NEX1a	NEX1b	NEX2	PEX1	PEX2
Proto UWE25	NA	NA	NA	7/7	0/2
Proto EI2	NA	NA	NA	5/14	0/3
Neo EPS4	19/13	6/1	3/5	NA	NA
Neo TUME1	4/20	0/3	5/6	NA	NA
Total Expansions	78	12	25	49	14
Total Losses	33	4	11	21	5



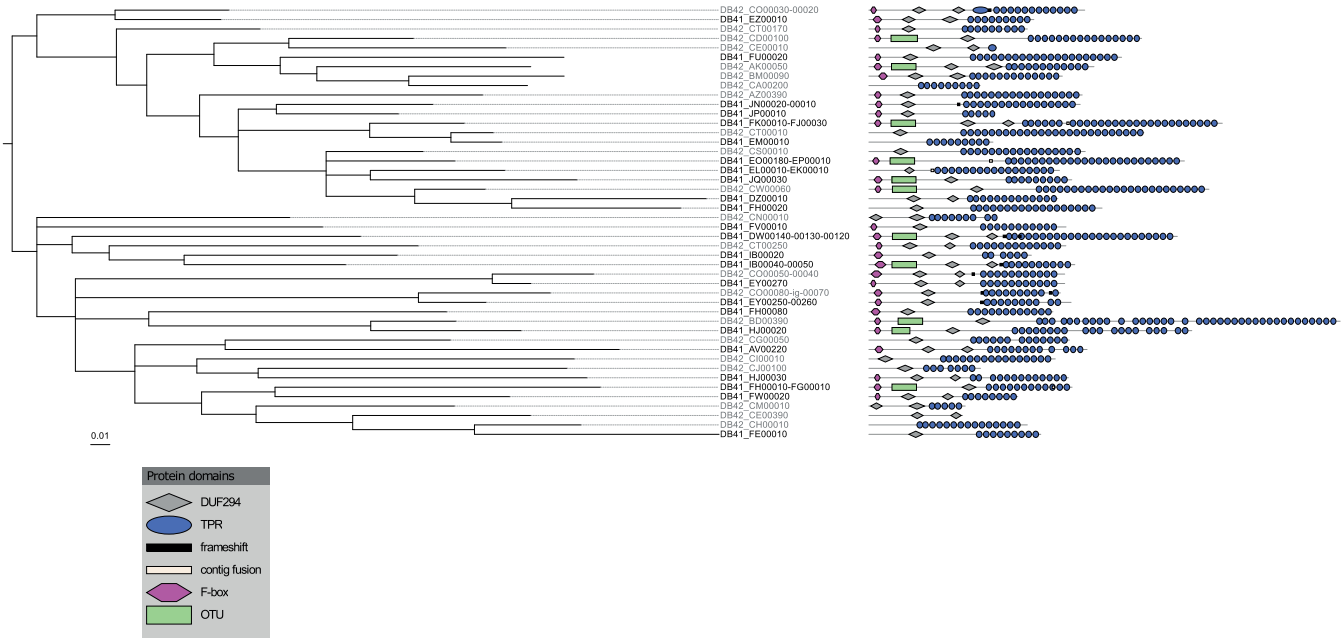
Supplementary Figure S1. **Type III secretion assay in *Shigella*.** A representative member from each of the gene families (NEX1: DB42_AK 00400; NEX2: DB42_CW00060; PEX1: YP_007742.1; PEX2: YP_007044.1) was tested for the presence of a functional type III secretion (T3S) signal in its N-terminus using *Shigella flexneri*. Chimera between the first 20 codons of the chlamydial genes and the reporter gene *cya* were transformed into a T3S competent strain (*ipaB*, T3S+) and a T3S deficient strain (*mxiD*, T3S-). Liquid cultures were fractionated into pellet (P) and supernatant (S), the protein extracts were run on a SDS-PAGE and transferred to a membrane. The membrane was probed with antibodies against IpaD, a known T3S substrate of *Shigella*, CRP, a cytosolic marker protein that serves as a control for cell lysis, and Cya. Each of the Cya chimeras tested was observed in the supernatant of the T3S+ strain (black triangles), and not of the T3S strain, demonstrating the presence of a functional T3S signal.



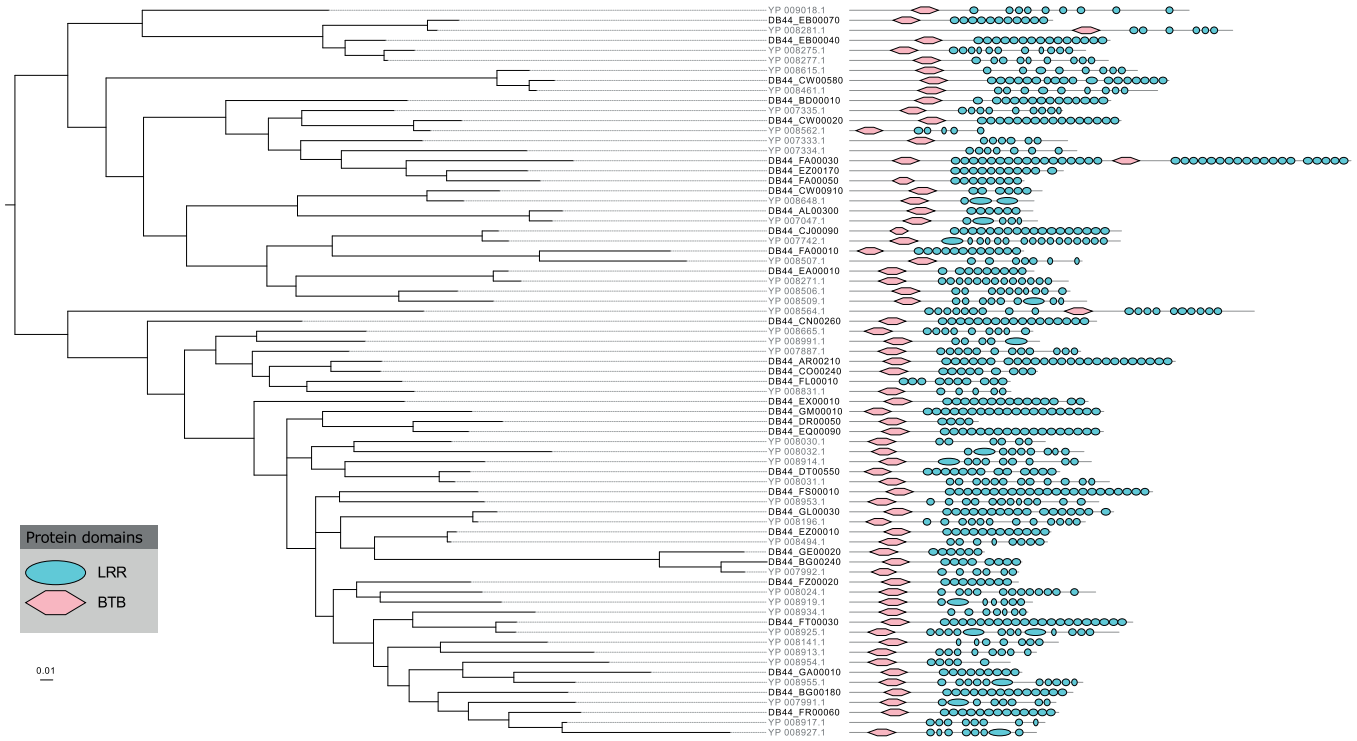
Supplementary Figure S2. **Phylogeny of the NEX1a gene family.**
A Bayesian reconstruction of the phylogenetic relationship between the NEX1a members is shown along with the corresponding protein domain architecture. There is a conserved F-box domain at the N-terminus followed by LRR domains. A region with no detectable domains between the F-box and LRR is conserved between the members. Shading in the sequence names differentiate between species, where *Neochlamydia* sp. TUME1 is colored in black and *Neochlamydia* sp. EPS4 is in grey.



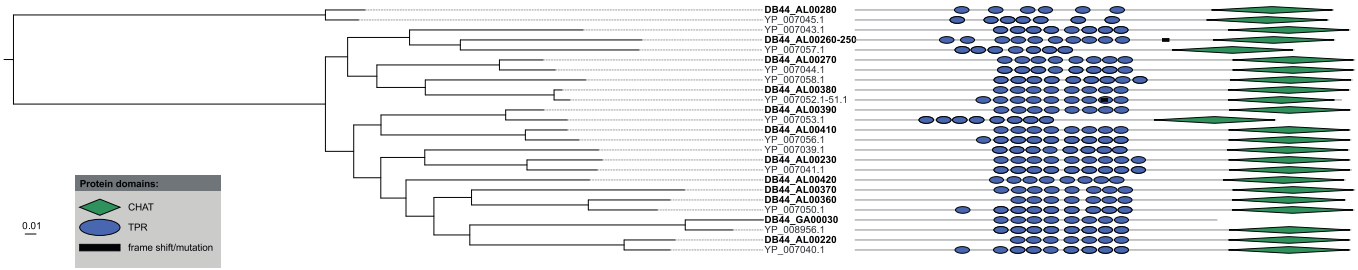
Supplementary Figure S3. **Phylogeny of the NEX1b gene family.**
A Bayesian reconstruction of the phylogenetic relationship between NEX1b members is shown along with the corresponding protein domain architecture. At the N-terminus there is a conserved RING/U-box domain followed immediately by a TPR domain. The C-terminus of the protein consists of various copies of LRR domains. Locus tags colored in black correspond to *Neochlamydia* sp. TUME1 and those in grey to *Neochlamydia* sp. EPS4.



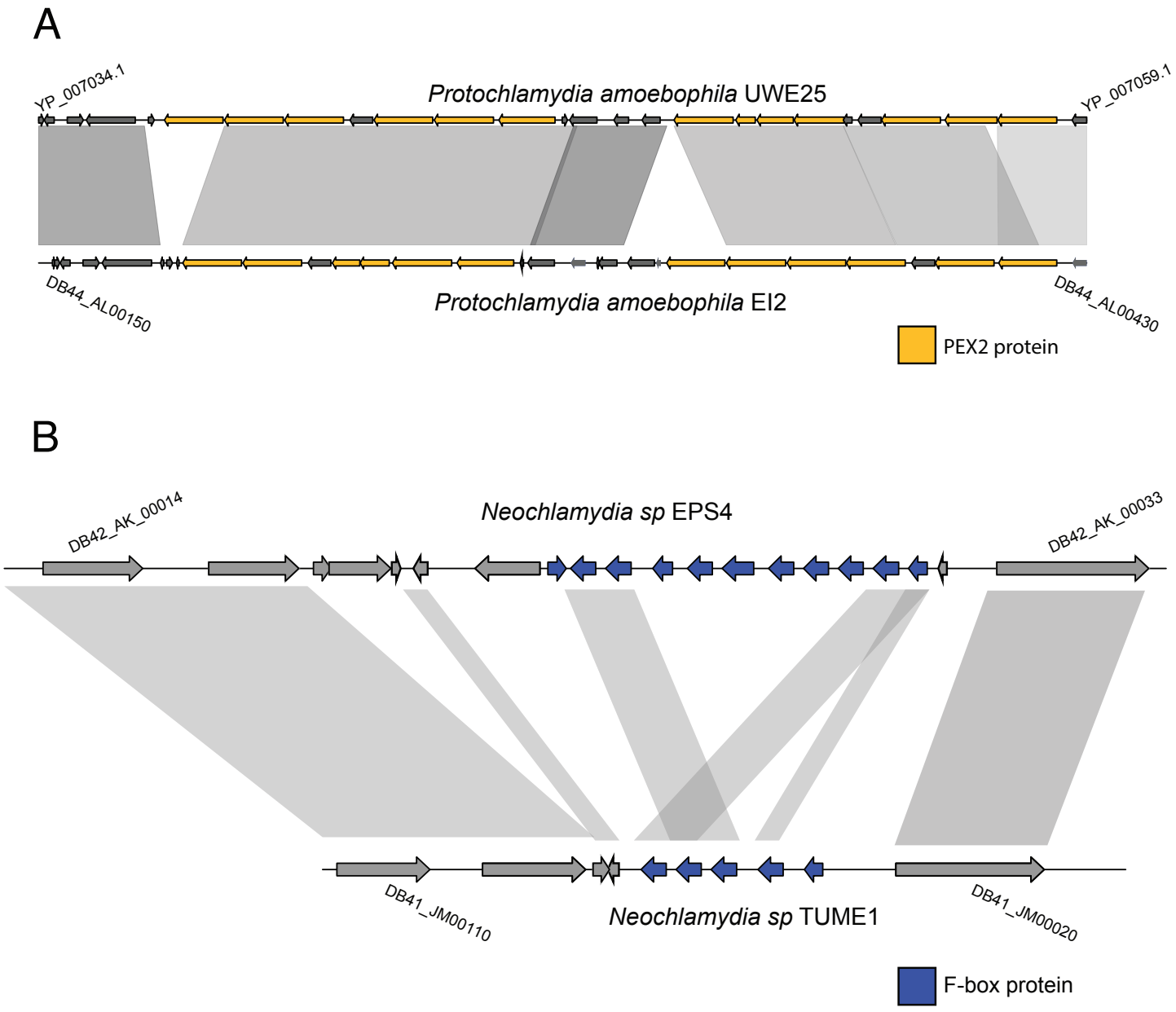
Supplementary Figure S4. **Phylogeny of the NEX2 gene family.**
A Bayesian reconstruction of the phylogenetic relationship between NEX2 members is shown along with the corresponding protein domain architecture. In the majority of members, there is an F-box domain at the N-terminus. In some cases (n=11) an OTU domain immediately follows. All members then harbor repeating DUF294 domains and a C-terminus of TPR domains. Contigs have been fused where indicated by a dash. Colors in the locus tags denote species differences where black and grey correspond to *Neochlamydia* sp. TUME1 and *Neochlamydia* sp. EPS4, respectively.



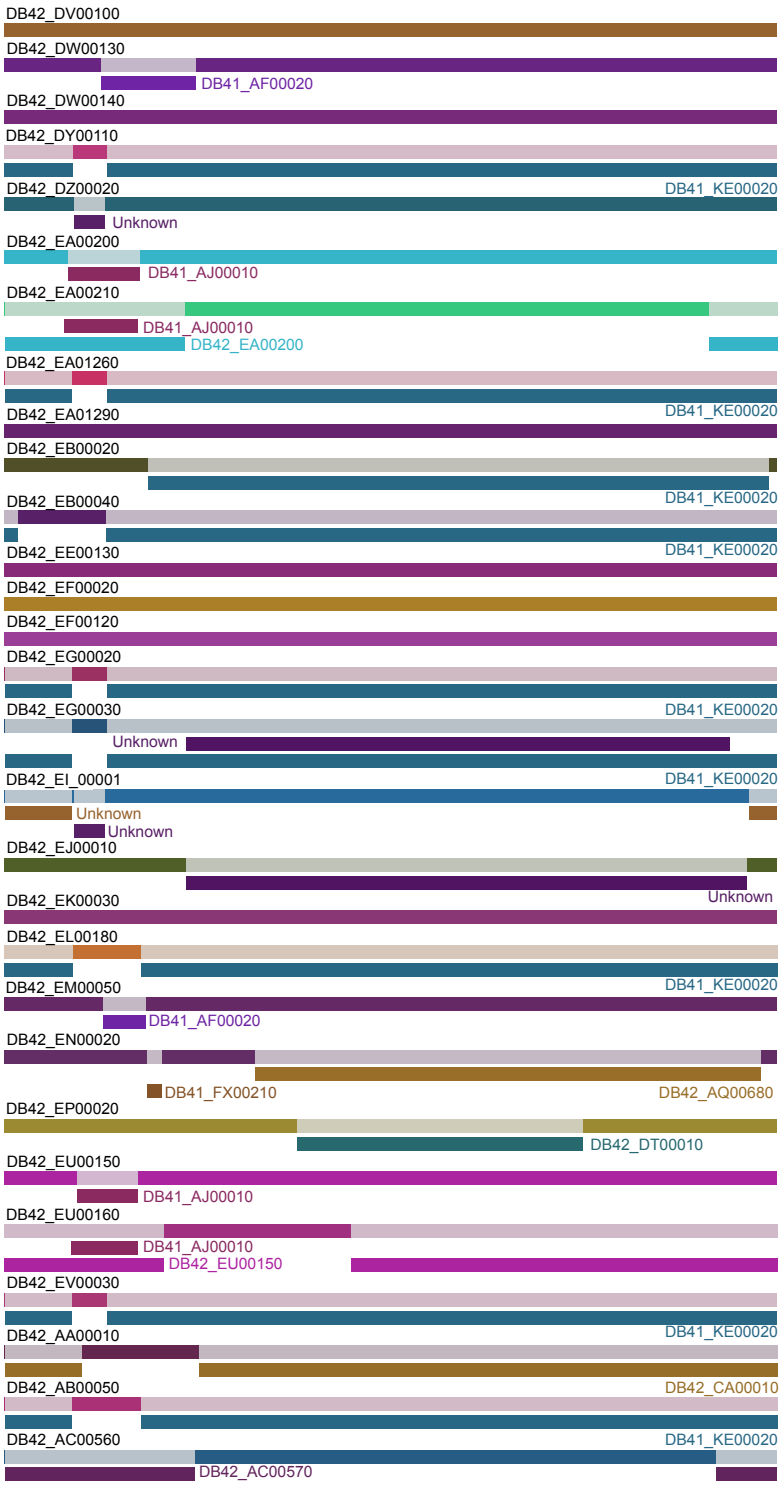
Supplementary Figure S5. **Phylogeny of the PEX1 gene family.**
A Bayesian reconstruction of the phylogenetic relationship between PEX1 members is shown along with the corresponding protein domain architecture. A BTB-box domain at the N-terminus is followed by multiple copies of LRR domains at the C-terminus. Locus tags colored in black correspond to *Protochlamydia amoebophila* EI2, while those in grey to *Protochlamydia ameobophila* UWE25.



Supplementary Figure S6. **Phylogeny of the PEX2 gene family.** A Bayesian reconstruction of the phylogenetic relationship between PEX2 members is shown along with the corresponding protein domain architecture. There is no apparent N-terminal domain present in these members, but various copies of TPR domains appear in the middle of the protein. The C-terminus is marked by the presence of a CHAT domain. Black locus tags correspond to *Protochlamydia amoebophila* EI2, while grey locus tags are used for *Protochlamydia amoebophila* UWE25.



Supplementary Figure S7. **Evidence for tandem duplications among expanded *Parachlamydiaceae* gene families.** Members of the PEX2 gene family (A) in *Protochlamydia* and F-box proteins (B) in *Neochlamydia* are found in tandem arrays, which are strongly indicative of gene duplication events. Synteny between genomes is depicted with gray bars.



Supplementary Figure S8. **Example of recombination detection within the NEX1a gene family.** The output of the RDP4 program for detecting recombination in the NEX1a gene family is presented. Each sequence is represented by a full bar, with possible recombination points and the donor sequence shown below. Most of the predicted recombination regions are weakly significant. Note that many of the predicted recombination sites are in the vicinity of the N-terminal F-box-like domain.

Chapter IV

Following the footsteps of chlamydial gene regulation

Title:

Following the footsteps of chlamydial gene regulation

Authors:

D. Domman[a] and M. Horn[a,1]

Author affiliations:

[a] Department of Microbiology and Ecosystem Science, University of Vienna, Althanstrasse 14, A-1090, Vienna, Austria.

Corresponding author:

Univ.-Prof. Dr. Matthias Horn
University of Vienna
Department of Microbiology and Ecosystem Science
Althanstr. 14
A-1090 Wien
Austria
phone: +43 1 4277 76608
email: horn@microbial-ecology.net

Abstract

Regulation of gene expression ensures an organism responds to stimuli and undergoes proper development. While the regulatory networks in bacteria have been investigated in model microorganisms, nearly nothing is known about the evolution and plasticity of these networks in obligate, intracellular bacteria. The phylum *Chlamydiae* contains a vast array of host-associated microbes, including several human pathogens. The *Chlamydiae* are unique among obligate, intracellular bacteria as they undergo a complex bi-phasic developmental cycle in which large swaths of genes are temporally regulated. Coupled with the low number of transcription factors, these organisms offer a model to study the evolution of regulatory networks in intracellular organisms. We provide the first comprehensive analysis exploring the diversity and evolution of regulatory networks across the phylum. We utilized a comparative genomics approach to construct predicted co-regulatory networks, which unveiled genus and family specific regulatory motifs and architectures, most notably those of virulence-associated genes. Surprisingly, our analysis suggests that few regulatory components are conserved across the phylum, and those that are conserved are involved in the exploitation of the intracellular niche. Our study thus lends insight into a component of chlamydial evolution that has otherwise remained largely unexplored.

Introduction

All organisms rely on regulatory mechanisms to control the expression of certain genes at certain times or in response to certain stimuli. In bacteria, regulation of gene expression is often carried out by DNA-binding proteins that recognize specific motifs found in promoter regions and serve to either activate or repress transcription via interactions with RNA polymerase. These transcription factors and their target genes thus comprise how a cell may respond to different environmental or developmental signals (Perez and Groisman 2009). These regulatory networks may be highly conserved between related organisms, but growing evidence suggests that many of these networks confer species-specific regulator and target gene associations (Price et al. 2007). The evolution of regulatory networks in a given organism is thus highly reflective of its environment, where free living bacteria harboring large and diverse gene sets also contain a proportional number of regulatory factors. For instance, the soil bacterium *Streptomycetes avermitilis* has a genome size of 9.1 Mb (7,582 predicted genes) and is predicted to harbor 623 regulatory proteins (Madan Babu et al. 2006). Contrarily, bacterial symbionts, which experience a stable intracellular environment, have reduced genomes and tend to harbor few regulatory elements, such as the aphid endosymbiont *Buchnera aphidicola* which is predicted to only encode 4 regulatory proteins for its 507 predicted genes (Madan Babu et al. 2006). Unique among obligate, intracellular bacteria are the *Chlamydiae*, which undergo a biphasic developmental cycle, in which hundreds of genes must be temporally regulated. This conserved developmental cycle, taken with the diversity of ecological niches occupied by chlamydiae and their respective hosts, and reduced impact of horizontal gene transfer in the phylum makes the *Chlamydiae* prime candidates to study the evolution of regulatory networks among intracellular bacteria.

All members of the phylum *Chlamydiae* are associated with eukaryotic hosts. The family *Chlamydiaceae* includes many well-known animal and human pathogens, including the largest contributor to bacterial sexually transmitted disease, *Chlamydia trachomatis*. Outside of this family lies a vast array of chlamydiae that are collectively referred to as "environmental chlamydia". There are at least eight described families outside of the

Chlamydiaceae, whose members are associated with a smorgasbord of eukaryotes, ranging from protists, enigmatic marine worms, arthropods, and fish (Horn 2008; Lagkouvardos et al. 2014; Taylor-Brown et al. 2015). Despite this tremendous diversity in host range, a paramount unifying feature is a shared biphasic developmental cycle in which an infectious, extracellular elementary body (EB) enters a host cell, and transitions into a replicative and fully metabolically active reticulate body (RB). Following replication, the RBs differentiate back to EBs and are subsequently released into the environment, usually as a result of host cell lysis. Several pioneering transcriptomics studies in the human pathogens *Chlamydia trachomatis* (Belland et al. 2003; Nicholson et al. 2003) and *Chlamydia pneumoniae* (Mäurer et al. 2007; Albrecht et al. 2011) illustrated that this developmental cycle is marked by differential temporal expression patterns of large sets of genes, which have been broadly characterized as early (EB to RB conversion), mid (RB replication), and late (RB to EB conversion).

Despite the wide importance of these organisms in animal and human health, the infancy of tools for genetic manipulation (Heuer et al. 2007; Nguyen and Valdivia 2013) and the difficulty of intracellular systems have made the elucidation of the major regulatory players arduous (Tan 2012). Only a handful of microarray and RNA sequencing studies are available (Belland et al. 2003; Nicholson et al. 2003; Mäurer et al. 2007; Albrecht et al. 2011) and within this subset even fewer provide the resolution needed to characterize expression profiles over the developmental cycle. Here, we utilize the power of phylogenomics to lend insights into the evolution and diversity of the regulatory proteins and schemes we find distributed throughout the phylum. We systematically predicted transcription factors found in the chlamydial phylum, and we provide the first comprehensive prediction of transcription regulatory networks for various members of the *Chlamydiae*.

Results and Discussion

Diversity of regulatory elements reflects host diversity and ecology

The genome size of members of the *Chlamydiae* varies over 2 Mb, from the smallest genome of *Chlamydia trachomatis* (1.04 Mb) to the largest of *Parachlamydia acanthamoeba* (3.07 Mb). Of the 9,933 gene families in the phylum, only 409 families are conserved between all *Chlamydiae*. This indicates a small subset of genes that likely function in core chlamydial biology, such as the developmental cycle, and a large repertoire of genes that likely serve more specific roles tailored to each organism's environment (Collingro et al. 2011). Diversity and expansion of gene content is often met with the need to regulate these genes (Perez and Groisman 2009). To see if this large diversity of genes was matched with increased regulatory elements, we exhaustively searched for predicted regulators in the genomes of all sequenced chlamydia by identifying all proteins containing DNA-binding domains and/or sequence homology to known regulators. Combined with the nine previously described regulators in the *Chlamydiaceae*, we uncovered a striking diversity of 73 putative regulators found with varying frequency throughout the phylum (Figure 1, Supplementary Table S1). Indeed, we observe that the more reduced genomes of the *Chlamydiaceae* harbor relatively few transcription factors (12-15), whereas the larger genomes of the environmental chlamydia harbor an extensive diversity of putative regulators. It is interesting to note, however, that the largest chlamydial genomes do not harbor the largest set of predicted regulatory elements, that honor is bestowed upon *Rubidus massiliensis*, which harbors 37 predicted regulators. *Protochlamydia naegleriophila* and *Simkania negevensis* both contain 31 predicted regulators, however they vary in genome size by nearly ~0.5 Mb (which in real terms is half the size of the genome of *C. trachomatis*). *S. negevensis* notably harbors the most unique set of predicted regulators, with 11 being unique to only this organism.

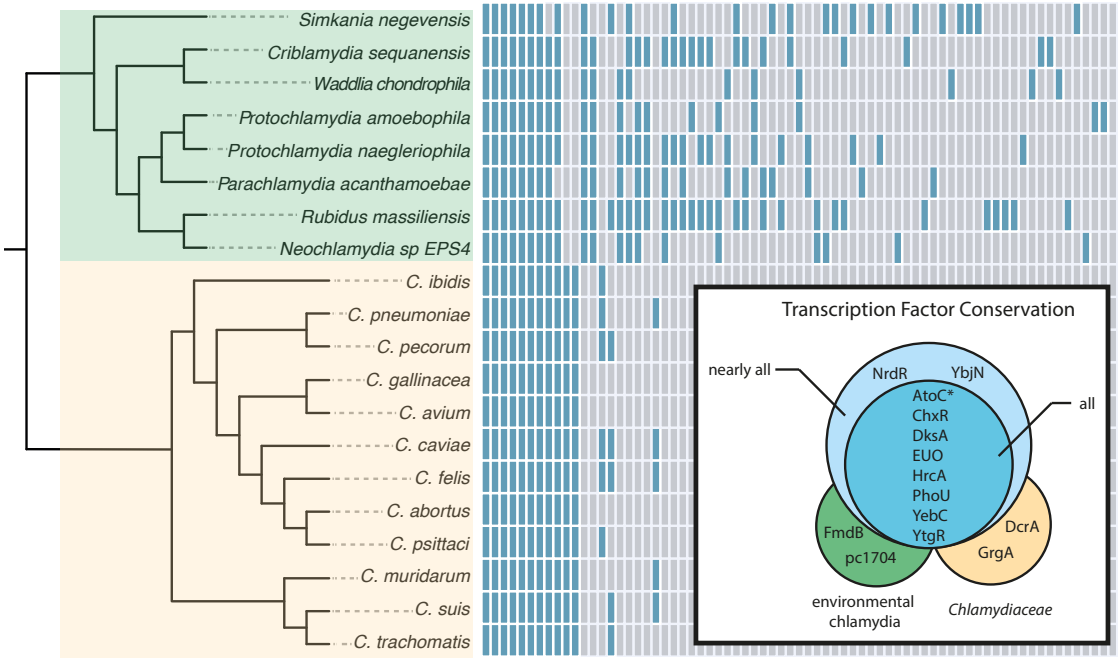


Figure 1. Distribution and conservation of putative transcriptional regulators in the *Chlamydiae*. The plot shows gene presence and absence of predicted regulators for members of the phylum *Chlamydiae* (yellow background indicates members of the family *Chlamydiaceae*; green background indicates environmental chlamydiae). The species phylogeny shown was calculated from a concatenation of 33 markers genes using PhyloBayes (Lartillot et al. 2013) under the CAT-GTR model. The box inlay displays those genes that are conserved throughout different taxonomic levels. We included AtoC as a globally conserved regulator, as it is only absent in the incomplete genome of *Criblamydia* (*). pc1704 refers to the *Protochlamydia amoebophila* UWE25 locus tag of the conserved, yet uncharacterized protein.

This large disparity in the predicted regulatory elements may reflect the differences in host and environmental niches of each organism. All members of the *Chlamydiaceae* infect higher animal hosts, such as humans, koala, birds, and reptiles (Horn 2008; Lagkouvardos et al. 2013; Taylor-Brown et al. 2015), which suggest that these organisms are well adapted

to this particular intracellular environment. This is in contrast with most of the sequenced environmental chlamydia, which primarily have been isolated from free-living amoeba in both soil and aquatic environments (Horn 2008; Lagkouvardos et al. 2013; Taylor-Brown et al. 2015). These environments are much more tumultuous and thus these chlamydial organisms must have a genetic repertoire to compete effectively against other facultative amoeba-associated organisms, such as *Legionella* species (Moliner et al. 2010), and to survive the harsh conditions while in the extracellular EB stage. From the current fully sequenced environmental chlamydia, only two organisms were not originally isolated from free-living amoeba. *Waddlia chondrophila* was first isolated from an aborted bovine fetus (Rurangirwa et al. 1999), and *Simkania negevensis* was originally discovered as a contaminant in human cell culture (Kahane et al. 1995), however both organisms grow well in *Acanthamoeba* species. (Horn 2008).

Here we find that several of the unique or sparsely distributed transcription factors are putatively involved in regulating operons that function in distinct metabolite metabolism or transport (Supplementary table S1), such as arsenic resistance (ArsR), amino acid metabolism (ArgR, TrpR), and carbon storage (CsrA, CsiR). Speculation on the actual roles of regulatory proteins is difficult, as orthologous transcription factors can have vastly different functions, even between closely related bacteria (Price et al. 2007). Most transcription factors acquired via horizontal gene transfer often are local regulators, meaning they only control a small, typically adjacent, subset of genes (Price et al. 2008). Indeed, most of the transcription factors that are species specific are adjacent to other non-conserved genes, although this does not necessarily mean these genes are under the control of the "local" regulator. However, we do find an apparent co-transfer of the mercury resistance repressor and the corresponding operon into the plasmid of *S. negevensis*, consisting of the regulator MerR, the mercuric reductase MerA, and one membrane spanning protein MerT (Boyd and Barkay 2012). Evidence for horizontal acquisition is also exemplified with a member of the LysR family of regulators that was transferred into the ancestor of *Simkania* and *Criblamydia* from members of the *Alphaproteobacteria* (Figure S1). The presence of LysR in only two chlamydiae suggests

that multiple losses also occurred, such as in *Waddlia* and all *Parachlamydia,* when we reconcile this data with the species tree (shown in Figure 2).

When we modeled the gene family history of all putative transcription factors (Csűös 2010), we find that there has been a veritable mix of gains and losses in every family except for the *Chlamydiaceae* (Figure 2). The *Chlamydiaceae* have only acquired three transcription factors when they split from the other families (DcrA, GrgA, and the plasmid regulator pGP4), and all other changes in the transcription factor repertoire have been differential losses. This is in stark contrast to the environmental chlamydia, where a total of 63 regulator genes were acquired in various lineages (Figure 2). Notably, we even find gains and losses among members of the same species, for instance between the two *Protochlamydia* species infecting different amoeba hosts, which may indicate that adaptation to novel niches may have been facilitated via changes in gene regulation.

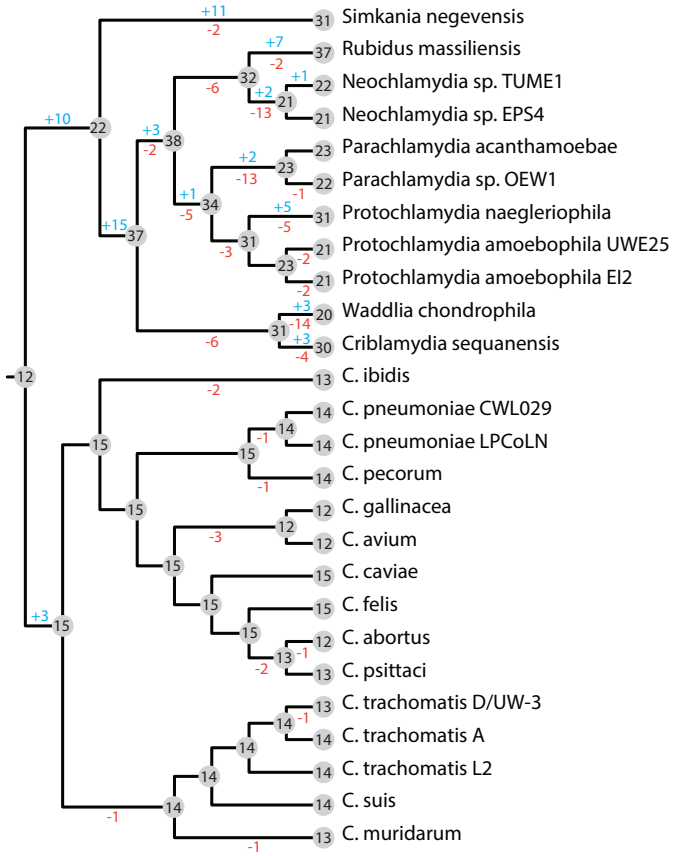


Figure 2. Gain and loss of transcription factor gene families during *Chlamydiae* evolution. Using the phyletic profile of each gene family containing a predicted regulator, we used the Dollo parsimony approach in COUNT (Csűös 2010) to model the gains and losses of transcription factors along the chlamydial species tree. The environmental chlamydial have undergone extensive gains and losses, whereas the *Chlamydiaceae* have only three gains at the initial family separation and are then marked by differential losses.

Conserved regulators are essential to chlamydial biology

Moving from the surprising diversity of regulators found in the phylum *Chlamydiae*, those regulators that are conserved are those that likely confer essential and fundamental roles in chlamydial biology. Out of all putative transcription factors, only eight (AtoC, ChxR, DksA, EUO, HrcA, PhoU, YebC, YtgR) are conserved among all members of the phylum (Figure 1; Supplementary table S1). An additional two (YbjN, NrdR) are nearly conserved (absent in < 2 genomes) in all *Chlamydiae*, and two described additional factors (DcrA, GrgA) are specific for the *Chlamydiaceae* (Rau et al. 2005; Bao et al. 2012). The role of DcrA as a transcription factor, however, is currently under debate (Kemege et al. 2011). Some of these phylum-wide conserved regulators have been implicated as major players in the chlamydial developmental cycle. For example, EUO has donned the title of the master regulator of late gene expression (RB to EB conversion) in *Chlamydiaceae*, where previous studies have nicely shown that EUO represses the transcription of late genes (Rosario and Tan 2012)(Rosario et al. 2014). Several studies have demonstrated that global regulators, that is those regulators which control large numbers of target genes, evolve more slowly than other regulators, and tend to be vertically inherited (Rajewsky et al. 2002; Price et al. 2008; Perez and Groisman 2009). Thus, the individual gene trees for global regulators tend to be concordant with species trees (Price et al. 2008). As EUO currently represents the only *bona fide* global regulator in the phylum we chose to reconstruct the phylogeny of this gene. Indeed, the gene tree for EUO is highly concordant with the chlamydial species tree (Figure S2), indicative that EUO has been vertically inherited throughout chlamydial evolution, providing further evidence for its role as global regulator in all chlamydiae.

Additionally, ChxR has been implicated as an activator of midcyle genes, where chlamydia are fully metabolically active and dividing as RBs (Koo et al. 2006; Hickey et al. 2011). YtgR has been shown to negatively regulate the *ytg* operon which is believed to function in metal ion transport (Akers et al. 2011), and likely does not have a major role in developmental cycle regulation. This is likely also the case for the acetate metabolism regulator AtoC and the heat shock response regulator HrcA, which is involved in response to cellular stress and has been experimentally characterized to regulate the *dnaK* and *groE* operons in *C. trachomatis* (Wilson and Tan 2004; Wilson et al. 2005). The putative phosphate regulator PhoU has not been investigated, but likely has a limited role specific to regulating genes under specific environmental stimuli, and has not been implicated as a major player in the developmental cycle. The role of YebC in chlamydia has not yet been investigated, and little can be derived about its function in these organisms. This gene is upregulated late in the *Chlamydia* developmental cycle (Nicholson et al. 2003), suggesting that it may function in processes involved in the conversion of RB to EB.

Loss of σ^{28} reveals plasticity in gene regulation

Sigma factors allow the differential binding of RNA polymerase to the promoter region of genes and thus are transcriptional regulators. Within the *Chlamydiae* only three sigma factors have been identified: the primary σ^{66}, the alternative σ^{54}, and a minor σ^{28} (Tan 2012). Thus far, the role ascribed to σ^{28} is a temporal regulator of late gene expression (Yu et al. 2006). Intriguingly, only members of the *Chlamydiaceae* contain the minor σ^{28}. The absence of σ^{28} in the environmental chlamydia is perplexing as several of the genes shown to be regulated by this protein in the *Chlamydiaceae* are still present in these organisms.

Several lines of evidence indicate that σ^{28} was lost in environmental chlamydia rather than acquired by the *Chlamydiaceae*. Phylogenetic analysis suggests that σ^{28} in the *Chlamydiaceae* was vertically inherited from the last common ancestor with the *Verrucomicrobia* (Figure 3), which is in line with our current understanding of the evolution of the *Planctomycetes-Verrucomicrobia-Chlamydiae* (PVC) superphylum (Wagner and Horn 2006; Kamneva et al. 2012). We exhaustively searched in the intergenic regions of the environmental chlamydia genomes to detect any remaining fragments of σ^{28}, but it

seems that such an ancient loss has left little remnants. Secondly, as it has been proposed that σ^{28} may be subject to regulation via a partner-switching mechanism (Hua et al. 2006), we find several members of this pathway, such as RsbV2, RsbW, and RsbU2, present in environmental chlamydial genomes. The losses of the other members of this regulatory cascade, such as RsbU and RsbV, which are conserved in the *Chlamydiaceae* (Hua et al. 2006), reflect that these proteins were likely no longer needed once σ^{28} was lost. The retention of the other Rsb proteins might suggest that these proteins were recruited in a regulatory cascade for one of the two remaining sigma factors present in the environmental chlamydia.

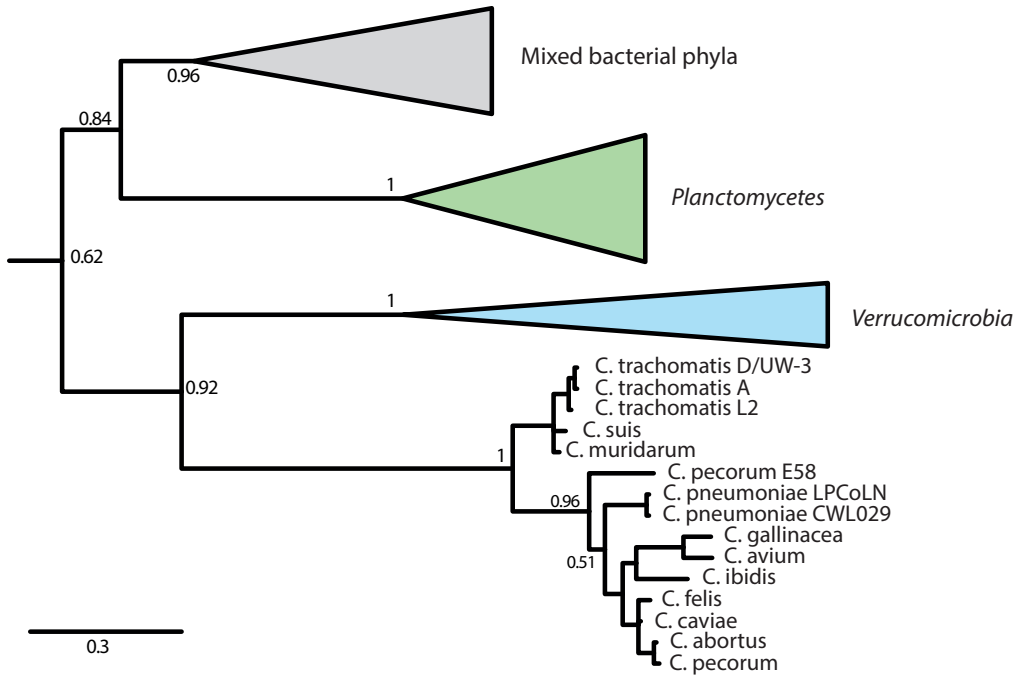


Figure 3. Loss of sigma factorσ^{28} in environmental chlamydia. Phylogenetic analysis under the GTR model in PhyloBayes supports a scenario in which the *Chlamydiaceae* σ^{28} was inherited from the *Verrucomicrobia* (PP = 0.92), and an ancient loss of this protein occurred in the environmental chlamydia. The topology of the gene tree for the *Chlamydiaceae* is largely congruent with the species phylogeny. Arrow indicates outgroup consisting of various representatives of different bacterial phyla.

The loss of σ^{28} would have prompted major changes in the transcriptional regulatory network within all environmental chlamydia. At the sequence level, these changes may be borne out as losses of transcriptional regulatory binding sites for genes once under the control of a σ^{28}. Since functional elements, such as transcription factor binding sites, tend to be under selective constraint, these elements may be highly conserved throughout evolution (Molina and Nimwegen 2008). Thus, one can scan the orthologous promoter regions in multiple species to find conserved sequence motifs, an approach called phylogenetic footprinting (Cliften et al. 2003; Katara et al. 2011). In this vein, the tail-specific protease, Tsp, is a σ^{28}-regulated gene expressed late in the *Chlamydiaceae* (Lad et al. 2007), but is present also in all environmental chlamydia genomes. When we look at the phylogenetic footprint in the promoter region of this gene, there is a significant motif found through all *Chlamydiaceae*, as would be expected (Figure 4A).

There is a notable absence of a motif shared between any members of the environmental chlamydia, suggesting major independent sequence evolution has occurred in these promoter regions. This is in stark contrast with the promoter region of the globally conserved heat shock response regulator, HrcA, which is known to self-regulate its own expression in addition to the other heat shock response genes in *Chlamydiaceae* (Wilson and Tan 2004; Wilson et al. 2005). Here, we find conserved motifs found throughout all members of the phylum *Chlamydiae* (Figure 4B), suggestive that this gene is regulated in the same manner throughout all members. Thus, the key question becomes: how are the σ^{28}-regulated genes in the *Chlamydiaceae* that are present in the environmental chlamydia regulated? Given that we cannot detect any significant conserved motifs in the promoters of these genes amongst the environmental chlamydia, this suggests a loss of strict regulation. Indeed, when we examine the preliminary transcriptome of *Protochlamydia amoebophila* (König et al., in prep.), we find that all of these "σ^{28}-late genes" are now constitutively expressed throughout the developmental cycle, again suggesting a loss of the temporal regulation of these genes as seen in the *Chlamydiaceae*.

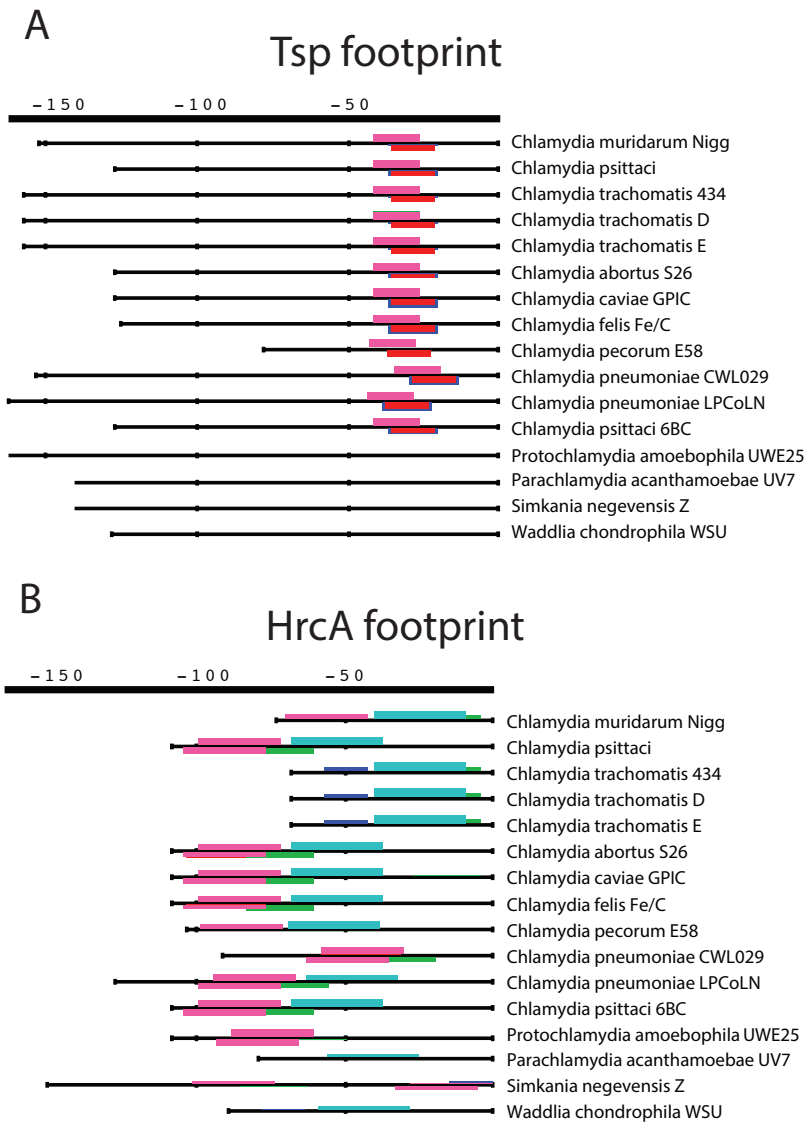


Figure 4. Phylogenetic footprints for *tsp* and *hrcA*. The conserved over-represented DNA motifs detected for (A) *tsp* and (B) *hrcA* are shown as the output of the program *matrix-scan* from the RSAT package (Thomas-Chollier et al. 2008). The colored boxes in the promoter regions indicate discovered motifs, and the height represents the statistical significance score. The promoters for each organism correspond to the direct orthologous promoter regions found in each species for either *tsp* or *hrcA*. The *hrcA* (B) promoter contains a well-conserved motif, which is found throughout the chlamydial phylum.

Transcription of *tsp* in the *Chlamydiaceae* is mediated by σ^{28} and we detect a well-defined motif (A) among these organisms. However, the loss of σ^{28} in environmental chlamydia is matched with the loss of this binding site for these organisms.

Evolutionary dynamics within chlamydial regulatory networks

To investigate the evolution of regulatory networks within the *Chlamydiae* we used a combinatorial approach of comparative genomics and existing transcriptomics data from various chlamydial organisms. Our approach focused around phylogenetic footprinting, and using the approach from (Brohee et al. 2011) we linked genes together that share similar footprints to construct predicted co-regulatory networks for each of the fully sequenced chlamydial genomes (n=17). We then used the transcriptomic studies from *C. trachomatis* (Belland et al. 2003; Nicholson et al. 2003) and *C. pneumoniae* (Mäurer et al. 2007; Albrecht et al. 2011) to further corroborate predicted regulatory schemes. This sequence based approach was shown to infer co-regulation networks just as well as microarray derived networks in yeast (Brohee et al. 2011), and thus can serve to elucidate regulatory schemes for non-model organisms.

Table 1. Properties of predicted co-regulatory networks

Network	DPbits score	Nodes	Edges	Avg # of neighbors
C. trachomatis	1 5	644 (422) 233 (89)	3968 (3617) 656 (468)	12.3 (17.1) 5.6 (10.5)
C. pneumoniae	1 5	733 (558) 286 (86)	4523 (4238) 679 (412)	12.3 (15.1) 4.74 (9.6)
P. amoebophila	1 3	710 (450) 350 (143)	2261 (1889) 627 (368)	6.3 (8.4) 3.6 (5.1)
S. negevensis	1 3	734 (266) 314 (100)	1668 (1060) 621 (250)	4.6 (7.9) 3.9 (5.0)
Chlamydiaceae	1	443 (134)	975 (581)	4.4 (8.7)
"Chlamydia" clade	5	194 (71)	488 (325)	5.0 (9.2)
"Chlamydophila" clade	5	230 (50)	346 (133)	3.0 (5.3)
environmental chlamydia	1	165 (NA)	169 (NA)	2.1 (NA)
***Chlamydiae* phylum**	1	122 (NA)	115 (NA)	1.9 (NA)

* Numbers inside parenthesis refer to counts within the large interconnected sub-network

Genes, i.e. nodes, are incorporated into the predicted co-regulation network if they have a significant motif shared in the promoter region with other genes or are predicted to be in an operon, where the edges are weighted by the strength of the similarity between these motifs, called the DPbits score (Brohee et al. 2011). Thus connections between nodes suggests that the respective genes are regulated in the same fashion, and large regulons would appear as highly connected subnetworks. Out of the 874 genes present in *C. trachomatis* 434/Bu, 644, or 76%, are represented in the inferred co-regulation network (Table 1). Similarly, the *C. pneumoniae* CWLO29 network comprises 733 genes (70%) out of the total 1,052 total genes (Figure 5A, Table 1). With this method, we detect sub-networks of previously well-defined regulons, such as the HrcA regulon. This sub-network in *C. pneumoniae* is comprised of well-characterized members of the regulon such as *hrcA*, *dnaK*, *groEL*, *groES*, and *grpE* (Wilson and Tan 2004; Wilson et al. 2005)(Figure 5B). Intriguingly, two genes (*phoH* and CPn0105/CT_016) are strongly predicted to be co-regulated within the HrcA regulon, possibly representing additional members within this regulatory network. These novel members are conserved in the predicted networks across the phylum and contain nearly perfect motifs matching the described CIRCE (Figure S3) element recognized by chlamydial HrcA (Wilson and Tan 2004) , strengthening the argument that these are likely part of this regulon. The EUO regulon currently consists of 15 members (Rosario and Tan 2012; Rosario et al. 2014), 10 of which are integrated into the networks (Figure S4). Six of these members, including *ltuB, omcA, hctB,* and *scc2*, have direct links to each other in the network, but all members are part of a tightly linked sub-network that is strongly indicative of a regulon (Figure S4). We additionally investigated if the known five members of the ChxR regulon (Koo et al. 2006; Hickey et al. 2011) were present in the network. Indeed, four members of the ChxR regulon are present, including *chxR*, *tufA*, *infA*, and CT_084 (Figure S4). Therefore, of the three defined regulons described for chlamydial organisms, we correctly predict that many of the respective members are co-regulated with each other.

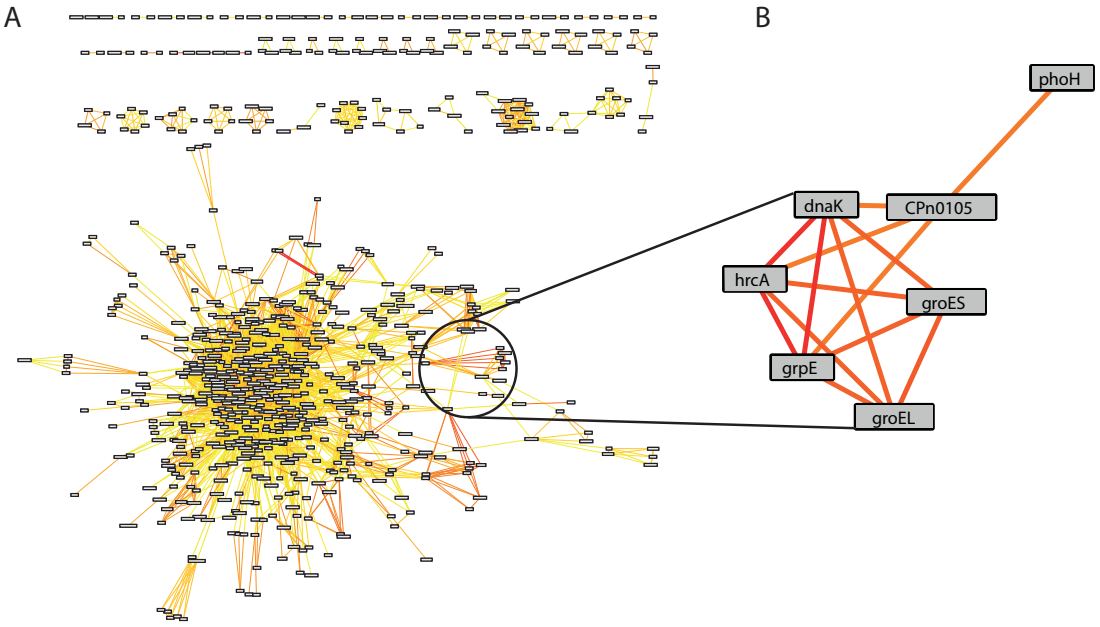


Figure 5. Predicted co-regulatory network and HrcA regulon for *C. pneumoniae*.
The predicted co-regulatory network (A) for *C. pneumonaie* CWL029 is shown where nodes in the network represent genes, and edges are predictions of genes to be co-regulated based on the similarity of phylogenetic footprints between genes. The color of the edges scale with the strength of a prediction, where dark red represents genes strongly predicted to be co-regulated and yellow for weaker predictions. The large interconnected set of genes contains many type III effector proteins and virulence-associated genes. Many of the sub-networks outside of this large "hairball" represent predicted operons. The HrcA regulon (B) is shown as an example of genes that are strongly predicted to be co-regulated.

The deep RNA-sequencing study of *C. pneumoniae* revealed that 70% of all genes detected could be affiliated with an operon (Albrecht et al. 2011), and thus operon prediction should be taken into consideration for network construction. The approach we used to infer operons (Thomas-Chollier et al. 2008) is based on a simple distance metric (default of 55 base pairs) of genes oriented in the same direction, which has been reported to be ~80% accurate (Janky and Helden 2008). To ensure that our operon prediction was reasonable,

we compared these predictions to that of the DOOR 2.0 database (Mao et al. 2014). Here operon prediction is based on a sophisticated algorithm considering a number of additional parameters, which correctly predicted 78.6% of the operons identified from the *C. pneumoniae* RNA-sequencing experiment (Albrecht et al. 2011), and was shown to be the best overall operon prediction software (Brouwer et al. 2008). We find excellent agreement between these two methods, as the percentage of RSAT operon predictions that were also predicted by the DOOR database was 98% for *C. trachomatis* 434/Bu, 98% for *C. pneumoniae* CWLO29, and 96% for *Protochlamydia amoebophila* UWE25 (Supplementary table S2). Therefore, the operon prediction applied is quite accurate for the dataset and – as it accounts for roughly 50% of the genes present in the networks (Supplementary table S2) – important for the correct prediction of co-regulated genes.

Untangling the hairball of virulence genes

One of the major goals of chlamydia research, both at a clinical and basic research level, is to identify those proteins translocated into the host cell in order to manipulate the host and subvert resources to the chlamydia. This is primarily achieved by the type III secretion system (Peters et al. 2007; Beeckman and Vanrompay 2010; Betts-Hampikian and Fields 2010; Mueller et al. 2014) . Quite strikingly, when we reduce the individual networks to only those edges that are the strongest predictions for being co-regulated (DPbits score >= 7), a sub-network of primarily virulence genes is preserved. In *C. trachomatis* 434/Bu, this network consists of 27 genes, 16 of which are either known type III secreted effector proteins or membrane proteins (Figure 6).

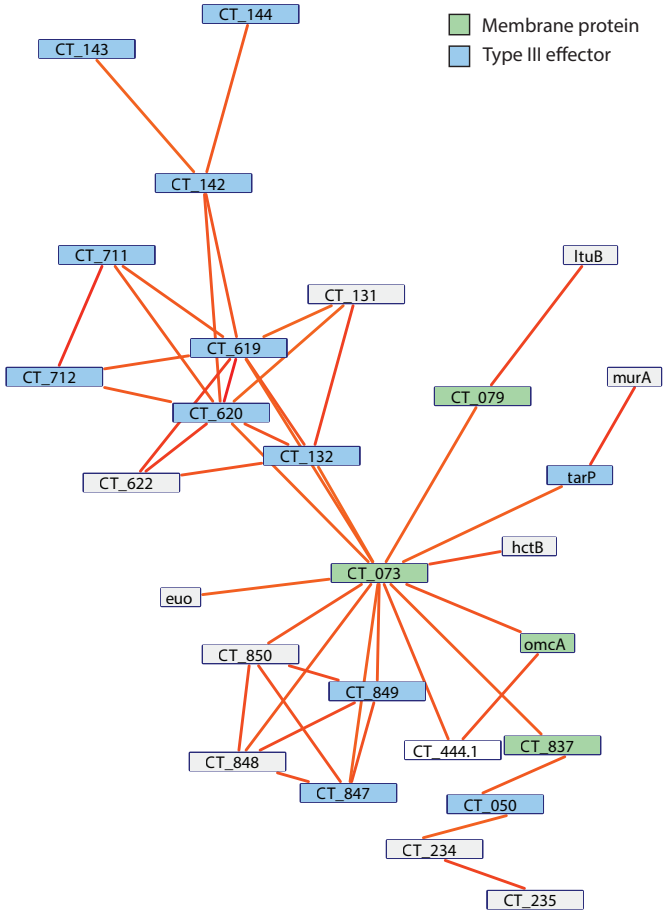


Figure 6. Putative virulence regulon of *C. trachomatis*. When we filter the predicted co-regulation network to only include those edges that are the most strongly predicted (DPbits score >= 7), a tightly connected sub-network appears, consisting mainly of virulence-associated genes. Of the 29 nodes in the sub-network, ten have been described as Type III effector proteins and five described as membrane proteins. Nodes within this sub-network likely represent strong candidates for a role in facilitating host-microbe interactions.

These gene include the actin modulating effector TarP (CT_456) (Clifton et al. 2004); the family of DUF582 proteins recently reported to be effectors (CT_620, CT_711, CT_712) (Muschiol et al. 2011); a protein that interacts with the host cell cycle regulator GCIP

(CT_847) (Chellas-Géry et al. 2007); the Pmp-like secreted protein CT_050 (Jorgensen and Valdivia 2008); CT_132 and CT_142-144, all of which either demonstrated translocation by a surrogate type III system or were computationally predicted to be effectors (da Cunha et al. 2014). In addition to these type III effectors, we find several genes encoding membrane proteins, such as a predicted membrane protein OMC1 (CT_073), the predicted virulence-associated inclusion membrane protein (Inc) CT_837 (Dehoux et al. 2011), and the cysteine-rich outer membrane protein OmcA (Everett and Hatch 1995). Given the inherent difficulties surrounding the prediction of type III effectors, it is rather striking that our approach uncovered such a highly connected network of virulence genes. By relaxing the threshold (DPbits score >= 5) this sub-network expands to 89 nodes (Figure S5), which includes other known type III secreted effectors such as GlgC (Ball et al. 2013) CT_365, CT_620, and CT_695 (Muschiol et al. 2011), and several other putative Inc proteins (CT_005, CT_814) (Dehoux et al. 2011). These top-scoring predictions connecting virulence genes holds over all organisms investigated (Figure S6).

The topologies in these networks may indicate differential regulation between effector protein sets. For example, in the *C. trachomatis* sub-network, there appear to be two main cliques: one containing the membrane proteins and those proteins connected to CT_073, and the other with those proteins clustering around CT_619 and CT_620 (Figure 6). These two cliques may represent two different sets of effectors, regulated at different times, by different factors, or for different functions. For instance, TarP and CT_849 represent effector proteins that are involved in initial inclusion formation and modification (Valdivia 2008), whereas the family of DUF582 proteins (CT_620, CT_712) in the other clique have been proposed to function in mid/late stages of the developmental cycle with a possible role facilitating exit from the host cells (Muschiol et al. 2011).

Conservation of co-regulatory networks across the phylum

Key questions we can ask using the individual predicted co-regulatory networks are how similar are they across differing taxonomic levels. If a particular prediction was conserved throughout these organisms, we would consistently recover these edges in the individual networks, and thus they would be present in a consensus network. Indeed, within-family

comparisons revealed many shared co-regulated genes for the *Chlamydiaceae*, for instance *C. trachomatis* and *C. pneumoniae* share 556 nodes out of 820 orthologs (Table 1). Among six members of the *Chlamydiaceae*, 443 nodes are present in the consensus network (Table 1, Figure S7). If we construct a consensus network where we only consider edges from individual networks if they are top predictions (DPbits >= 5), we uncover certain network properties that are different between two groups within the *Chlamydiaceae*, represented by *C. trachomatis, C. muridarum*, and *C. suis* ("Chlamydia" clade), and the group containing *C. pneumoniae* and relatives ("Chlamydophila" clade, previously classified as a separate genus; (Stephens et al. 2009) . Despite the "Chlamydophila" clade having more total nodes in the networks (230 to 194), the large putative virulence regulon (i.e. the main network) is comprised of fewer members (50 to 71) and has far fewer edges than that of the "Chlamydia" clade network (133 to 325). Another parameter we can assess between these networks is the average number of neighbors a node has, which is 5.3 in the "Chlamydophila" and 9.15 for the "Chlamydia", again confirming a higher degree of conservation for the "Chlamydia" clade. This disparity suggests that the regulatory network controlling the genes in the "Chlamydia" clade is more conserved than that of the former "Chlamydophila" clade. The members of the "Chlamydophila" have a wider breadth of hosts than that of the "Chlamydia" clade, and thus this difference might be borne out here in that certain members of the "Chlamydophila" clade have more specialized network architecture. In summary, the consensus network of the *Chlamydiaceae* is still similar to the individual organisms' networks. The presence and retention of many of the genes of the "virulence"-subnetwork indicates family- and genus-specific regulons involved in host manipulation.

When we ask which predicted interactions are conserved across the whole phylum (*C. trachomatis*, *C. pneumoniae*, and all environmental chlamydia; n=6), 122 genes remain in the network. Most of these represent genes that are in predicted operons whose gene order has been preserved, such as the ribosomal proteins, the ATP-synthase subunits, cell wall components, and type III secretion machinery. The highly conserved HrcA regulon, whose various members are not in an operon, was recovered, serving almost as an internal positive control for this analysis. If we relax our stringency on conservation to an edge

being present in 5 of the 6 genomes analyzed, the consensus network doubles in size, to 240 genes (Figure 7A). Here we recover a tightly connected sub-network of 49 genes primarily involved in host manipulation and acquisition/processing of nutrients (Figure 7B). This includes the ATP/ADP translocase 1 (*tlcA*), adenylate kinase (*adk*), the well-studied type II effector CPAF (Zhong et al. 2001) , glycogen metabolism genes (*glgC*, *glgX*), and several other genes involved in nucleotide metabolism (*dut*, *surE, pyrG)*. This consensus network is notably void of the virulence associated genes found in the stringently filtered individual networks. The phylum-wide sub-network (Figure 7B) is enriched in eggNOG functional categories in metabolism compared to the equivalent (i. e. the "hairball") in *C. trachomatis* (44% to 26%, respectively; Figure S8).

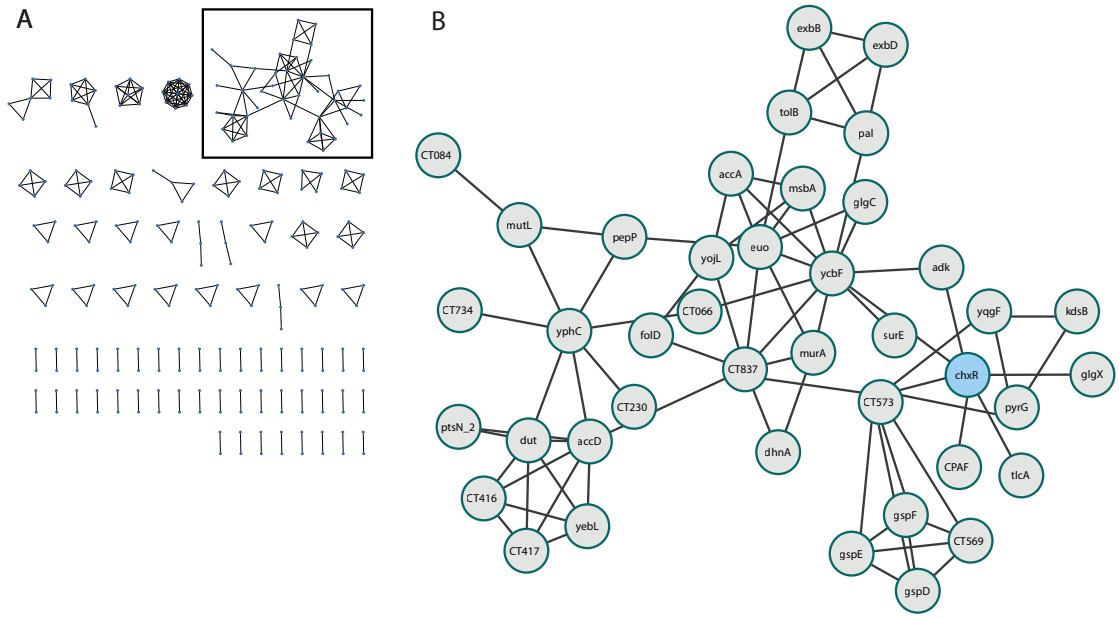


Figure 7. Phylum wide conservation of predicted co-regulations in the *Chlamydiae*. The consensus network was created by comparison of individual organism's predicted co-regulation networks (n=6). An edge was kept if it was present in five or more networks. The 240 genes present in this network (A) mainly represent conserved operons, with the exception of a putative regulon of 41 genes (B) that mainly have function in exploitation of the intracellular niche. The activator of this sub-network may be ChxR, which is indicated in blue. Notably, the virulence genes detected in individual organism's networks are absent in this consensus network.

This conserved sub-network thus seems to comprise genes needed for exploitation of the intracellular niche and nutrient metabolism. Unlike the establishment of infection, which may require highly specialized effectors for distinct host species, once inside a eukaryotic cell it seems the ways to exploit this niche by chlamydial species are rather conserved.

Intriguingly, ChxR and EUO, two of the conserved and previously described chlamydial transcriptional regulators, are highly integrated in this network (Figure 7B). Given that ChxR is known to autoregulate its own expression, it is tempting to suggest that those promoters linked with ChxR in this regulon may be under its control. Of the conserved predictions of co-regulation with ChxR we find several genes associated with type II secretion, including the type II secreted effector CPAF and the type II secretion machinery operon (*gspDEF* and the conserved hypothetical protein CT_573). Although there is debate as to the biological function CPAF serves in chlamydial infections (Chen et al. 2012), it is intriguing that we uncover the conservation of predictions involving ChxR with a type II secreted substrate and the type II secretion system, both considered mid-cycle genes. ChxR is also connected to ATP/ADP translocase (*tlcA*) and the adenylate kinase (*adk*), both of which are involved in nucleotide metabolism and appear to be mid-cycle genes (Belland et al. 2003). The interconnectedness of all nodes in the conserved phylum-wide network suggests that, indeed, these genes all have shared motifs and may be under the control of the same regulator. The presence of one of the experimentally demonstrated targets of ChxR, CT_084 (Koo et al. 2006), in the network offers more evidence that this may represent genes under the control of this activator. As it has been proposed that ChxR may function as the activator of mid-cycle genes (Koo et al. 2006), our networks support this notion, and suggest that ChxR may be an even more important regulator of global gene expression than previously thought.

Conclusions

Here, we have investigated the evolution and diversity of the transcriptional regulatory architecture at a phylum wide level. We systematically identified putative transcription factors and demonstrated that there have been extensive gains and losses of these factors during chlamydial evolution. The conserved regulatory players, especially EUO and ChxR,

likely play fundamental roles in regulating gene expression, as demonstrated by their conservation across the phylum and their central placement within our regulatory networks. Further investigations, for instance by CHiP-Seq of various transcription factors, within and between chlamydial organisms, will allow us to fully characterize the regulatory schemes present in the phylum. As we work towards this goal, the comparative genomics approach we implemented here remains a powerful tool to explore this component of evolution that would otherwise remain vastly unexplored. In this vein, we provide the first description of regulatory networks for members of the *Chlamydiae*, including those with direct relevance for human health. Our analysis revealed that major players involved in host-cell manipulation and virulence are co-regulated and are largely genus and family specific in their network organization. Additionally, we uncovered that the regulatory network architecture is not well conserved throughout the phylum, but those connections that are conserved are primarily involved in the exploitation of the intracellular niche, such as nucleotide and ATP scavenging. An invaluable corollary of this network approach is that genes integrated into these networks represent prime candidates as novel virulence-associated genes, and provide the chlamydial research community a solid starting point for investigating the roles of hypothetical proteins. This approach can easily be expanded to other non-model systems to elucidate putative functions for hypothetical proteins and determination of virulence factors (Brohee et al. 2011).

Materials and Methods

Identification of conserved transcription factors

The proteome of each organism was scanned using InterProScan v5 (Jones et al. 2014), and hits matching the DNA-binding domain families from the curated DBD database (Wilson et al. 2008) to PFAM (Finn et al. 2013) and SUPERFAMILY (Gough et al. 2001) were extracted and further curated to remove false-positive matches. We additionally searched for GO terms associated with gene regulation and those previously described in the literature. Orthologous groups of proteins were determined by OrthoMCL (Li et al. 2003) using default parameters. We inferred the evolutionary history of the transcription factors along the species tree using COUNT with the Dollo parsimony option (Csűös 2010).

Chlamydial species phylogeny

Using AMPHORA2 (Wu and Scott 2012) we extracted 31 phylogenetic marker genes from each chlamydial proteome. Each gene family was aligned with MAFFT (Katoh and Standley 2013) using the LINSI algorithm, followed by removal of poorly aligned sites using BMGE (Criscuolo and Gribaldo 2010). The individual alignments were concatenated together using SCaFoS (Roure et al. 2007). Phylogenetic analysis was performed using the CAT-GTR model in PhyloBayes-MPI (Lartillot et al. 2013) running two independent chains. We determined the chains had converged when the maximum discrepancies in bipartition frequencies (bpcomp) dropped below 0.1 and effective sampling size of parameters (tracecomp) was at least 100 between the chains, as per the recommendation in the PhyloBayes manual (Lartillot et al. 2009). We additionally performed a maximum likelihood analysis using RAxML (Stamatakis 2006) under the "PROTGAMMALGF" model with 1000 bootstraps. The PhyloBayes and RAxML tree topologies were nearly congruent.

Gene family phylogenies

EUO protein sequences were obtained from the OrthoMCL data. Protein sequences for the *lysR (pecT)* and σ^{28} analysis were obtained via BLAST against the UniRef 90 database (Suzek et al. 2007). To account for compositional heterogeneity between species in the σ^{28} analysis, we recoded the alignment into the six DayHoff categories. All alignments were

performed with MAFFT (Katoh and Standley 2013) using the LINSI algorithm, followed by removal of poorly aligned sites using BMGE (Criscuolo and Gribaldo 2010).We reconstructed the EUO gene family phylogenetic tree using RAxML (Stamatakis 2006) under the "PROTGAMMALGF" model with 1000 bootstraps. Phylogenetic trees for LysR and σ^{28} were calculated with PhyloBayes under the GTR model and convergence checks were performed the same as in the species tree analysis.

Phylogenetic footprinting and predicted co-regulatory networks

We used the Regulatory Sequence Analysis Tools (RSAT) (Thomas-Chollier et al. 2008) suite to construct both the phylogenetic footprints and predicted co-regulatory networks. Briefly, for each organisms considered, we determined if there was a significant phylogenetic footprint for each gene in that genome by detecting over-represented motifs in promoter regions via the program *dyad-analysis* (Defrance et al. 2008; Janky and Helden 2008). We then created the co-regulation networks in RSAT by linking similar phylogenetic footprints together as previously described (Brohee et al. 2011) via the *footprint-discovery* program within RSAT with the following default parameters: "-lth occ 1 -lth occ_sig 0 -uth rank 50 -bg_model taxfreq -all_genes -sep_genes -filter -infer_operons -task all". Networks were also constructed without predicting operons by omitting the "-infer_operons" option. Consensus networks were created using a custom Python script using the 'NetworkX' package. Networks were viewed and processed using Cytoscape v.3.2.1 (Shannon et al. 2003). Sequence logos were created using WebLogo 3 (Crooks et al. 2004). The MEME software (Bailey et al. 2009) was also used to scan sequence groups, such as the putative EUO regulon for over-represented motifs. Operon predictions were downloaded from the DOOR 2.0 database (Mao et al. 2014) for comparison against the RSAT operon predictions. Functional categories were assigned to genes via BLAST of the eggNOG 4.0 database (Powell et al. 2014). All networks may be downloaded as GML files from figshare.com/s/0b0b4ebe046a11e59e9c06ec4bbcf141 .

Acknowledgement

This work was supported by the Austrian Science Fund FWF grant I1628-B22, the European Research Council StG EVOCHLAMY (281633), and the Marie Curie Initial Training Network Symbiomics.

References

Akers JC, HoDac H, Lathrop RH, Tan M. 2011. Identification and functional analysis of CT069 as a novel transcriptional regulator in *Chlamydia*. J. Bacteriol. 193:6123–6131.

Albrecht M, Sharma CM, Dittrich MT, Müller T, Reinhardt R, Vogel J, Rudel T. 2011. The transcriptional landscape of *Chlamydia pneumoniae*. Genome Biol. 12:R98.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME Suite: tools for motif discovery and searching. Nucleic Acids Res. 37:W202–W208.

Ball SG, Subtil A, Bhattacharya D, Moustafa A, Weber APM, Gehre L, Colleoni C, Arias M-C, Cenci U, Dauvillée D. 2013. Metabolic effectors secreted by bacterial pathogens: Essential facilitators of plastid endosymbiosis? Plant Cell Online 25:7–21.

Bao X, Nickels BE, Fan H. 2012. *Chlamydia trachomatis* protein GrgA activates transcription by contacting the nonconserved region of σ66. Proc. Natl. Acad. Sci. U. S. A. 109:16870–16875.

Beeckman DSA, Vanrompay DCG. 2010. Bacterial secretion systems with an emphasis on the chlamydial Type III secretion system. Curr. Issues Mol. Biol. 12:17–41.

Belland RJ, Zhong G, Crane DD, Hogan D, Sturdevant D, Sharma J, Beatty WL, Caldwell HD. 2003. Genomic transcriptional profiling of the developmental cycle of *Chlamydia trachomatis*. Proc. Natl. Acad. Sci. 100:8478–8483.

Betts-Hampikian HJ, Fields KA. 2010. The chlamydial Type III secretion mechanism: Revealing cracks in a tough nut. Front. Microbiol. 1.

Boyd ES, Barkay T. 2012. The mercury resistance operon: From an origin in a geothermal environment to an efficient detoxification machine. Front. Microbiol. [Internet] 3. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3466566/

Brohee S, Janky R, Abdel-Sater F, Vanderstocken G, Andre B, van Helden J. 2011. Unraveling networks of co-regulated genes on the sole basis of genome sequences. Nucleic Acids Res. 39:6340–6358.

Brouwer RWW, Kuipers OP, Hijum SAFT van. 2008. The relative value of operon predictions. Brief. Bioinform. 9:367–375.

Chellas-Géry B, Linton CN, Fields KA. 2007. Human GCIP interacts with CT847, a novel *Chlamydia trachomatis* type III secretion substrate, and is degraded in a tissue-culture infection model. Cell. Microbiol. 9:2417–2430.

Chen AL, Johnson KA, Lee JK, Sütterlin C, Tan M. 2012. CPAF: A chlamydial protease in search of an authentic substrate. PLoS Pathog 8:e1002842.

Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M. 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. Science 301:71–76.

Clifton DR, Fields KA, Grieshaber SS, Dooley CA, Fischer ER, Mead DJ, Carabeo RA, Hackstadt T. 2004. A chlamydial type III translocated protein is tyrosine-phosphorylated at the site of entry and associated with recruitment of actin. Proc. Natl. Acad. Sci. U. S. A. 101:10166–10171.

Collingro A, Tischler P, Weinmaier T, Penz T, Heinz E, Brunham RC, Read TD, Bavoil PM, Sachse K, Kahane S, et al. 2011. Unity in variety—The pan-genome of the *Chlamydiae*. Mol. Biol. Evol. 28:3253–3270.

Criscuolo A, Gribaldo S. 2010. 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.

Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: A sequence logo generator. Genome Res. 14:1188–1190.

Csűös M. 2010. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics 26:1910–1912.

Da Cunha M, Milho C, Almeida F, Pais SV, Borges V, Maurício R, Borrego MJ, Gomes JP, Mota LJ. 2014. Identification of type III secretion substrates of *Chlamydia trachomatis* using *Yersinia enterocolitica* as a heterologous system. BMC Microbiol. 14:40.

Defrance M, Janky R, Sand O, Helden J van. 2008. Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences. Nat. Protoc. 3:1589–1603.

Dehoux P, Flores R, Dauga C, Zhong G, Subtil A. 2011. Multi-genome identification and characterization of chlamydiae-specific type III secretion substrates: the Inc proteins. BMC Genomics 12:109.

Everett KD, Hatch TP. 1995. Architecture of the cell envelope of *Chlamydia psittaci* 6BC. J. Bacteriol. 177:877–882.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, et al. 2013. Pfam: the protein families database. Nucleic Acids Res. 42:D222–D230.

Gough J, Karplus K, Hughey R, Chothia C. 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J. Mol. Biol. 313:903–919.

Heuer D, Kneip C, Mäurer AP, Meyer TF. 2007. Tackling the intractable - approaching the genetics of *Chlamydiales*. Int. J. Med. Microbiol. IJMM 297:569–576.

Hickey JM, Lovell S, Battaile KP, Hu L, Middaugh CR, Hefty PS. 2011. The atypical response regulator protein ChxR has structural characteristics and dimer interface interactions that are unique within the OmpR/PhoB subfamily. J. Biol. Chem. 286:32606–32616.

Horn M. 2008. *Chlamydiae* as Symbionts in Eukaryotes. Annu. Rev. Microbiol. 62:113–131.

Hua L, Hefty PS, Lee YJ, Lee YM, Stephens RS, Price CW. 2006. Core of the partner switching signalling mechanism is conserved in the obligate intracellular pathogen *Chlamydia trachomatis*. Mol. Microbiol. 59:623–636.

Janky R, Helden J van. 2008. Evaluation of phylogenetic footprint discovery for predicting bacterial cis-regulatory elements and revealing their evolution. BMC Bioinformatics 9:37.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. 2014. InterProScan 5: genome-scale protein function classification. Bioinforma. Oxf. Engl.

Jorgensen I, Valdivia RH. 2008. Pmp-like proteins Pls1 and Pls2 are secreted into the lumen of the Chlamydia trachomatis inclusion. Infect. Immun. 76:3940–3950.

Kahane S, Metzer E, Friedman MG. 1995. Evidence that the novel microorganism "Z" may belong to a new genus in the family *Chlamydiaceae*. FEMS Microbiol. Lett. 126:203–207.

Kamneva OK, Knight SJ, Liberles DA, Ward NL. 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.

Katara P, Grover A, Sharma V. 2011. Phylogenetic footprinting: a boost for microbial regulatory genomics. Protoplasma 249:901–907.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in performance and usability. Mol. Biol. Evol. 30:772–780.

Kemege KE, Hickey JM, Lovell S, Battaile KP, Zhang Y, Hefty PS. 2011. Ab initio structural modeling of and experimental validation for *Chlamydia trachomatis* protein CT296 reveal structural similarity to Fe(II) 2-oxoglutarate-dependent enzymes. J. Bacteriol. 193:6517–6528.

Koo IC, Walthers D, Hefty PS, Kenney LJ, Stephens RS. 2006. ChxR is a transcriptional activator in *Chlamydia*. Proc. Natl. Acad. Sci. U. S. A. 103:750–755.

Lad SP, Yang G, Scott DA, Wang G, Nair P, Mathison J, Reddy VS, Li E. 2007. Chlamydial CT441 is a PDZ domain-containing tail-specific protease that interferes with the NF-κB pathway of immune response. J. Bacteriol. 189:6619–6625.

Lagkouvardos I, Weinmaier T, Lauro FM, Cavicchioli R, Rattei T, Horn M. 2014. Integrating metagenomic and amplicon databases to resolve the phylogenetic and ecological diversity of the Chlamydiae. ISME J. 8:115–125.

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

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

Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.

Madan Babu M, Teichmann SA, Aravind L. 2006. Evolutionary dynamics of prokaryotic transcriptional regulatory networks. J. Mol. Biol. 358:614–633.

Mao X, Ma Q, Zhou C, Chen X, Zhang H, Yang J, Mao F, Lai W, Xu Y. 2014. DOOR 2.0: presenting operons and their functions through dynamic and integrated views. Nucleic Acids Res. 42:D654–D659.

Mäurer AP, Mehlitz A, Mollenkopf HJ, Meyer TF. 2007. Gene expression profiles of *Chlamydophila pneumoniae* during the developmental cycle and iron depletion–mediated persistence. PLoS Pathog 3:e83.

Molina N, Nimwegen E van. 2008. Universal patterns of purifying selection at noncoding positions in bacteria. Genome Res. 18:148–160.

Moliner C, Fournier P-E, Raoult D. 2010. Genome analysis of microorganisms living in amoebae reveals a melting pot of evolution. FEMS Microbiol. Rev. 34:281–294.

Mueller KE, Plano GV, Fields KA. 2014. New frontiers in Type III secretion biology: the *Chlamydia* perspective. Infect. Immun. 82:2–9.

Muschiol S, Boncompain G, Vromman F, Dehoux P, Normark S, Henriques-Normark B, Subtil A. 2011. Identification of a family of effectors secreted by the Type III secretion system that are conserved in pathogenic *Chlamydiae*. Infect. Immun. 79:571–580.

Nguyen BD, Valdivia RH. 2013. Forward genetic approaches in *Chlamydia trachomatis*. J. Vis. Exp. JoVE:e50636.

Nicholson TL, Olinger L, Chong K, Schoolnik G, Stephens RS. 2003. Global stage-specific gene regulation during the developmental cycle of *Chlamydia trachomatis*. J. Bacteriol. 185:3179–3189.

Perez JC, Groisman EA. 2009. Evolution of transcriptional regulatory circuits in bacteria. Cell 138:233–244.

Peters J, Wilson DP, Myers G, Timms P, Bavoil PM. 2007. Type III secretion à la *Chlamydia*. Trends Microbiol. 15:241–251.

Powell S, Forslund K, Szklarczyk D, Trachana K, Roth A, Huerta-Cepas J, Gabaldón T, Rattei T, Creevey C, Kuhn M, et al. 2014. eggNOG v4.0: nested orthology inference across 3686 organisms. Nucleic Acids Res. 42:D231–D239.

Price M, Dehal P, Arkin A. 2008. Horizontal gene transfer and the evolution of transcriptional regulation in *Escherichia coli*. Genome Biol. 9:R4.

Price MN, Dehal PS, Arkin AP. 2007. Orthologous Transcription factors in bacteria have different functions and regulate different genes. PLoS Comput Biol 3:e175.

Rajewsky N, Socci ND, Zapotocky M, Siggia ED. 2002. The evolution of DNA regulatory regions for Proteo-gamma bacteria by interspecies comparisons. Genome Res. 12:298–308.

Rau A, Wyllie S, Whittimore J, Raulston JE. 2005. Identification of *Chlamydia trachomatis* genomic sequences recognized by chlamydial divalent cation-dependent regulator A (DcrA). J. Bacteriol. 187:443–448.

Rosario CJ, Hanson BR, Tan M. 2014. The transcriptional repressor EUO regulates both subsets of *Chlamydia* late genes. Mol. Microbiol. 94:888–897.

Rosario CJ, Tan M. 2012. The early gene product EUO is a transcriptional repressor that selectively regulates promoters of *Chlamydia* late genes. Mol. Microbiol. 84:1097–1107.

Roure B, Rodriguez-Ezpeleta N, Philippe H. 2007. SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. BMC Evol. Biol. 7:S2.

Rurangirwa FR, Dilbeck PM, Crawford TB, McGuire TC, McElwain TF. 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. Int. J. Syst. Bacteriol. 49 Pt 2:577–581.

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

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

Stephens RS, Myers G, Eppinger M, Bavoil PM. 2009. Divergence without difference: phylogenetics and taxonomy of *Chlamydia* resolved. FEMS Immunol. Med. Microbiol. 55:115–119.

Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23:1282–1288.

Tan M. 2012. Temporal gene regulation during the chlamydial developmental cycle. In: Tan M, Bavoil PM, editors. Intracellular Pathogens I: Chlamydiales. p. 149–169.

Taylor-Brown A, Vaughan L, Greub G, Timms P, Polkinghorne A. 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*. Pathog. Dis. 73:1–15.

Thomas-Chollier M, Sand O, Turatsinze J-V, Janky R, Defrance M, Vervisch E, Brohée S, van Helden J. 2008. RSAT: regulatory sequence analysis tools. Nucleic Acids Res. 36:W119–W127.

Valdivia RH. 2008. *Chlamydia* effector proteins and new insights into chlamydial cellular microbiology. Curr. Opin. Microbiol. 11:53–59.

Wagner M, Horn M. 2006. The *Planctomycetes*, *Verrucomicrobia*, *Chlamydiae* and sister phyla comprise a superphylum with biotechnological and medical relevance. Curr. Opin. Biotechnol. 17:241–249.

Wilson AC, Tan M. 2004. Stress response gene regulation in *Chlamydia* is dependent on HrcA-CIRCE interactions. J. Bacteriol. 186:3384–3391.

Wilson AC, Wu CC, Yates JR, Tan M. 2005. Chlamydial GroEL autoregulates its own expression through direct interactions with the HrcA repressor protein. J. Bacteriol. 187:7535–7542.

Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA. 2008. DBD––taxonomically broad transcription factor predictions: new content and functionality. Nucleic Acids Res. 36:D88–D92.

Wu M, Scott AJ. 2012. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics 28:1033–1034.

Yu HHY, Kibler D, Tan M. 2006. In silico prediction and functional validation of σ28-regulated genes in *Chlamydia* and *Escherichia coli*. J. Bacteriol. 188:8206–8212.

Zhong G, Fan P, Ji H, Dong F, Huang Y. 2001. Identification of a chlamydial protease-like activity factor responsible for the degradation of host transcription factors. J. Exp. Med. 193:935–942.

Following the footsteps of chlamydial gene regulation

Domman and Horn

Supplementary Information

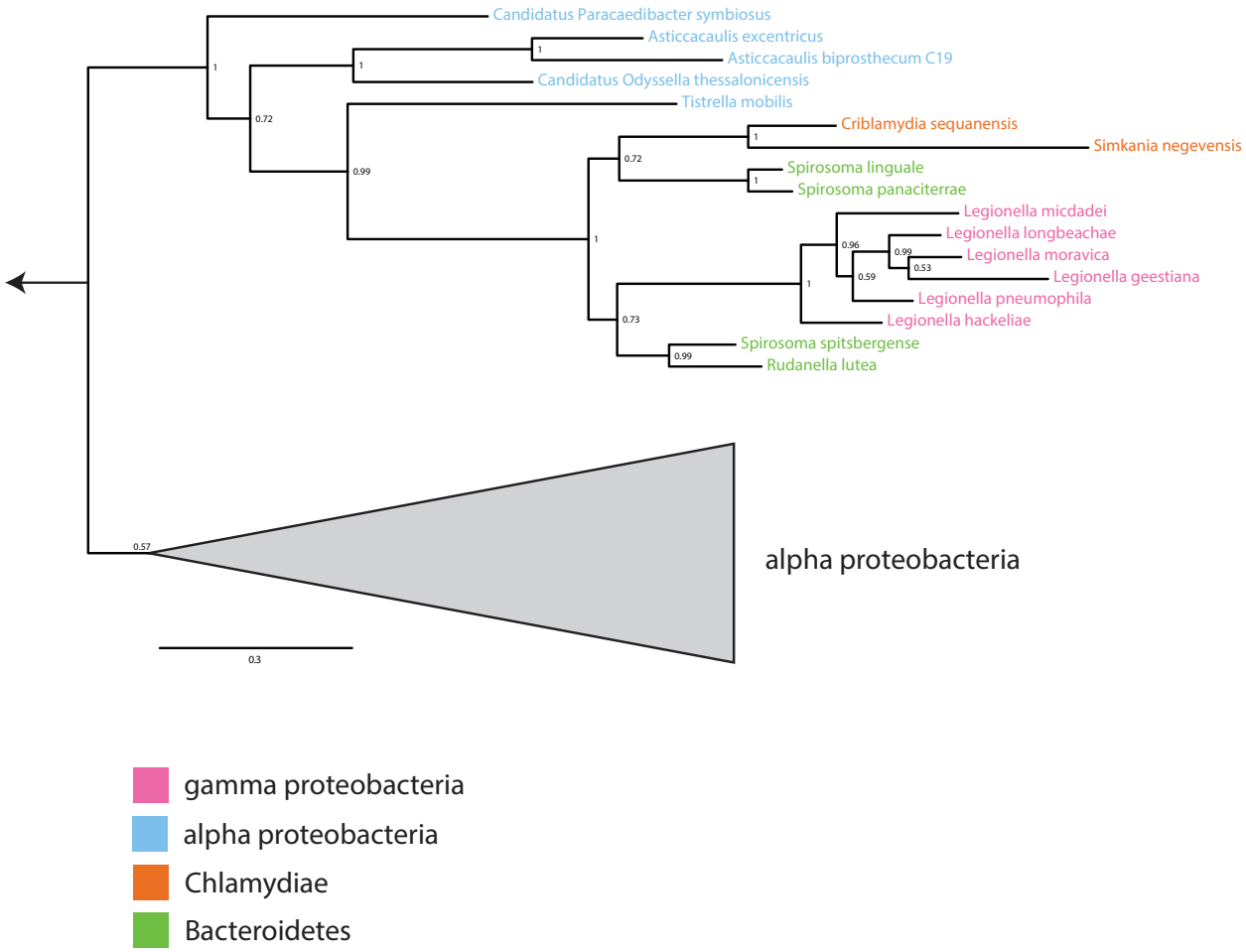
Table S1. Putative regulators within the Chlamydiae

OG	LocusTag	Gene_name	UniProtID	Description	CHLAB	CHLAIB	CHLAVI	CHLCV	CHLFF	CHLGAL	CHLMU	CHLP6	CHLPE	CHLPN	CHLPP	CHLSUIS	CHLTR	CHLT2	CHLTA	Crib	EPS4	TUME1	PARAV	OEW1	EI2	PARUW	ProNeg	Rubis	SIMNZ	WADCW	Totals
ChlamOG_171	CT_630	chaR	O84635	putative regulator of midcycle genes	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	26
ChlamOG_264	CT_069	ytgR	O84072	acts as an iron-binding repressor of diphteria toxin gene expression and may serve as a global regulator of gene expression	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	26
ChlamOG_36	CT_407	dksA/traR	O84412	zinc finger-containing members of the DksA/TraR family; dnaK suppressor	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	26
ChlamOG_509	CT_691	phoU	O84697	Putitive phosphate transport regulator	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	26
ChlamOG_567	CT_394	hrcA	P36426	Heat-inducible transcription repressor HrcA	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	26
ChlamOG_600	CT_446	euo	O84452	repressor of late genes	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	26
ChlamOG_88	CT_457	yebC/pmpR	O84463	Probable transcriptional regulatory protein	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	26
ChlamOG_14	CT_468	atoC	O84474	sigma-54 interaction	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	25
ChlamOG_659	CT_260	YbjN	O84262	putative sensory transduction regulator	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	25
ChlamOG_384	CT_406	NrdR	O84411	Transcriptional repressor NrdR	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	24
ChlamOG_1092	CT_504	grgA	O84512		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	15
ChlamOG_1084	CT_296	dcrA	O84298	Divalent cation-dependent regulator A	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	15
ChlamOG_1464	pc1704	WHTH	Q9MAH1	WHTH_DNA-bd_dom	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	11
ChlamOG_840	pc0029	FmdB	Q9MP96	Putative regulatory protein, FmdB, Zinc ribbon domain	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	10
ChlamOG_1031	pc1344	AnsC	Q9MBI1	arsC family also comprises the Spx proteins which are Gram-positive bacterial transcription factors that regulate the transcription of multiple genes in response to disulphide stress	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	0	0	0	9
ChlamOG_1849	CT_169	trpR	O84171	Trp operon repressor	0	0	0	1	1	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	1	1	0	9
ChlamOG_746	pc1269	lexA	Q9MBQ6	LexA repressor	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	0	1	9
ChlamOG_2270	CPn0193	ArgR	Q9Z8Z1	Regulates arginine biosynthesis genes	0	1	0	1	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	8
ChlamOG_2274	pCTA_0004	pGP4-D	Q3MH80	virulence gene regulator	0	0	0	1	1	0	0	1	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	8
ChlamOG_1016	PUV_06280	csoR	F8KXA6		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	2	2	0	9
ChlamOG_2102	SNE_A03110	SNE_A03110	F8L4N3		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	1	1	1	1	7
ChlamOG_2705	pc1813	pc1813	Q9MA62		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	1	1	1	0	7
ChlamOG_4177	PUV_15180	ydcH	F8KDV9		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0	1	1	1	0	5
ChlamOG_1991	PUV_18110	mntR	F8KWZ1		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	1	0	4
ChlamOG_2006	pc1801	fur	Q6MA74		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	1	4
ChlamOG_2459	PUV_02730	ylaA	F8KVT4		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	1	0	0	4
ChlamOG_4058	PUV_18090	PUV_18090	F8LDL0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	1	4
ChlamOG_4199	SNE_A08630	yobU	F8L7K7		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	1	0	4
ChlamOG_5328	pc0065	pc0065	Q6MF60		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	1	0	4
ChlamOG_5388	BN1013_00774	BN1013_00774	A0A078L4R7		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	1	1	0	4
ChlamOG_2722	SNE_A16180	SNE_A16180	F8L5I9		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	3
ChlamOG_5179	SNE_B24610	higA	F8L2X4		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	3
ChlamOG_6615	pc1472	pc1472	Q6MB53		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	3
ChlamOG_7055	BN1013_02379	BN1013_02379	A0A078LD90		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	3
ChlamOG_7103	CSEC_p0022	CSEC_p0022	A0A090D144		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	3
ChlamOG_7116	pc1922	pc1922	Q6M9V3		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	3
ChlamOG_7592	SNE_A12490	SNE_A12490	F8L8J5		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	3
ChlamOG_10148	pc1022	pc1022	Q6M868		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	2
ChlamOG_10183	pc0518	ada	Q6MDV7		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	2
ChlamOG_10288	D842_C000240	D842_C000240	A0A0C1H691		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	2
ChlamOG_2514	SNE_A13880	SNE_A13880	F8L8W8		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	2
ChlamOG_3739	CSEC_0909	CSEC_0909	A0A090C1U1		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	2
ChlamOG_4255	SNE_A14140	SNE_A14140	F8L8Y7		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	2
ChlamOG_4839	PUV_24240	PUV_24240	F8L248		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	2
ChlamOG_5420	PUV_14500	PUV_14500	F8K2P9		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	2
ChlamOG_6393	D841_C000010	D841_C000010	A0A0C1H087		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	0	3
ChlamOG_7024	SNE_A08330	sigA-A	F8L6B0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	2
ChlamOG_7336	BN1013_01721	BN1013_01721	A0A078L3R0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	2
ChlamOG_9707	PUV_06250	PUV_06250	F8KXA4		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	2
ChlamOG_6558	SNE_A16270	pecT	F8L9G7	HTH-type transcriptional regulator pecT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	2
ChlamOG_10577	SNE_A16330	padR	F8L9F7		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
ChlamOG_10588	BN1013_01811	csiR	A0A078L828	Carbon starvation induced regulator	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
ChlamOG_1669	wcw_0617	wcw_0617	D6YV24		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
ChlamOG_2194	CSEC_p0050	arsR	A0A090D3A3	Arsenical resistance operon repressor	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
ChlamOG_2474	CSEC_0873	csrA	A0A090CYR8	Putative carbon storage regulatory protein	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
ChlamOG_2564	wcw_0478	wcw_0478	D6YUN8		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
ChlamOG_27338	WP_032124883	WP_032124883	UPI0005099A10		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
ChlamOG_2814	BN1013_02431	BN1013_02431	A0A078L571	Anaerobic benzoate catabolism transcriptional regulator	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
ChlamOG_31216	BN1013_00085	BN1013_00085	A0A078R0T5		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
ChlamOG_31291	BN1013_00626	BN1013_00626	A0A078L4N8	Response regulator	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
ChlamOG_31294	BN1013_00631	BN1013_00631	A0A078L4N9	Growth regulator	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
ChlamOG_31784	SNE_A20610	SNE_A20610	F8L5P1		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
ChlamOG_31800	SNE_A16440	sigA-B	F8L9I4	SigA-B protein	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
ChlamOG_31881	SNE_B24190	SNE_B24190	F8L2T2		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
ChlamOG_34846	wcw_p0015	wcw_p0015	D6YXJ0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
ChlamOG_3662	WP_032126008	WP_032126008	UPI0005096218	Fis family transcriptional regulator	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
ChlamOG_4743	SNE_A23700	SNE_A23700	F8L671		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	4
ChlamOG_4904	BN1013_01172	slyA	A0A078LLY3	Salmolysin	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
ChlamOG_4917	SNE_A04940	mraZ	F8L6M3	Transcriptional regulator MraZ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
ChlamOG_5273	CSEC_0758	CSEC_0758	A0A090DKR8	Transcriptional regulator	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
ChlamOG_7539	SNE_B24440	merR1	F8L2V7		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
ChlamOG_8690	WP_032125098.1	WP_032125098.1	UPI0005093E1FE		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	2
ChlamOG_9631	SNE_A09120	SNE_A09120	F8L7Q1		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
Totals :					12	13	12	15	15	12	13	13	14	14	14	14	13	14	14	30	21	22	23	22	21	21	31	37	31	20	481

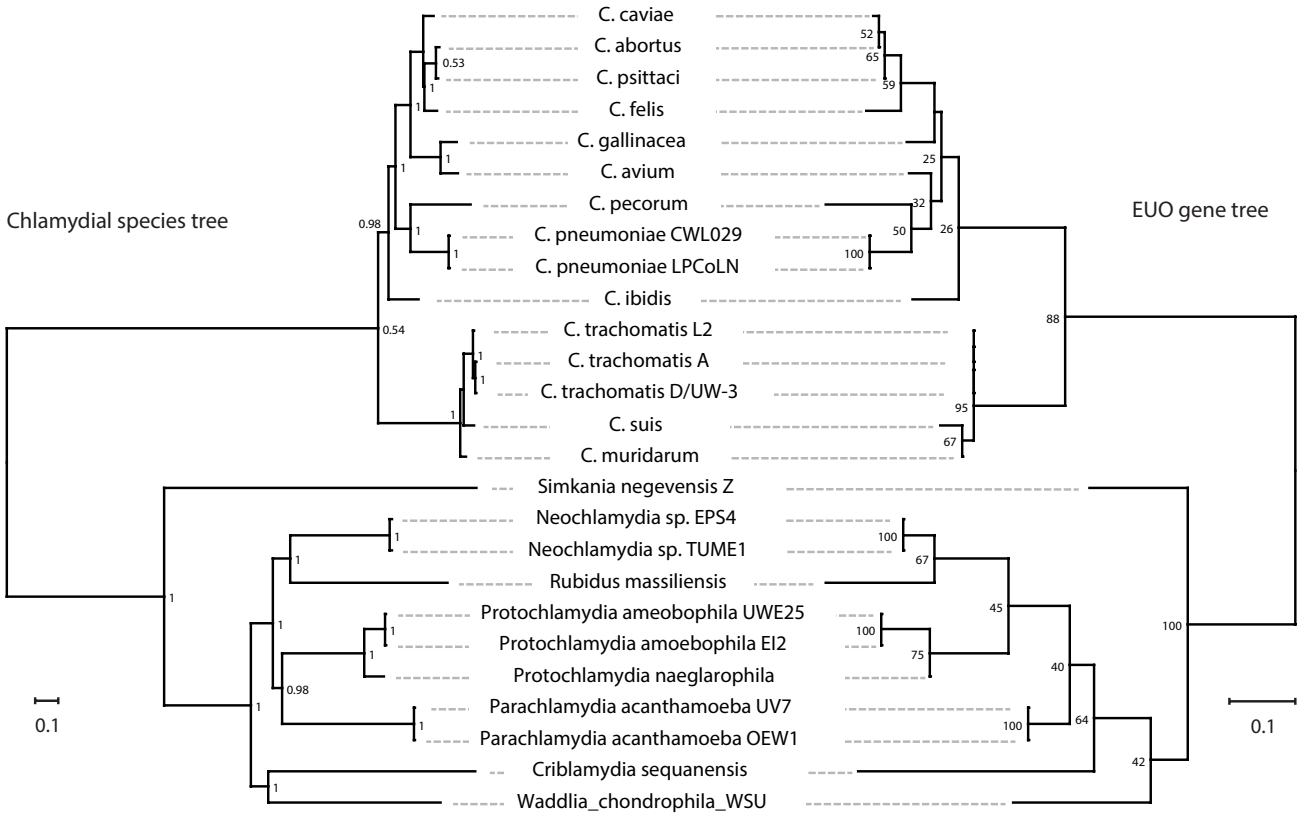
Species abbreviation	Species name
CHLAB	Chlamydia abortus S26
CHLAIB	Chlamydia ibidis
CHLAVI	Chlamydia avium 10DC88
CHLCV	Chlamydia caviae GPIC
CHLFF	Chlamydia felis Fe/C-56
CHLGAL	Chlamydia gallinacea 08-1274/3
CHLMU	Chlamydia muridarum MoPn/Nigg
CHLP6	Chlamydia psittaci ATCC VR-125/6BC
CHLPE	Chlamydia pecorum ATCC VR-628/ES8
CHLPN	Chlamydia pneumoniae CWL029
CHLPP	Chlamydia pneumoniae LPCoLN
CHLSUIS	Chlamydia suis MD56
CHLTR	Chlamydia trachomatis D/UW-3/Cx
CHLT2	Chlamydia trachomatis L2 434/Bu/ATCC VR-902B
CHLTA	Chlamydia trachomatis A ATCC VR-571B/DSM19440/HAR-13
Crib	Criblamydia sequanensis
EPS4	Neochlamydia sp EPS4
TUME1	Neochlamydia sp TUME1
PARAV	Parachlamydia acanthamoebae UV-7
OEW1	Parachlamydia acanthamoebae OEW1
EI2	Protochlamydia amoebophila EI2
PARUW	Protochlamydia amoebophila UWE25
ProNeg	Protochlamydia naegleriophila
Rubis	Rubidus massiliensis
SIMNZ	Simkania negevensis Z
WADCW	Waddlia chondrophila WSU

Table S2. **Comparison of operon predictions.**

	***Chlamydia trachomatis* 434/Bu**	***Chlamydia pneumoniae* CWL029**	***Protochlamydia amoebophila* UWE25**
Total genes	874	1033	2031
Genes in RSAT operon prediction	537	634	947
Genes in DOOR operon prediction	575	691	995
RSAT predictions in DOOR predictions	531	624	917
Genes in non-operon network	311	385	384
Genes in operon network	644	733	710



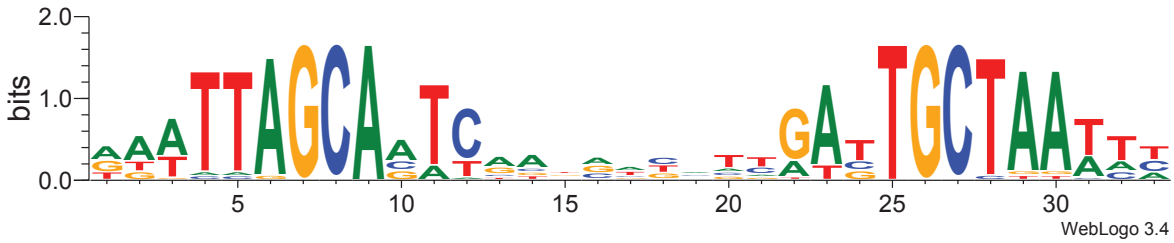
Supplementary Figure S1. **Acquisition of transcription factors for members of the *Chlamydiae*.** The transcription factor *pecT* is part of the general LysR transcriptional regulator family, and seems to have been aquired by *Simkania negevensis* (F8L9G7)and *Criblamydia sequanensis* (A0A090E2E0) from members of the *Baceteroidetes*. The gene history is complicated, as the gene was transfered first from alpha-proteobacteria into the clade with the *Chlamydiae* and *Bacteroidetes*. There was a subsequent transfer of this gene to members of the *Legionella*, which are known to infect free living ameoba.



Supplementary Figure 2. **Gene trees of global regulators, such as EUO, largely follow the species tree.** Shown is a comparison of the species phylogeny (based on 33 concatenated marker proteins) to the gene tree of the late gene regulator EUO. Congruence between species and gene trees indicates vertical inheritance of this global regulator throughout chlamydial evolution. The EUO gene tree was calculated via RAxML under the LG +gamma+F model with 1000 bootstraps.

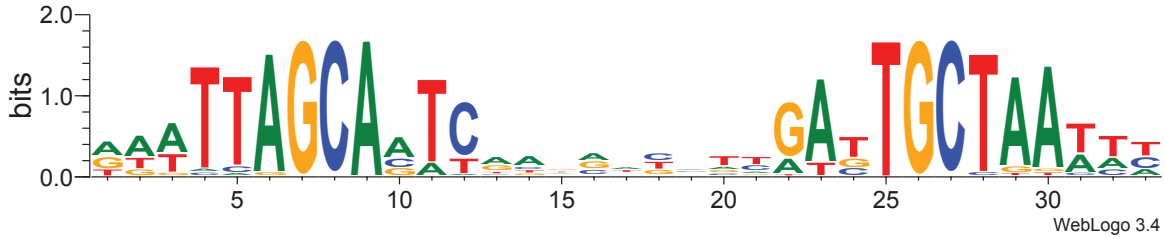
A

CT_015/CPn0106 (phoH)



B

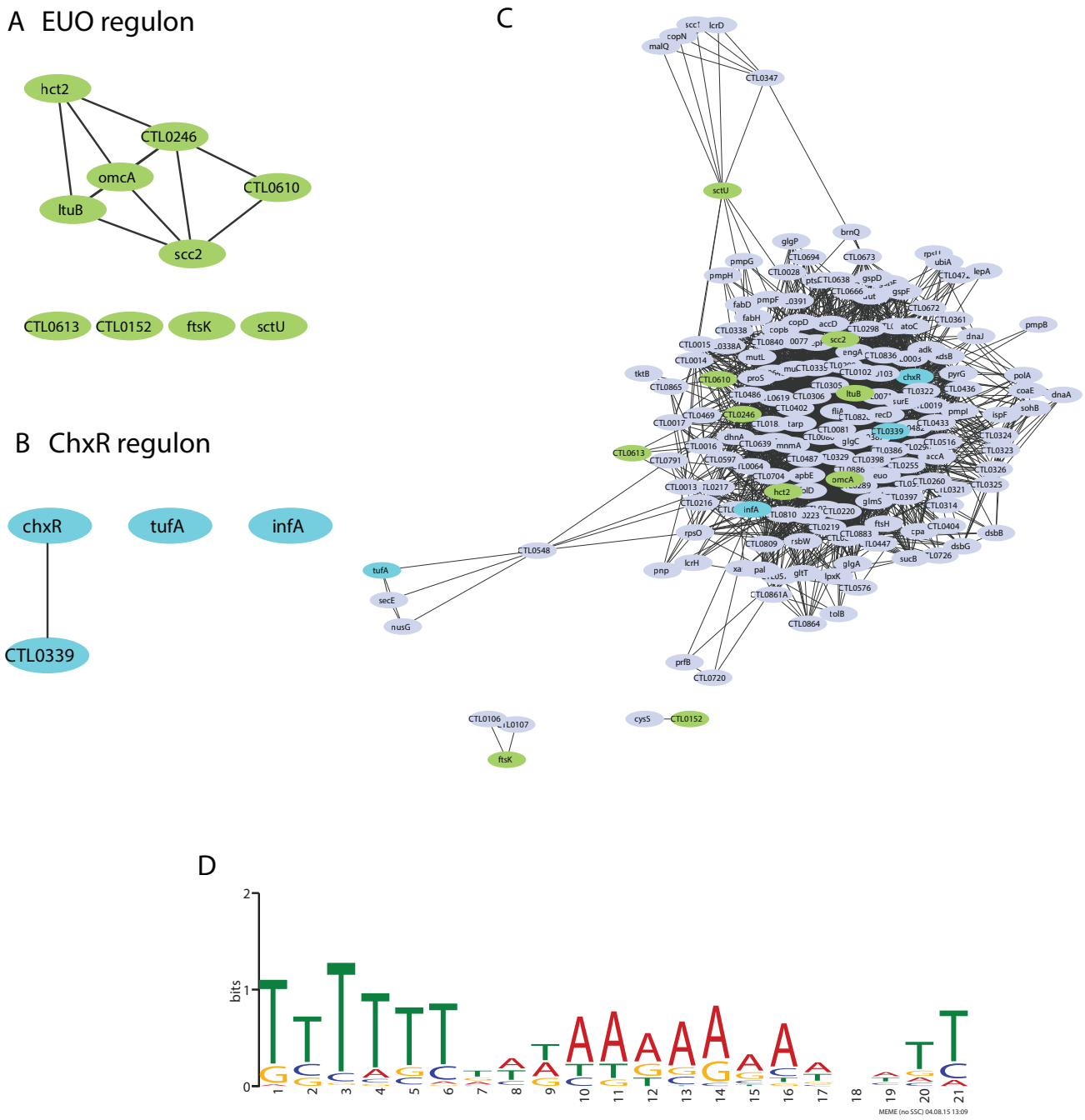
CT_016/CPn0105



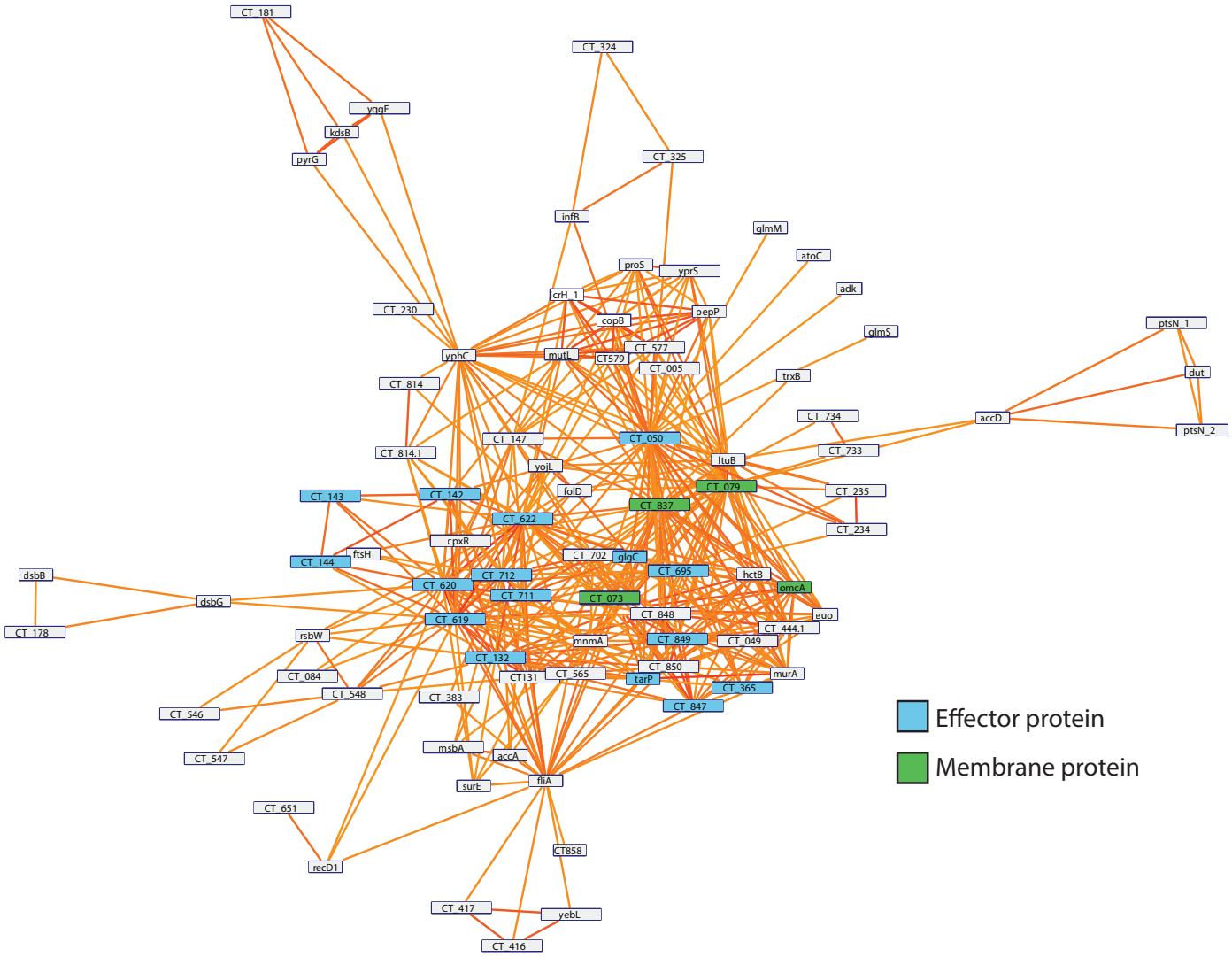
Consensus CIRCE

TTAGCACTC-N_9-GAGTGCTAA

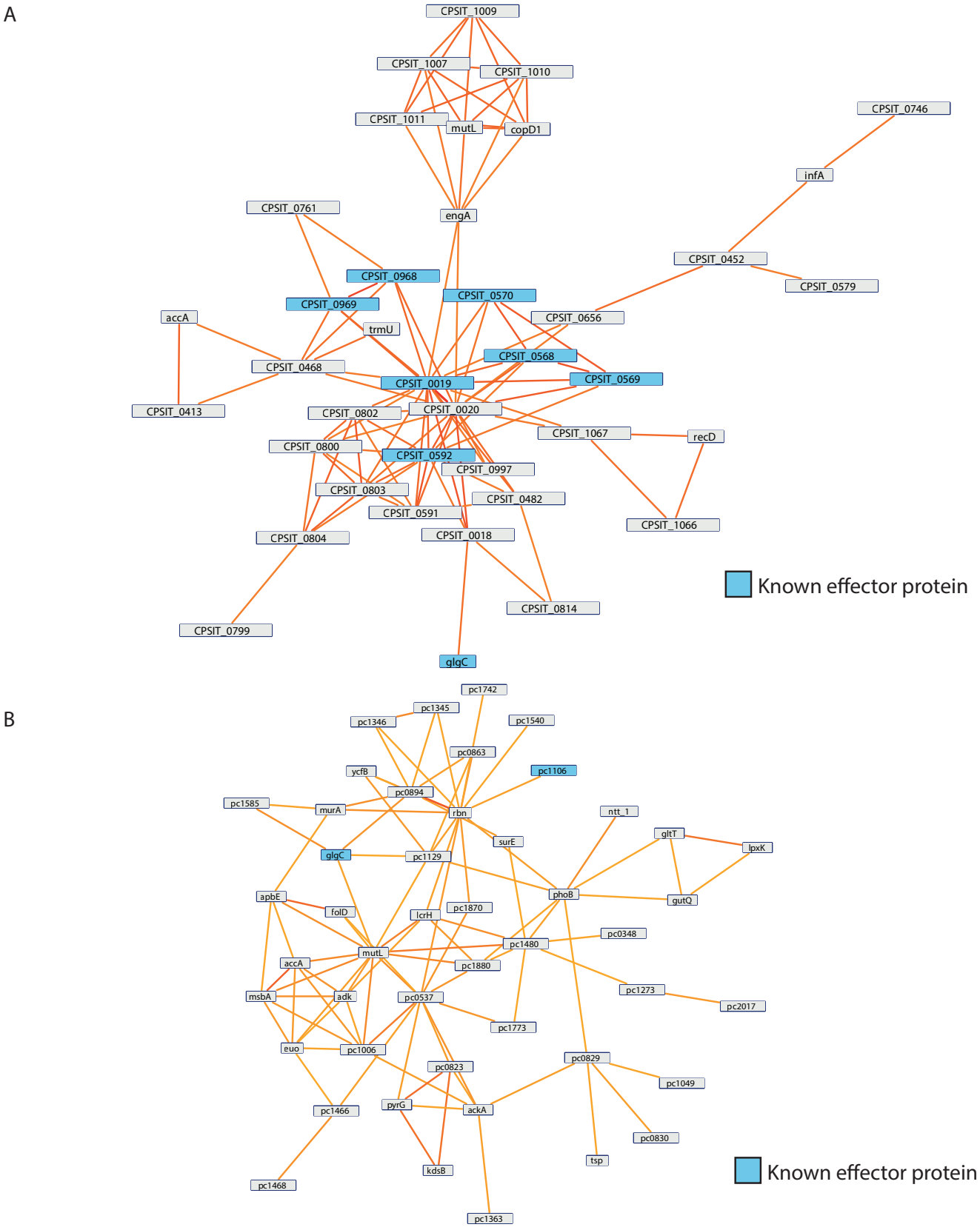
Supplementary Figure S3. **Conserved motifs for predicted novel HrcA regulon members.** The discovered motifs for both *phoH* (A) and CPn0105 (B) are near perfect matchers to the consensus CIRCE element reported in Wilson and Tan (2004), strongly suggesting that these genes are under the control of HrcA.



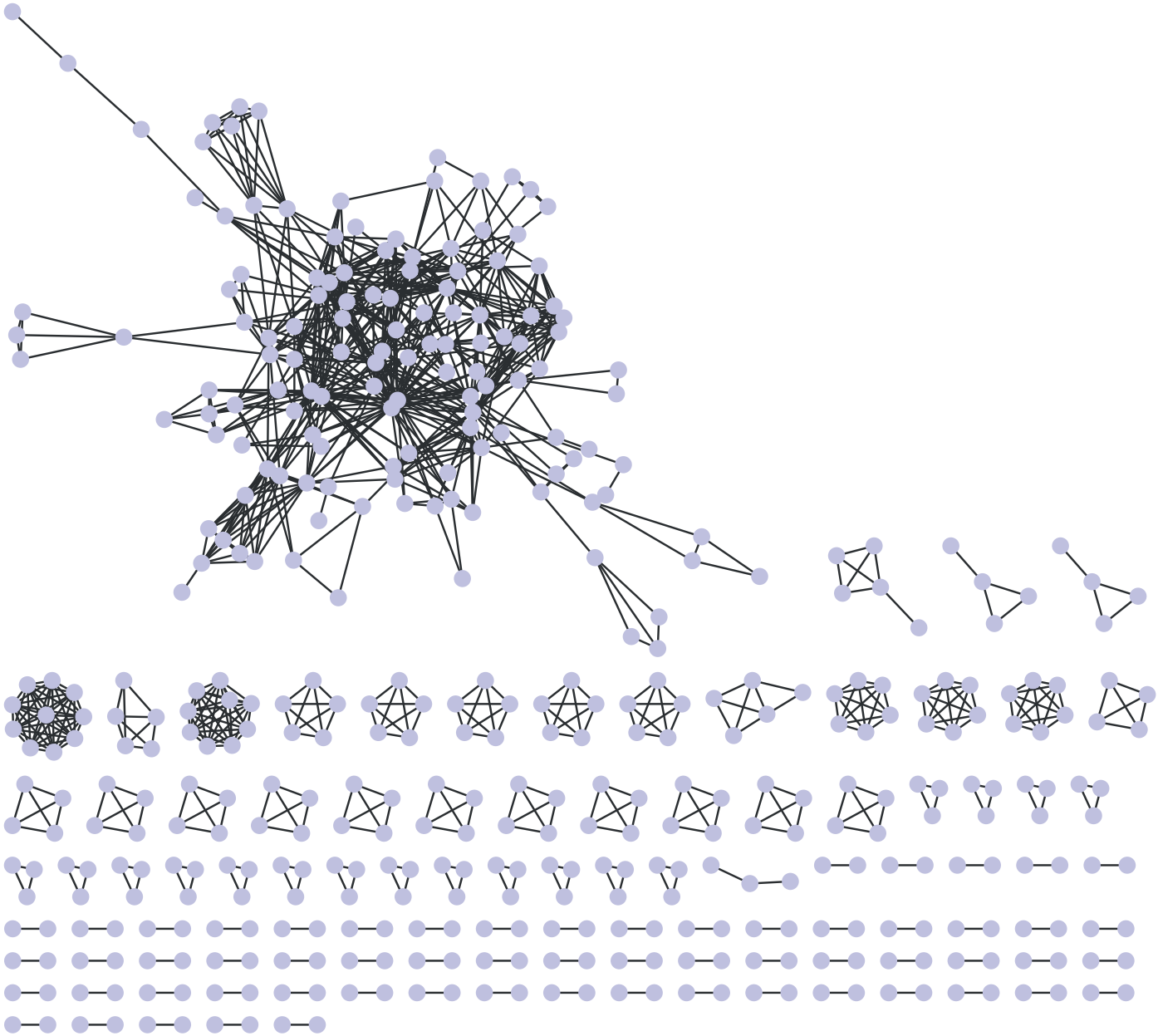
Supplementary Figure S4. **EUO and ChxR regulons found in predicted networks.** The co-regulatory network predictions from *Chlamydia trachomatis* 434/Bu of the known members of the (A) EUO regulon and (B) ChxR regulon. The known EUO regulon currently consists of 15 members, while the known ChxR regulon is comprised of five. EUO members are colored in green and ChxR are colored in blue. Six of members of the EUO network have direct connections with each other in the network. Only chxR and CTL0339 (CT_084) are linked among the ChxR regulon. (C) Extracting the nearest neighbors from the complete network for each known member of both the EUO and ChxR regulons, one observes a tightly connected subnetwork of 178 nodes. This tight clustering is strongly indicative of these genes being part of the same regulon. It may be that some members are duely regulated, and thus the EUO and ChxR networks are linked together. (D) Logo plot of the motif detected using MEME in the promoter regions of genes in (C).



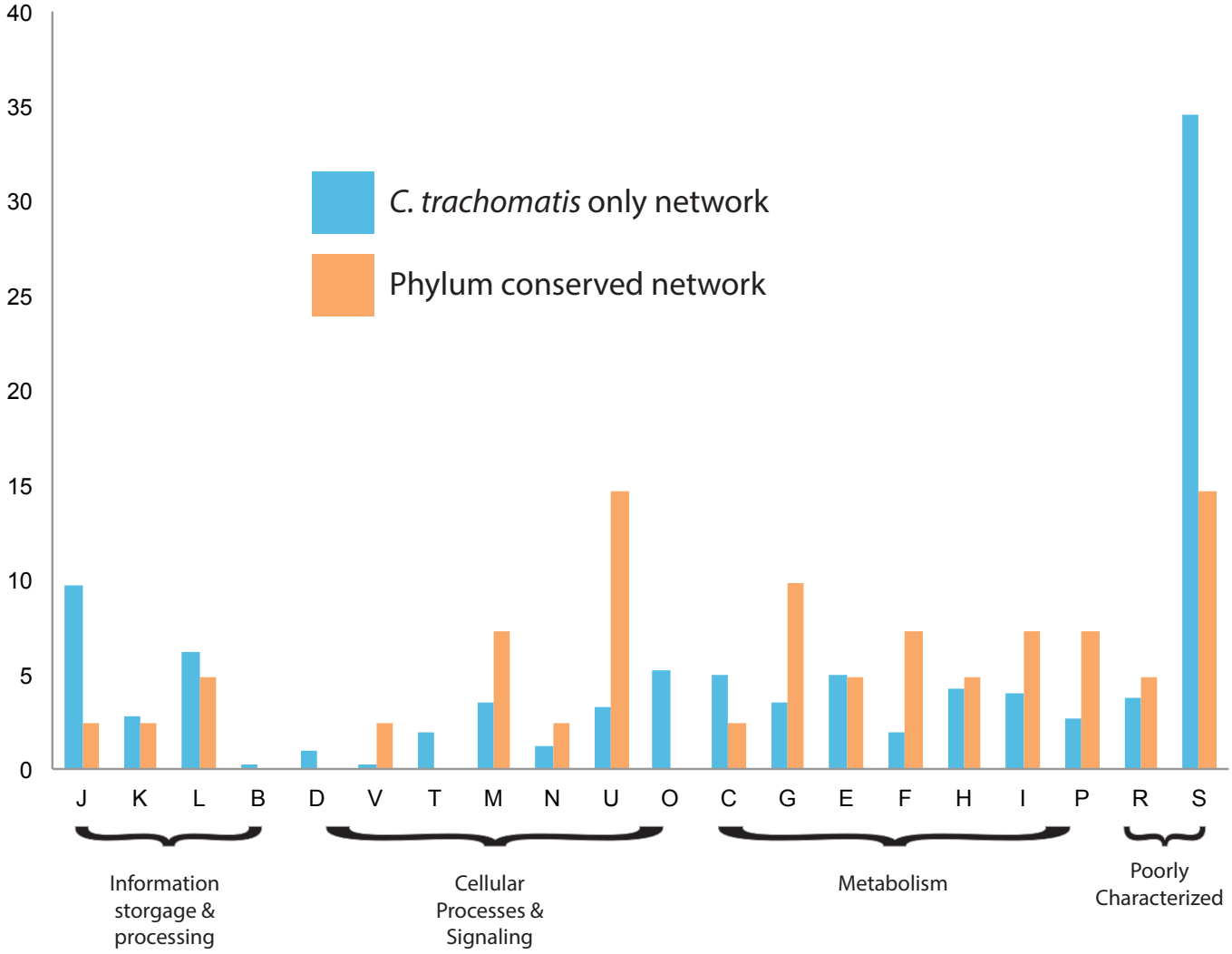
Supplementary Figure 5. **Predicted virulence co-regulation sub-network of *C. trachomatis*.** The original network was filtered for edges with a DPbits score >= 5 and the putative virulence regulon extraced. This network is comprised of 89 nodes, including many known Type III secretion effector proteins and other genes involved in virulence. The strength of the predictions are given by the weight and color of the edges, where the more red and thicker equals better prediction.



Supplementary Figure S6. **Predicted virulence co-regulatory networks for (A) *Chlamydia psittaci* 6BC and (B) *Protochlamydia amoebophila UWE25*.** The network for C. psittaci was filtered at a DPbits score >=6, while the the Protochlamdyia networks was filtered at DPbits score >=4. Orthologs to known effector proteins are colored in blue. Many of the effector proteins found in *C. trachomatis* are still highly connected in the *C. psittaci* network, indicating genus level conservation of virulence genes. Protochlamydia still share some known effectors, but many of the genes linked here are of unknown function, possibly revelaing a different set of virulence factors for this organism.



Supplementary Figure S7. **Consensus co-regulatory network from members of the *Chlamydiaceae*.** This network was created by only incorporating edges that are present in all six *Chlamydiaceae* genomes analyzed. The network contains 443 nodes and 975 edges.



Supplementary Figure S8. **Comparison of eggNOG functional categories for the *C. trachomatis* and phylum conserved sub-networks.** The phylum conserved network is enriched in eggNOG functional categories that are involved in metabolism and cellular proccesses and signaling versus that of the *C. trachomatis* network. Note the dramatic reduction in the poorly characterized proteins, as this largely represents the Type III effector proteins used for host cell entry, rather than those used by chlamydia to exploit the intracellular niche. eggNOG functional categries are descbied here: ftp://eggnog.embl.de/version_4.0.beta/data/downloads/eggnogv4.funccats.txt.

Chapter V

Plastid establishment did not require a chlamydial partner

nature COMMUNICATIONS

ARTICLE

Received 5 Dec 2014 | Accepted 27 Jan 2015 | Published 11 Mar 2015

DOI: 10.1038/ncomms7421 OPEN

Plastid establishment did not require a chlamydial partner

Daryl Domman[1], Matthias Horn[1], T. Martin Embley[2] & Tom A. Williams[2]

Primary plastids descend from the cyanobacterial endosymbiont of an ancient eukaryotic host, but the initial selective drivers that stabilized the association between these two cells are still unclear. One hypothesis that has achieved recent prominence suggests that the first role of the cyanobiont was in energy provision for a host cell whose reserves were being depleted by an intracellular chlamydial pathogen. A pivotal claim is that it was chlamydial proteins themselves that converted otherwise unusable cyanobacterial metabolites into host energy stores. We test this hypothesis by investigating the origins of the key enzymes using sophisticated phylogenetics. Here we show a mosaic origin for the relevant pathway combining genes with host, cyanobacterial or bacterial ancestry, but we detect no strong case for *Chlamydiae* to host transfer under the best-fitting models. Our conclusion is that there is no compelling evidence from gene trees that *Chlamydiae* played any role in establishing the primary plastid endosymbiosis.

[1] Department of Microbiology and Ecosystem Science, University of Vienna, A-1090 Vienna, Austria. [2] Institute for Cell and Molecular Biosciences, Newcastle University, Newcastle upon Tyne NE2 4HH, UK. Correspondence and requests for materials should be addressed to T.A.W. (email: tom.williams2@ncl.ac.uk).

Endosymbiosis is key to the evolutionary success of eukaryotes[1], from the ancient endosymbiotic origins of mitochondria and chloroplasts[2] to the methanogenic symbionts of anaerobic ciliates[3] and the nutritional symbioses of sap-feeding insects[4]. Cell biology and phylogenetics testify to the prokaryotic origins of these endosymbiotic organelles, but the molecular mechanisms by which their free-living progenitors were originally recruited and integrated with a host cell remain poorly understood. The endosymbiotic capture of a cyanobacterium by a heterotrophic eukaryotic host cell at the origin of the Archaeplastida marked one of the most important events in evolutionary history, for through this symbiosis all plant life would emerge. Other photosynthetic eukaryotes obtained their plastids through secondary endosymbiosis of one of these primary lineages, implying that—with a single exception[5]—all photosynthetic eukaryotes trace the origin of their photosynthetic machinery to the primary cyanobacterial endosymbiosis[6]. However, despite substantial progress on the evolution of plastids and their relationships to free-living cyanobacteria[7,8], the initial selective pressure that drove the acquisition and retention of the cyanobacterial endosymbiont remains unclear. Modern plastid and host metabolisms are intimately intertwined, with the chloroplast providing primarily fixed carbon to the host in exchange for a multitude of metabolites, including phosphate derivatives and NAD[9]. However, present-day host–plastid interactions are the product of more than a billion years of co-evolution and the situation may have been very different at the time of the initial endosymbiosis. In addition to the provision of carbohydrates to the host[10], nitrogen fixation[8] and the production of molecular oxygen for use by host mitochondria[11] have also been proposed as initial selective drivers for the retention of the cyanobacterial endosymbiont.

Recently, a detailed, metabolically explicit hypothesis for the initial selective pressure driving endosymbiosis was proposed in which the heterotrophic host cell that engulfed the cyanobacterial endosymbiont was already infected with an ancient member of the *Chlamydiae*[12–16]. In this 'ménage à trois'[16] (Fig. 1), named with reference to the proposed tripartite nature of the endosymbiosis, the chlamydial partner secreted a series of effectors that manipulated the host cell, rerouting host energy through glycogen metabolism for subsequent conversion to maltotetraose and import into the pathogen[16]. The proposed first step in this process was the conversion of host glucose-1-phosphate to the bacterial metabolite adenine diphosphoglucose (ADP-glucose) by the chlamydial effector GlgC; ADP-glucose was subsequently polymerized to glycogen and then processed for import by the pathogen through a series of downstream reactions catalysed by the effectors GlgA, GlgP and GlgX, all secreted by the pathogen into the host cytoplasm. In this scenario, an engulfed cyanobacterium could have provided immediate relief to the infected host cell through the provision of ADP-glucose generated as a byproduct of its own metabolism, preventing further depletion of host energy stores, that is, the energy sink represented by the chlamydial pathogen would provide the initial selective pressure for capture and retention of the cyanobiont. Although the immediate effect would have been to rescue the host cell, this tripartite metabolic interaction might also have potentiated the development of long-term endosymbiosis by establishing an initial metabolic link between host and cyanobiont, through the incorporation of cyanobacterial ADP-glucose into host glycogen stores.

The ménage à trois idea is a useful hypothesis, because it makes explicit cell biological and phylogenetic predictions that can be tested against currently available data. Modern *Chlamydiae* have a broad host range, from humans (where *Chlamydia trachomatis* is a major cause of sexually transmitted disease) to cattle, fish, isopods and protists[17]. However, extant *Chlamydiae* are not known to infect any members of the Archaeplastida, although the situation may have been different in the distant past[18]. The 'smoking gun' for the mitochondrial and plastid endosymbioses was the detection of an organelle[19], and although there is currently no evidence for a chlamydia-derived organelle in modern Archaeplastida, the chlamydial partner might have been lost from the consortium following horizontal transfer (HGT) of the key metabolic genes to the host nucleus[13,14,16]. In support of the hypothesis, some modern pathogenic *Chlamydiae* appear to manipulate host metabolism by the secretion of glycolytic enzymes[16,20] and some of the homologues of these enzymes from environmental *Chlamydiae* were shown to be secreted by the type III secretion system in a *Shigella* assay[16]. Further, published gene trees for some archaeplastidal enzymes involved in carbohydrate metabolism show the archaeplastidal sequences

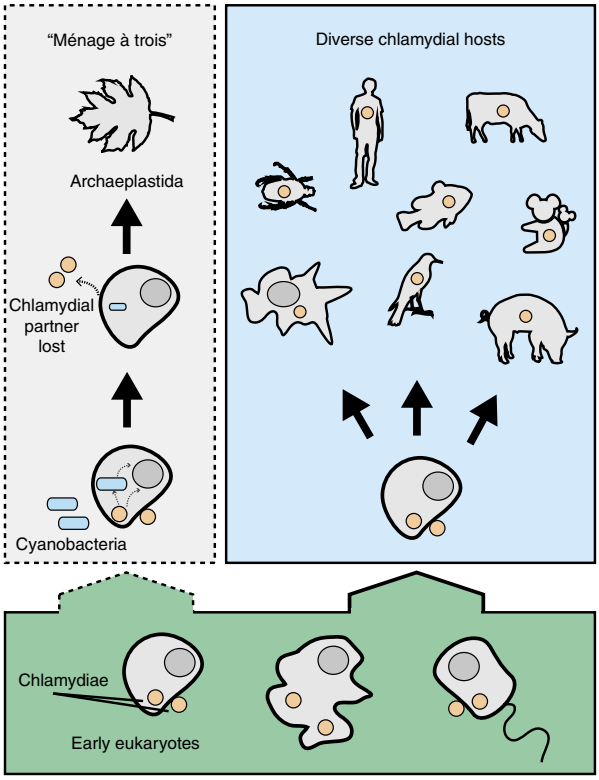


Figure 1 | The evolutionary history of chlamydia–eukaryote interactions and the ménage à trois hypothesis for plastid establishment. Chlamydiae have probably been associated with eukaryotes for at least 700 million years (17) so it appears reasonable to suggest that they also infected even more ancient eukaryotes. Extant *Chlamydiae* can infect a tremendously diverse range of eukaryotic hosts such as humans, cattle, pigs, birds, koala, fish, insects and unicellular protists. Notably, *Chlamydiae* have not been found infecting any member of the Archaeplastida. A proposed evolutionary scenario, coined the 'ménage à trois' hypothesis[16], posits that an early eukaryotic cell was host to both a chlamydial and cyanobacterial partner. Key metabolic genes that enabled the symbiotic capture of the cyanobacterium are proposed to have been horizontally transferred from chlamydia primarily to the host, but also to the cyanobacterium. Once these genes were transferred, the chlamydial partner was no longer needed and was subsequently lost. The newly formed relationship between cyanobacterium and host led to the modern plastid and the evolution of Archaeplastida.

emerging from within, or clustering with, the *Chlamydiae*. These include the genes encoding the putative effectors GlgX and GlgA mentioned above. These trees are compatible with a chlamydial origin for key components of plant carbohydrate metabolism, providing phylogenetic support for the ménage à trois hypothesis.

However, the deep internal branches of single gene trees are notoriously difficult to reconstruct, because they are often highly sensitive to the methods used, particularly when inferring phylogenies from anciently diverged sequences[21,22]. Standard phylogenetic models make simplifying assumptions about the evolutionary process that are often not met, with potential consequences for the relationships inferred. Here we re-evaluate the phylogenetic evidence for the ménage à trois hypothesis using a range of more complex evolutionary models that incorporate additional features of the sequence data shown to be important by statistical tests of model fit. Analyses using the best-fitting phylogenetic models reveal a mosaic origin for archaeplastidal storage polysaccharide metabolism, raise the possibility that some previous analyses have been misled by simple evolutionary models and suggest that there is no need to invoke a chlamydial contribution to the plastid endosymbiosis.

Results and Discussion

Simple methods do not adequately model sequence evolution. Under the ménage à trois hypothesis, archaeplastidal GlgC, GlgA, GlgP and GlgX originated as chlamydial effectors whose coding sequences were later transferred to the host nucleus; as a consequence, their modern-day archaeplastidal homologues are expected to cluster within, or as the sister to, chlamydial genes in single gene trees. Published phylogenies of these genes have made use of the single-matrix substitution models JTT[13], WAG[14,15,23] and LG[16], which all share the simplifying assumptions that the process of evolution is homogeneous across the sites of the alignment and the branches of the tree. These assumptions are frequently violated by real molecular sequences, in which sequence composition, and by inference evolutionary process, often varies extensively in both of these dimensions. Violation of these assumptions results in poor model fit and can lead to phylogenetic artefacts such as long-branch attraction, in which fast-evolving sequences (long branches) cluster together irrespective of true historical relationships; as a result, analyses with poorly fitting models can potentially lead to the recovery of strongly supported but incorrect phylogenetic trees[24]. To evaluate whether the assumptions of single-matrix models are met by the enzymes key to the ménage à trois hypothesis, we performed posterior predictive simulations[25] on alignments of GlgC, GlgX, GlgA, GlgP and UhpC under the LG model, which according to analyses using the model selection tool ProtTest 3.4 (ref. 26) was the best-fitting single-matrix model in all cases. Posterior predictive simulations provide a test of model adequacy by comparing the properties of data simulated under the model to the real alignment; significant compositional differences between the observed and simulated data suggest that the assumptions of the model are unrealistic for the data at hand. Our simulations indicated that all five alignments contained significant across-site and across-branch compositional variation that was not adequately accounted for by the single-matrix LG model (see Fig. 2 and Supplementary Table 1). These results suggested that LG provided an inadequate fit to the data with respect to sequence composition, raising the possibility that the resulting phylogenies might be affected by phylogenetic artefacts such as long-branch attraction and motivating the use of more complex models.

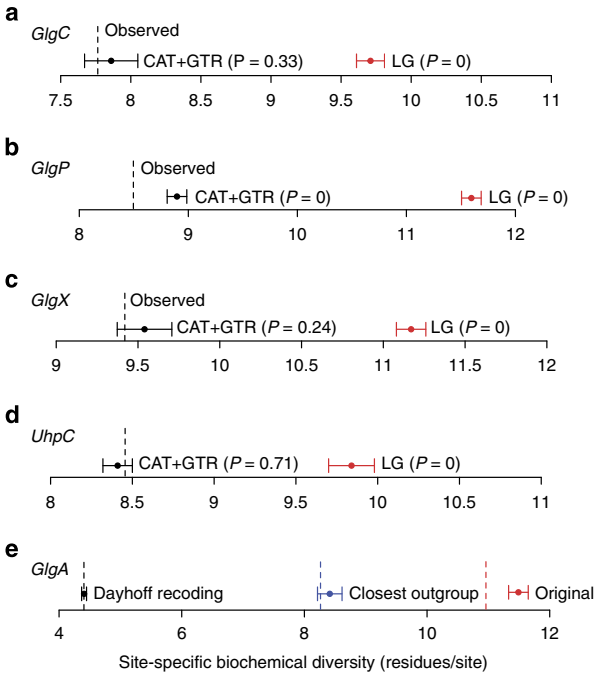


Figure 2 | Bayesian posterior predictive simulations for assessing model fit to the key enzymes implicated in the ménage à trois. Posterior predictive simulations[25] are a technique for assessing model adequacy with respect to key properties of the sequence alignment, which has an impact on phylogenetic inference. Here we compared the ability of the LG and CAT + GTR models to adequately capture the site-specific biochemical constraints experienced by the genes implicated in the 'ménage à trois' hypothesis. In sequence alignments, these constraints are manifest in the reduced number of amino acids observed in any one alignment column, which is usually much less than the theoretical maximum of 20. (**a**) The mean observed number of different amino acids per site in the GlgC alignment was 7.78. Data simulated under the LG model showed mean per-site diversity values (dot) much higher than the real data, suggesting this model did a poor job of modelling site-specific constraints. In contrast, the range (bars) of site-specific diversities predicted under the CAT + GTR model was comparable to that of the real data ($P = 0.33$), suggesting adequate model fit with respect to this important metric. (**b**–**d**) The results for our analyses of GlgP, GlgX and UhpC were similar, with the CAT + GTR model better able to capture site-specific constraints, although neither model produced realistic predictions for the GlgP alignment. (**e**) Analyses of three different GlgA alignments under the CAT + GTR model. The original data set contained a large and highly diverse outgroup, leading to a high per-site diversity and poor model fit. An outgroup consisting only of the sequences most closely related to the relevant GlgA clade reduced per-site diversity and enabled adequate model fit; Dayhoff recoding of the original alignment also resulted in improved model fit relative to the unrecoded data. In both analyses in which adequate model fit was achieved, we did not recover a specific *Chlamydiae*/Archaeplastida clade, as discussed in the main text. Error bars represent s.e. and *P*-values were calculated using the 'ppred' and 'readpb_mpi' programmes in the PhyloBayes and PhyloBayes-MPI packages, respectively.

Better-fitting models do not support the ménage à trois. In the last decade, growing recognition of the problems of systematic error in phylogenetics[24], improvements in computational power and the increasing popularity of Bayesian approaches have stimulated the development of more complex phylogenetic models that can accommodate across-site and across-branch compositional variation[27–29]. These are pervasive features of real sequence data that, when not adequately modelled, are known to

lead to topological errors in inferred trees. In particular, variation in sequence composition across the sites of an alignment is a ubiquitous feature of sequence data that arises from the site-specific selective constraints experienced by functional biological molecules; failure to account for the impact of these constraints on sequence evolution often results in poor modelling of the substitution process and can lead to phylogenetic artefacts such as long-branch attraction (LBA)[24]. One of the most useful approaches to modelling these site-specific constraints is the CAT family of substitution models[29] that accommodate across-site compositional variation by allowing sites to be fit by distinct equilibrium composition profiles; as a result, these models have been shown to be more resistant that standard single-matrix models to systematic phylogenetic error and LBA[30]. We therefore applied these methods to the archaeplastidal genes predicted to trace their ancestry to the chlamydial partner in the 'ménage à trois'. We compared the fit of these more complex models to the single-matrix models previously applied to these genes using posterior predictive simulations; the results of these tests are summarized in Fig. 2 and Supplementary Table 1, and are discussed on a per-gene basis below.

The first step in manipulation of the heterotrophic host cell by the ancient chlamydial pathogen is suggested to be the conversion of host energy, in the form of glucose-1-phosphate, to ADP-glucose via the ADP-pyrophosphorylase GlgC. However, our phylogenetic analyses of GlgC homologues from Archaeplastida, *Chlamydiaceae*, *Cyanobacteria* and other bacterial groups recovered the archaeplastid sequences clustering with the *Cyanobacteria* with maximal posterior support (Posterior probability, PP = 0.99 in the CAT + GTR analysis; see Fig. 3a and Supplementary Fig. 1). Within this clade, the archaeplastid sequences (with the exception of those from the green algae *Chlamydomonas* and *Ostreococcus*) emerged from within the *Cyanobacteria*, albeit with more modest support (PP = 0.84). The simplest interpretation of these results is that glgC of modern Archaeplastida was obtained directly from the cyanobiont by endosymbiotic gene transfer[31].

Following the generation of ADP-glucose by GlgC and its incorporation into host glycogen by GlgA (our analysis of which is discussed below), the next step in the exploitation of host energy by the ancient chlamydial pathogen is proposed to be the priming of glycogen for attack by parasite isoamylase (GlgX)[16]. The enzyme that performs this step is a glycogen phosphorylase, GlgP, which catalyses glycogen breakdown by recessing glycogen chains to within four residues of an α-1,6 branch. Our phylogenetic analyses did not recover a chlamydial, or indeed a cyanobacterial, origin for archaeplastidal glgP under any of the models used. Instead, the archaeplastidal sequences grouped with some other eukaryotes away from both the cyanobacterial and chlamydial clades (Fig. 3b and Supplementary Fig. 2), consistent with vertical descent of GlgP from the heterotrophic host cell for the cyanobacterial endosymbiont.

Under the ménage à trois hypothesis, the original role of chlamydial GlgX (isoamylase) was the generation of maltotetraose from host glycogen for import into the pathogen. Published trees inferred under the LG model[16] recovered a *Chlamydiae*/archaeplastid clade clearly distinct from other bacteria, but support for the relationships within this clade were weak. Phylogenetic inference under the better-fitting CAT + GTR model ($P = 0.24$ for across-site compositional heterogeneity, suggesting adequate model fit; see Fig. 2c and Supplementary Table 1) recovered a well-resolved *Chlamydiae*/archaeplastid clade in which the *Chlamydiae* emerged from within the Archaeplastida with high posterior support (PP = 0.98 for CAT + GTR; see Fig. 3c and Supplementary Fig. 3). These results are consistent with the horizontal transfer of GlgX between *Chlamydiae* and Archaeplastida, but suggest that the direction was to, rather than from, the *Chlamydia* and hence they do not support the ménage à trois hypothesis.

Origins of the UhpC hexose phosphate transporter. Genome analysis of the glaucophyte *Cyanophora paradoxa* has recently identified a homologue of UhpC, a hexose transporter of potential chlamydial origin[32]. This finding prompted a revision of the ménage à trois scenario in which an initial horizontal transfer of the *uhpC* gene from the chlamydial pathogen to the genome of the cyanobacterial endosymbiont provided the transporter needed for the export of photosynthate from the cyanobiont, in the form of glucose-1-phosphate[33]. It is worth pointing out that as the heterotrophic host cell would then have been able to make use of glucose-1-phosphate directly, this extension of the ménage à trois might seem to obviate the subsequent need for the chlamydial partner. Nonetheless, the revised theory locates both the cyanobiont and chlamydial partner in an inclusion within the eukaryotic host cell and posits that the glucose-1-phosphate exported from the cyanobiont by the chlamydial transporter was subsequently converted to ADP-glucose by chlamydial GlgC and transported into the host cytoplasm by a host-derived transporter, where it then followed the same fate as in the original ménage à trois model[33].

A key issue when inferring and interpreting phylogenies of multigene families is the placement of the root, in particular for a transporter such as UhpC that is a member of the major facilitator superfamily and is related to the glycerol-3-phosphate transporter GlpT[34]. The published tree[32] includes UhpC sequences from *Proteobacteria*, *Chlamydiae* and Archaeplastida and, when rooted on or within the *Proteobacteria*, produces a topology consistent with horizontal transfer of the gene from *Chlamydiae* to Archaeplastida. Our UhpC phylogeny, inferred under the better-fitting CAT + GTR model, would also support a transfer from *Chlamydiae* into plants but only if it too was rooted within the proteobacteria. By contrast, inclusion of the GlpT sequences as an outgroup placed the root between the archaeplastidal and bacterial sequences as a whole (Fig. 3d and Supplementary Fig. 4), eliminating any compelling case for specific gene transfer from *Chlamydiae* to Archaeplastida.

High levels of compositional heterogeneity in GlgA. In the ménage à trois, the original role of the glycogen synthase GlgA is proposed to have been the incorporation of ADP-glucose generated by GlgC into host glycogen for later exploitation by the chlamydial pathogen. This enzyme would therefore have established the intial link between host and cyanobiont metabolism by providing a route for the incorporation of a bacterial metabolite (ADP-glucose) into host energy stores. In agreement with the recent analyses of Ball *et al.*[16], a phylogeny inferred under the LG model provides moderate support (PP = 0.83) for chlamydial ancestry of the archaeplastid sequences, although this model was rejected in our posterior predictive simulations both for across-site and across-branch compositional heterogeneity ($P = 0$ for across-site compositional heterogeneity, $P = 0.002$ for across-branch heterogeneity; see Supplementary Table 1). Indeed, the GlgA alignment proved to be extremely heterogeneous in both across-site and across-branch composition; unusually, even the most general substitution model currently available (CAT + GTR) failed to provide an adequate fit with respect to across-site compositional variation ($P = 0$, see Fig. 2e and Supplementary Table 1). Although the tree inferred under CAT + GTR did not fit the data, it did weakly (PP = 0.74) support a *Chlamydiae* plus Archaeplastida clade, consistent with horizontal exchange (Fig. 4a and Supplementary Fig. 5).

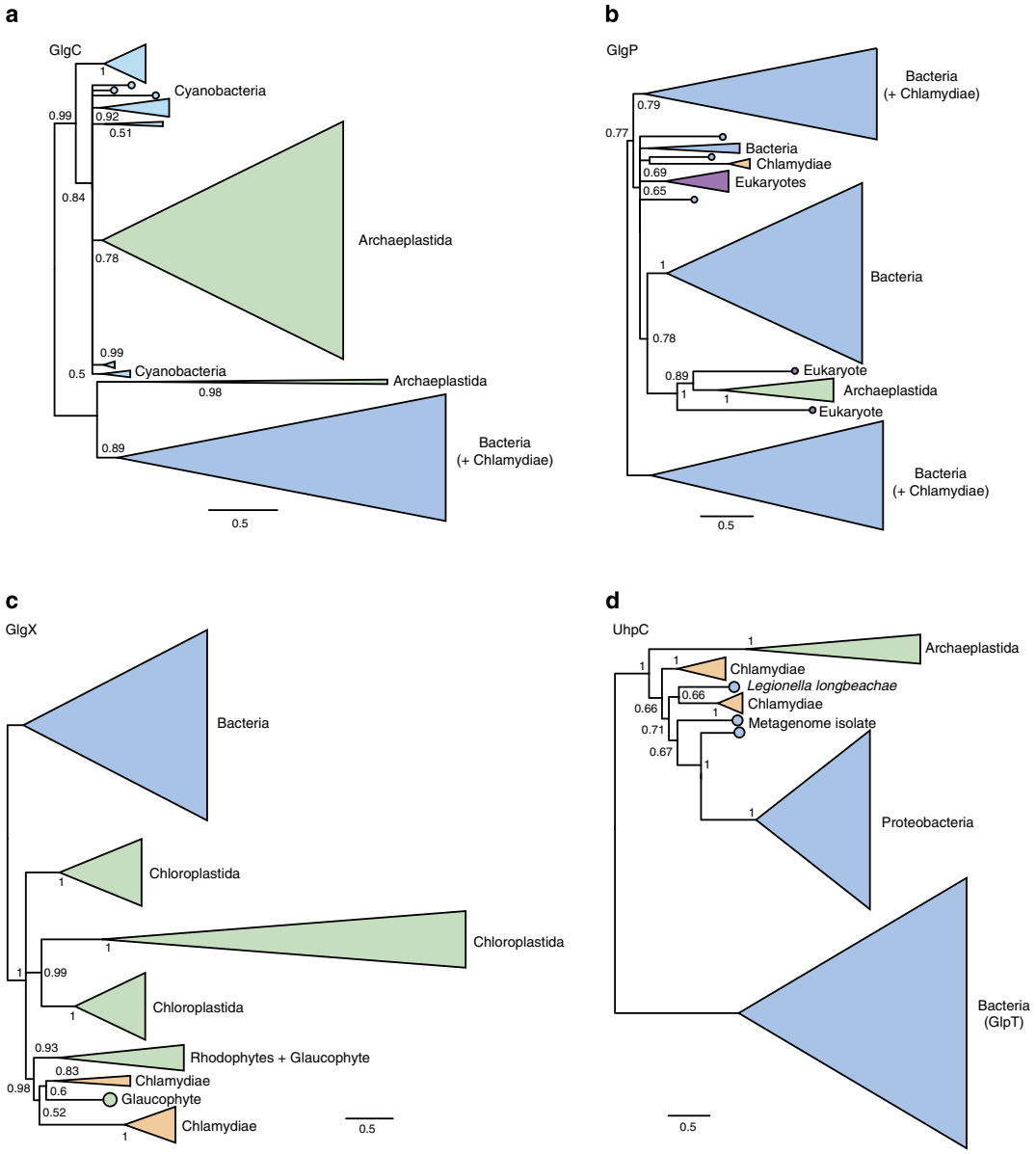


Figure 3 | Single gene trees for key components of archaeplastidal carbohydrate metabolism implicated in the ménage à trois. (**a**–**d**) Phylogenies for GlgC, GlgP, GlgX and UhpC. These trees were inferred under the CAT + GTR model in PhyloBayes, which performed better in our analyses of model fit than the single-matrix models originally used to analyse these genes. With the exception of the *Chlamydomonas* and *Ostreococcus* GlgC sequences, the Archaeplastida were recovered as a monophyletic group in all of these trees, suggesting that this pathway was already present in its current form in the last common ancestor of the group. However, the closest outgroup to the Archaeplastida varies among the individual gene trees, as discussed in the main text. We rooted the tree in panel (**d**) between UhpC and its paralogue GlpT. In the other panels, we oriented the trees to most clearly visualize the key relationships between the archaeplastid and chlamydial sequences, and to test the predictions of the ménage à trois hypothesis. Support values are summarized as Bayesian posterior probabilities and branch lengths are proportional to the expected number of substitutions per site, as indicated by the scale bar.

The original GlgA alignment of Ball *et al.*[16] contains, in addition to the chlamydial and archaeplastidal sequences that are key to the ménage à trois hypothesis, an extensively sampled outgroup that includes distantly related, functionally divergent paralogues of these enzymes from Archaeplastida and bacteria. We reasoned that the large evolutionary distances, functional shifts and associated long branches that characterize this outgroup might be a contributing factor to the failure of the CAT + GTR model to adequately capture the compositional heterogeneity evident in the data set, potentially interfering with the inference of in-group relationships[35]. To test this idea, we removed most of the outgroup sequences, retaining only the clade that branched closest to the key *Chlamydiae*/Archaeplastida clade in the initial CAT + GTR analysis, and performed inference on this reduced data set under the same model. The removal of the more distant outgroup sequences significantly reduced the

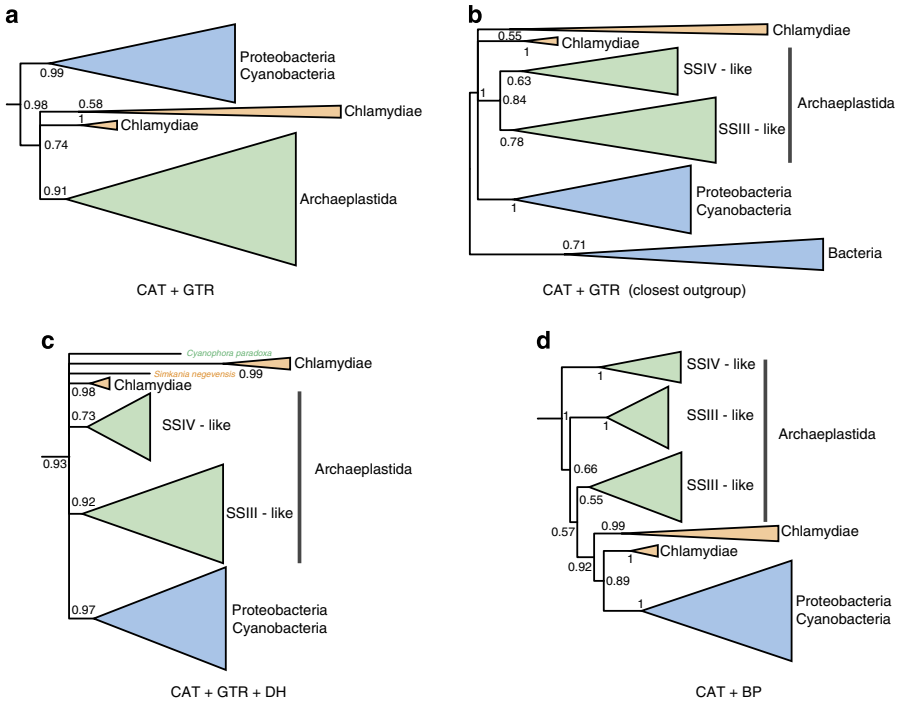


Figure 4 | Phylogenetic analyses of the glycogen synthase GlgA. (**a**) Inference under the CAT + GTR model recovers a weakly supported (PP = 0.74) clade comprising the chlamydial and Archaeplastidal sequences, but does not support horizontal transfer from *Chlamydiae* to Archaeplastida. This alignment was unusually heterogeneous in terms of sequence composition, and the CAT + GTR model failed our posterior predictive test for across-site compositional heterogeneity ($P = 0$). (**b**) Inclusion of only the closest outgroup sequences improved the fit of the CAT + GTR model and collapsed this relationship, recovering an in-group trichotomy between the sequences from Archaeplastida, *Chlamydiae* and other bacteria. (**c**) Analysis of the Dayhoff-recoded data set under the CAT + GTR model; Dayhoff recoding ameliorated the observed compositional heterogeneity and also failed to recover a specific *Chlamydiae*/Archaeplastida relationship. (**d**) Joint modelling of across-site and across-branch compositional variation using the non-stationary CAT + BP model, which also failed to recover a specific relationship. These panels represent sub-trees derived from larger analyses showing the portion of the tree containing the chlamydial and archaeplastidal sequences; the root positions indicated are based on the topology of the complete analyses. Support values are summarized as Bayesian posterior probabilities, and branch lengths are proportional to the expected number of substitutions per site.

compositional heterogeneity in the data set so that the CAT + GTR model now provided an adequate fit for across-site compositional variation ($P = 0.21$, see Fig. 2e), although it still failed the across-branch test (see Supplementary Table 1). Interestingly, this analysis no longer recovered a specific *Chlamydiae*/Archaeplastida clade (Fig. 4b and Supplementary Fig. 6), suggesting that the weakly supported relationship observed in the original tree may have been the result of poor model fit.

Given the poor fit of CAT + GTR to the GlgA sequences, we also evaluated two alternative approaches for modelling the evolution of GlgA: Dayhoff recoding and joint modelling of across-site and across-branch compositional variation using the CAT + BP model. In Dayhoff recoding[36], the 20 amino acids are clustered into 6 bins such that the substitution rates among amino acids in the same bin are higher than between bins. By recoding amino acid data into these six classes and only modelling substitutions between bins, the degree of substitutional saturation and compositional heterogeneity in the data is greatly reduced, often helping to ameliorate poor model fit and the effects of systematic phylogenetic error[37]. Dayhoff recoding inevitably results in some information loss, but the net effect is often a substantial improvement in phylogenetic accuracy[37], especially on extremely heterogeneous data sets such as the GlgA alignment[16] that we analyse here. Indeed, posterior predictive tests on the complete alignment of Ball *et al.*[16] analysed under the CAT + GTR model showed adequate fit with respect to both across-site and across-branch composition after Dayhoff recoding ($P = 0.49$ and 0.1, respectively; see Fig. 2e and Supplementary Table 1), demonstrating a large improvement in model fit over the un-recoded data. As with our analysis on the unrecoded data using only the closest outgroup, the phylogeny inferred under this model did not recover a specific *Chlamydiae*/Archaeplastida relationship (Fig. 4c and Supplementary Fig. 7), instead recovering a clade (PP = 0.93) in which the relationships among the *Chlamydiae*, Archaeplastida SSIII and SSIV-like, and other bacterial sequences were unresolved.

Finally, we attempted to jointly model the across-branch and across-site compositional variation in the GlgA alignment using the non-stationary CAT + BP model[27], which combines modelling of across-site composition in the same way as the CAT model with a process in which composition can change at breakpoints (BPs) across the phylogenetic tree, leading to across-branch compositional variation. These analyses also failed to recover a *Chlamydiae*/Archaeplastida clade (Fig. 4d and Supplementary Fig. 8). Overall, our analyses demonstrate that the evolution of the GlgA gene is unusually difficult to model, given the high levels of both across-site and across-branch compositional variation observed. Nonetheless, our analyses using a series of better-fitting models suggest that there is no convincing support for a specific *Chlamydiae*/Archaeplastida relationship.

In summary, our phylogenetic analyses suggest a mosaic origin for archaeplastidal carbohydrate metabolism: the ADP-pyrophosphorylase GlgC descends from within the cyanobacteria, consistent with an origin from the cyanobacterial endosymbiont; the glycogen phosphorylase GlgP may descend from the eukaryotic host cell for that endosymbiont, and the glycogen synthase GlgA, the debranching enzyme GlgX and the hexose phosphate transporter UhpC appear to have bacterial, but not necessarily chlamydial, origins. Thus, in contrast to the predictions of the ménage à trois hypothesis, our analyses suggest that there is no compelling evidence that any of the key genes of archaeplastidal carbohydrate metabolism were acquired from an ancient chlamydial partner.

Implications for the plastid endosymbiosis. In addition to the genes directly implicated in the ménage à trois hypothesis that we discuss above, support for chlamydial involvement in the establishment of the plastid has also been derived from the observation that nearly 60 archaeplastidal genes group with *Chlamydiae* in genomic surveys of single-gene trees[12–16,38]. These trees have been interpreted as evidence of a batch horizontal transfer from *Chlamydiae* to Archaeplastida that could also reflect a long period of infection, symbiosis or co-habitation of the same ecological niche[13,14,16]. For reasons of computational speed, phylogenomic screens have employed single-matrix methods, such as the LG model discussed above, and are therefore subject to the same caveats as the gene trees analysed here. Beyond these methodological concerns, there is a deeper problem with inferring a special explanation for the presence of putative chlamydial genes on plant genomes, in the absence of any physical evidence of the proposed chlamydial partner. The problem is that recent studies have demonstrated that in addition to organellar genes shared with *Cyanobacteria* and *Alphaproteobacteria*, the Archaeplastida share more genes with *Gammaproteobacteria*, *Actinobacteria*, *Deltaproteobacteria*, *Bacilli*, *Bacteroidetes* and *Betaproteobacteria* than with *Chlamydiae*[8]. Given the extent of HGT, particularly of metabolic genes, among major cellular groups[39] and the demonstrated limitations of standard phylogenetic models for the archaeplastidal genes we analysed here, these patterns of gene sharing—including those involving *Chlamydiae*—are most simply explained as a mixture of genuine HGT events and tree reconstruction artefacts. Thus, in the absence of cytological evidence for a *chlamydia*-derived organelle, or support for the ménage à trois hypothesis from better-fitting phylogenetic models, we conclude that there is no compelling need to invoke a chlamydial partner in the establishment of the primary plastid endosymbiosis.

Methods

Sequences and alignments. The GlgA and GlgX alignments were those used in Ball *et al.*[16] For the other genes, gene families were downloaded from the HOGENOM[40] (UhpC and GlgC) or OMA[41] (GlgP) databases and augmented with their orthologues from a set of newly sequenced chlamydial genomes (*Neochlamydia sp.* TUME1 and EPS4, *Protochlamydia sp.* EI2 and *Parachlamydia sp.* OEW1) as well as additional cyanobacterial orthologues. Sequences for GlgC and GlgP were aligned using Muscle 3.8 (ref. 42) and poorly aligning regions were detected and removed using BMGE[43] with the BLOSUM30 scoring matrix. UhpC sequences were collected by extracting the top 250 BLAST hits of the *Protochlamydia amoebophila* homologue (Q6ME88_PARUW) against the UniRef90 database[44] and supplemented with the aforementioned chlamydial genomes. The alignment was performed using clustalOmega[45] and filtered using GBlocks[46] by using the parameters '-b4 = 3 -b5 = a'. All sequence sets, alignments and Newick tree files have been deposited in FigShare (http://dx.doi.org/10.6084/m9.figshare.1257740).

Phylogenetic analyses. Analyses using the CAT + GTR and CAT + GTR + Dayhoff models were performed using PhyloBayes-MPI[47] 1.5a and analyses using CAT-BP were performed using nhPhyloBayes[27]. Bayesian analyses using the LG model were performed in PhyloBayes 3.3 (ref. 48). For each analysis, two chains were run in parallel, and the bpcomp and tracecomp programmes were used to assess convergence. We judged that analyses had converged when the maximum discrepancies in bipartition frequencies (bpcomp) and summary statistics (tracecomp) between the two chains had all dropped below 0.1, and the effective sample size of each parameter was at least 100, as recommended in the PhyloBayes manual (http://www.phylobayes.org).

Posterior predictive simulations. Posterior predictive simulations were performed using converged runs to evaluate model fit. We used the ppred (PhyloBayes 3.3) and readpb_mpi (PhyloBayes-MPI 1.5a) programmes to perform tests of across-site (site-specific biochemical diversity) and across-branch (compositional homogeneity) tests for the LG, CAT + GTR and CAT + GTR + Dayhoff models. We judged that a model failed a particular test if the test statistic calculated on the real data fell outside the central 95% of the simulated distribution.

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. Zimorski, V., Ku, C., Martin, W. F. & Gould, S. B. Endosymbiotic theory for organelle origins. *Curr. Opin. Microbiol.* **22,** 38–48 (2014).
3. Embley, T. M. & Finlay, B. J. The use of small subunit rRNA sequences to unravel the relationships between anaerobic ciliates and their methanogen endosymbionts. *Microbiology* **140,** 225–235 (1994).
4. McCutcheon, J. P. & Moran, N. A. Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.* **10,** 13–26 (2011).
5. Marin, B., M. Nowack, E. C. & Melkonian, M. A plastid in the making: evidence for a second primary endosymbiosis. *Protist* **156,** 425–432 (2005).
6. Rodríguez-Ezpeleta, N. *et al.* Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. *Curr. Biol.* **15,** 1325–1330 (2005).
7. Deusch, O. *et al.* Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. *Mol. Biol. Evol.* **25,** 748–761 (2008).
8. Dagan, T. *et al.* Genomes of Stigonematalean cyanobacteria (subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids. *Genome Biol. Evol.* **5,** 31–44 (2013).
9. Weber, A. P. M. & Linka, N. Connecting the plastid: transporters of the plastid envelope and their role in linking plastidial with cytosolic metabolism. *Annu. Rev. Plant Biol.* **62,** 53–77 (2011).
10. Martin, W. & Kowallik, K. Annotated English translation of Mereschkowsky's 1905 paper 'Über Natur und Ursprung der Chromatophoren imPflanzenreiche'. *Eur. J. Phycol.* **34,** 287–295 (1999).
11. Martin, W. & Müller, M. The hydrogen hypothesis for the first eukaryote. *Nature* **392,** 37–41 (1998).
12. Brinkman, F. S. L. *et al.* Evidence that plant-like genes in chlamydia species reflect an ancestral relationship between *Chlamydiaceae*, cyanobacteria, and the chloroplast. *Genome Res.* **12,** 1159–1167 (2002).
13. Huang, J. & Gogarten, J. Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? *Genome Biol.* **8,** R99 (2007).
14. Moustafa, A., Reyes-Prieto, A. & Bhattacharya, D. *Chlamydiae* has contributed at least 55 genes to plantae with predominantly plastid functions. *PLoS ONE* **3,** e2205 (2008).
15. Becker, B., Hoef-Emden, K. & Melkonian, M. Chlamydial genes shed light on the evolution of photoautotrophic eukaryotes. *BMC Evol. Biol.* **8,** 203 (2008).
16. Ball, S. G. *et al.* Metabolic effectors secreted by bacterial pathogens: essential facilitators of plastid endosymbiosis? *Plant Cell Online* **25,** 7–21 (2013).
17. Horn, M. *Chlamydiae* as symbionts in eukaryotes. *Annu. Rev. Microbiol.* **62,** 113–131 (2008).
18. Subtil, A., Collingro, A. & Horn, M. Tracing the primordial *Chlamydiae*: extinct parasites of plants? *Trends Plant Sci.* **19,** 36–43 (2014).
19. Gray, M. W. & Doolittle, W. F. Has the endosymbiont hypothesis been proven? *Microbiol. Rev.* **46,** 1–42 (1982).
20. Lu, C. *et al.* Chlamydia trachomatis GlgA is secreted into host cell cytoplasm. *PLoS One* **8,** e68764 (2013).
21. Foster, P. G., Cox, C. J. & Embley, T. M. The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. *Philos. Trans. R. Soc. B Biol. Sci.* **364,** 2197–2207 (2009).
22. Philippe, H. & Roure, B. Difficult phylogenetic questions: more data, maybe; better methods, certainly. *BMC Biol.* **9,** 91 (2011).
23. Tyra, H. M., Linka, M., Weber, A. P. & Bhattacharya, D. Host origin of plastid solute transporters in the first photosynthetic eukaryotes. *Genome Biol.* **8,** R212 (2007).
24. Philippe, H. *et al.* Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol.* **9,** e1000602 (2011).
25. Bollback, J. P. Bayesian model adequacy and choice in phylogenetics. *Mol. Biol. Evol.* **19,** 1171–1180 (2002).
26. Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21,** 2104–2105 (2005).

27. Blanquart, S. & Lartillot, N. A site- and time-heterogeneous model of amino acid replacement. *Mol. Biol. Evol.* **25,** 842–858 (2008).
28. Foster, P. G. Modeling compositional heterogeneity. *Syst. Biol.* **53,** 485–495 (2004).
29. 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).
30. 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,** S4 (2007).
31. Martin, W. *et al.* Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc. Natl Acad. Sci.* **99,** 12246–12251 (2002).
32. Price, D. C. *et al. Cyanophora paradoxa* genome elucidates origin of photosynthesis in algae and plants. *Science* **335,** 843–847 (2012).
33. Facchinelli, F., Colleoni, C., Ball, S. G. & Weber, A. P. M. Chlamydia, cyanobiont, or host: who was on top in the ménage à trois? *Trends Plant Sci.* **18,** 673–679 (2013).
34. Lemieux, M. J., Huang, Y. & Wang, D. N. Crystal structure and mechanism of GlpT, the glycerol-3-phosphate transporter from E. coli. *J. Electron Microsc. (Tokyo)* **54**(Suppl 1): i43–i46 (2005).
35. Hirt, R. P. *et al.* Microsporidia are related to fungi: evidence from the largest subunit of RNA polymerase II and other proteins. *Proc. Natl Acad. Sci.* **96,** 580–585 (1999).
36. Hrdy, I. *et al.* Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. *Nature* **432,** 618–622 (2004).
37. Susko, E. & Roger, A. J. On reduced amino acid alphabets for phylogenetic inference. *Mol. Biol. Evol.* **24,** 2139–2150 (2007).
38. Collingro, A. *et al.* Unity in variety—the pan-genome of the *Chlamydiae*. *Mol. Biol. Evol.* **28,** 3253–3270 (2011).
39. Alsmark, C. *et al.* Patterns of prokaryotic lateral gene transfers affecting parasitic microbial eukaryotes. *Genome Biol.* **14,** R19 (2013).
40. Penel, S. *et al.* Databases of homologous gene families for comparative genomics. *BMC Bioinformatics* **10,** S3 (2009).
41. Altenhoff, A. M., Schneider, A., Gonnet, G. H. & Dessimoz, C. OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res.* **39,** D289–D294 (2011).
42. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
43. 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).
44. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23,** 1282–1288 (2007).
45. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7,** 539 (2011).
46. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17,** 540–552 (2000).
47. 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).
48. Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25,** 2286–2288 (2009).

Acknowledgements

This work was supported by a Wellcome Trust Program Grant (number 045404) and the European Research Council Advanced Investigator Programme (ERC-2010-AdG-268701) to T.M.E, a Marie Curie Initial Training Network 'SYMBIOMICS' to D.D. and M.H., and a European Research Council ERC StG EVOCHLAMY (281633) to M.H.

Author contributions

The project was conceived, designed and all analyses performed by D.D. and T.A.W. The manuscript was written by T.A.W. and D.D., with significant input and editing by M.H. and T.M.E.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

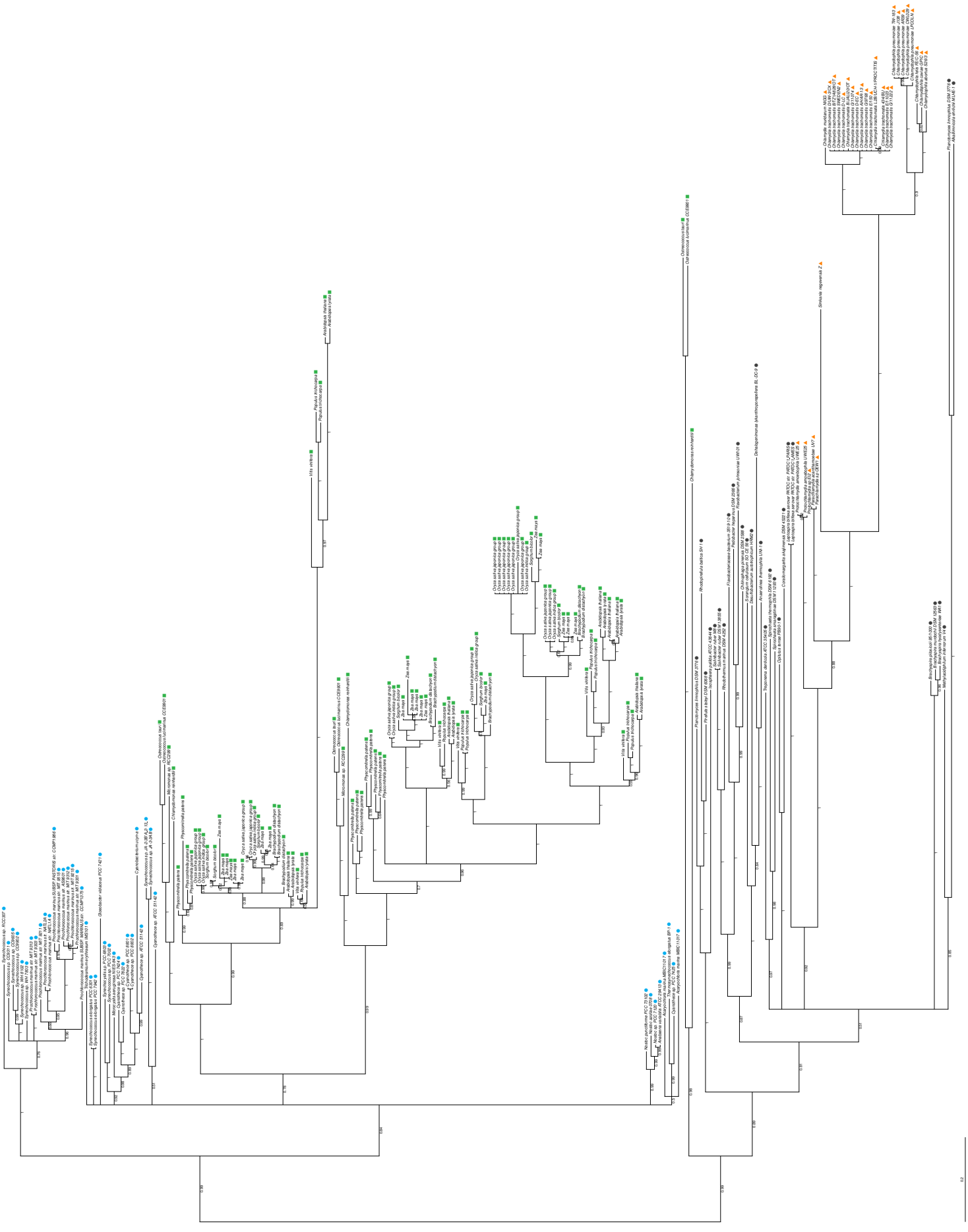
Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Domman, D. *et al.* Plastid establishment did not require a chlamydial partner. *Nat. Commun.* 6:6421 doi: 10.1038/ncomms7421 (2015).

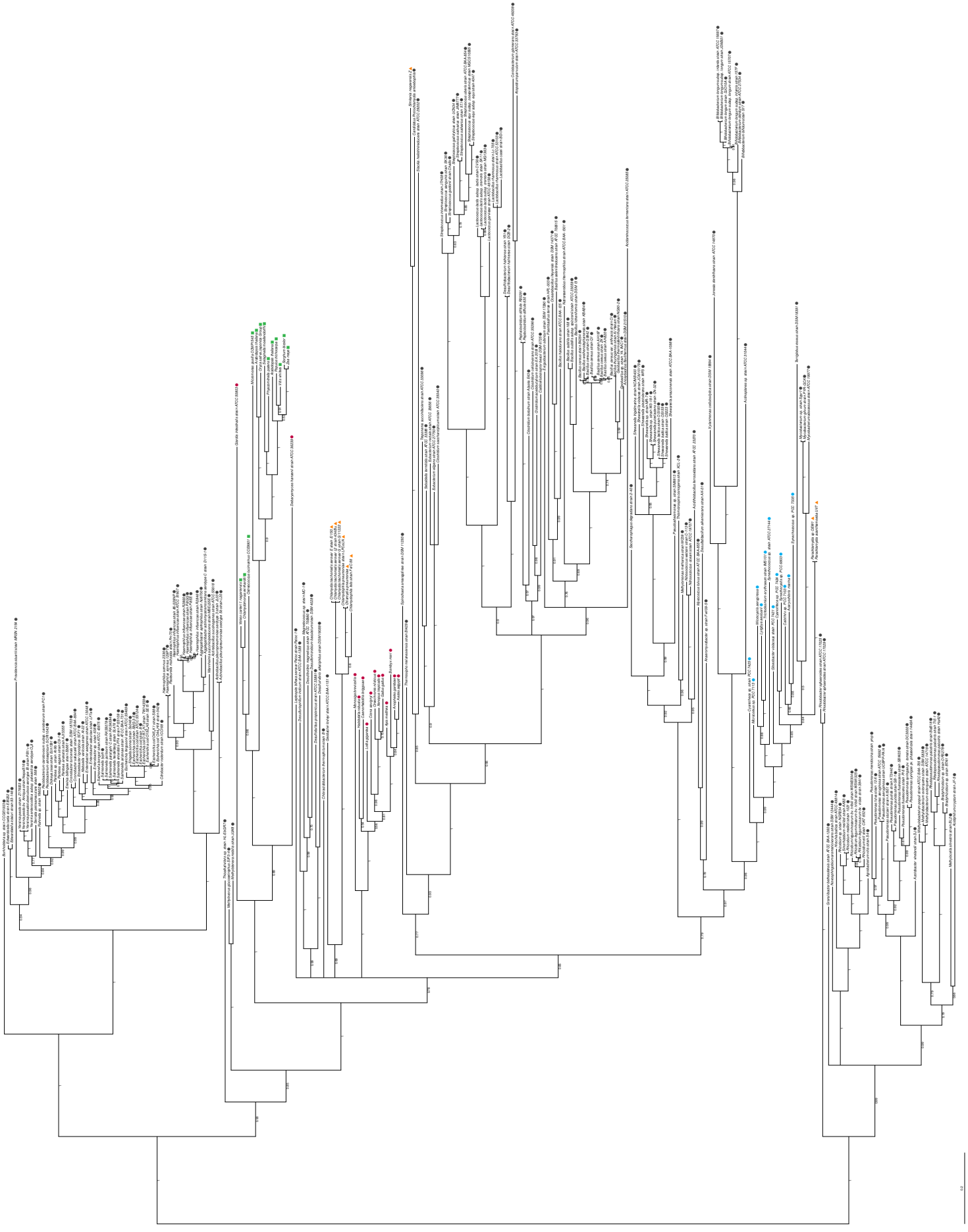
Plastid establishment did not require a chlamydial partner

Domman et al.

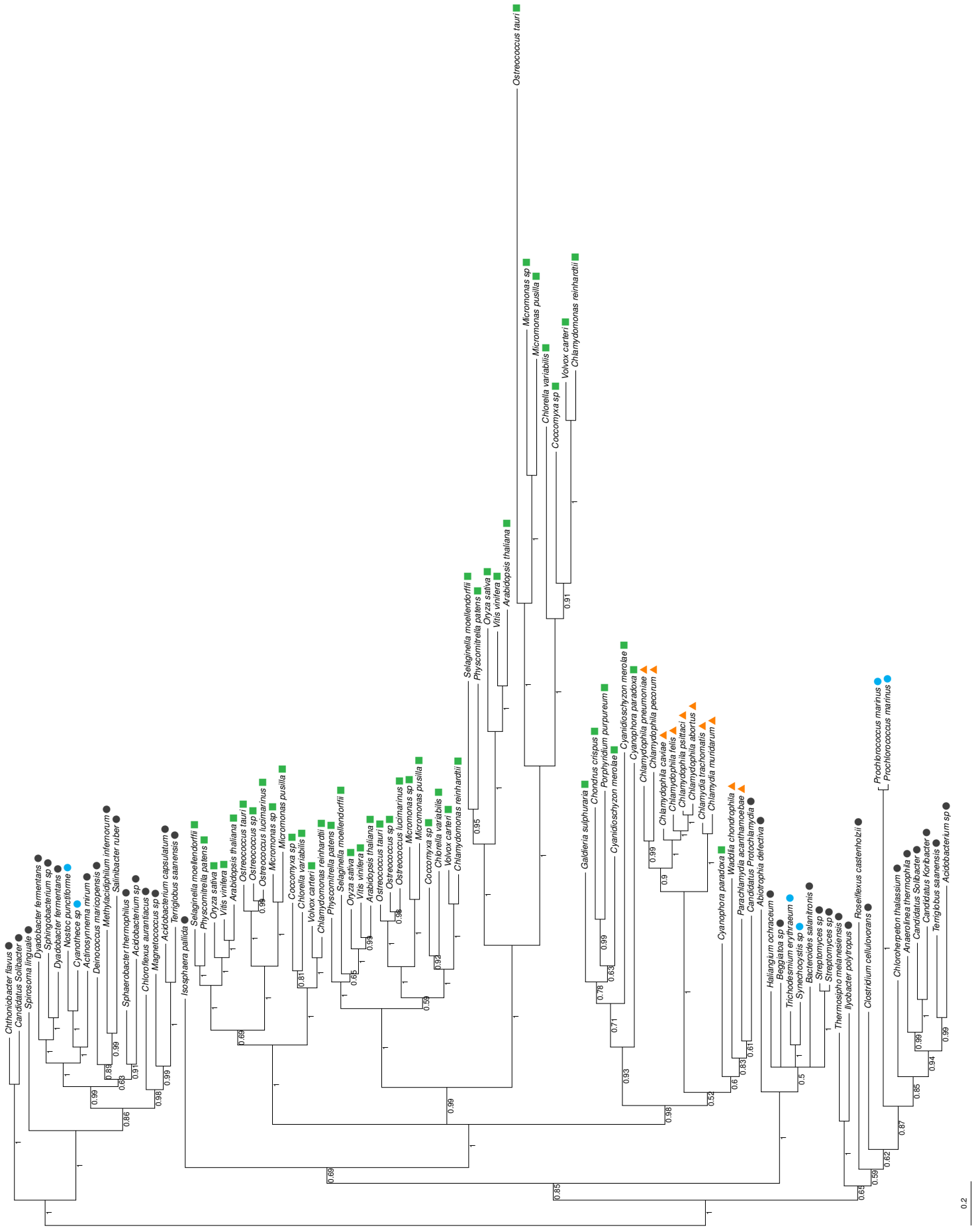
Supplementary Information



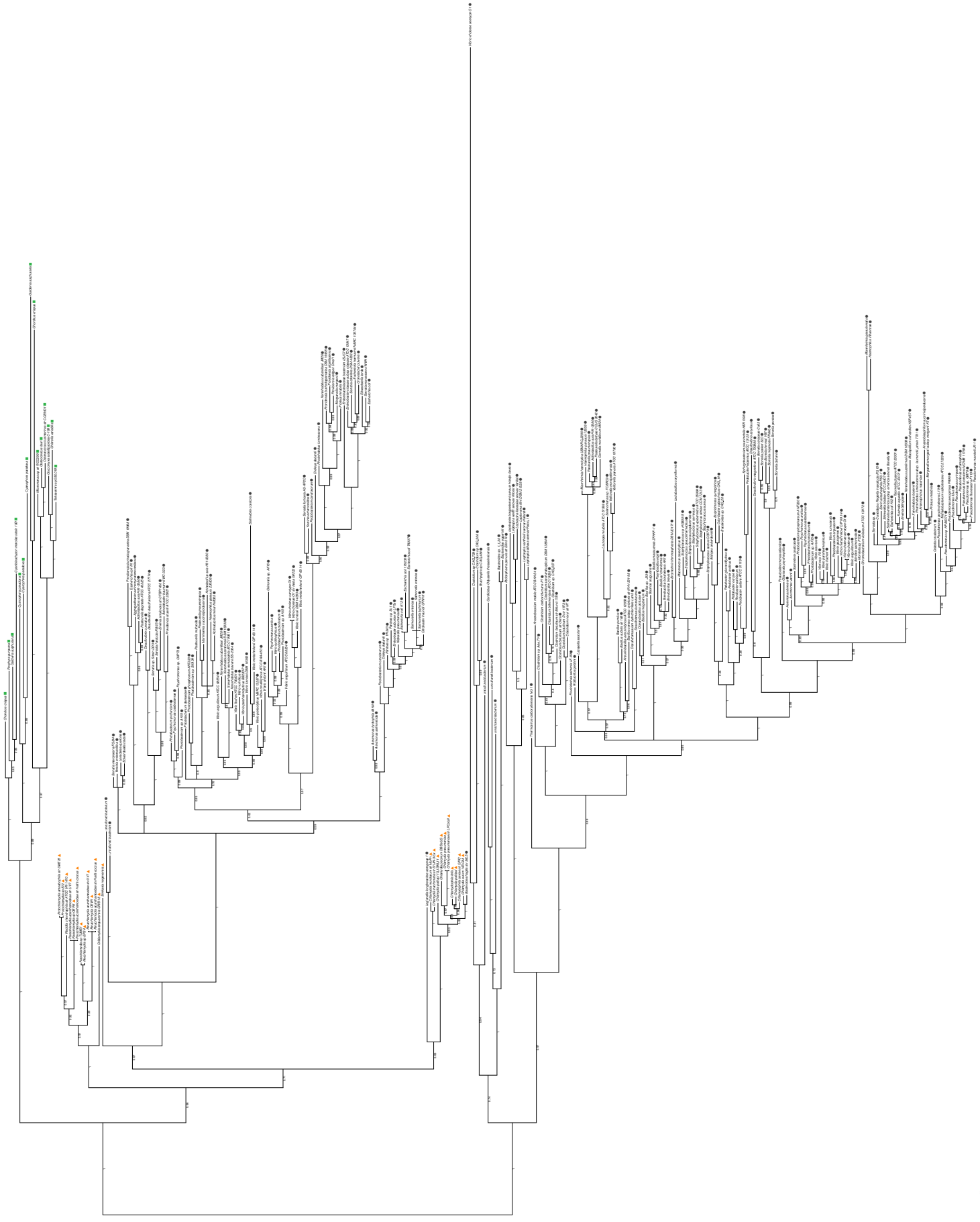
Supplementary Figure 1: **Phylogenetic analysis of the GlgC gene under the CAT+GTR model.** This is the complete tree upon which Fig. 3(a) is based. Archaeplastida sequences are denoted with green squares, *Chlamydiae* with orange triangles, *Cyanobacteria* with cyan circles, and other bacterial groups with black circles. Branch supports are Bayesian posterior probabilities, and branch lengths are proportional to the expected number of substitutions per site, as indicated by the scale bar.



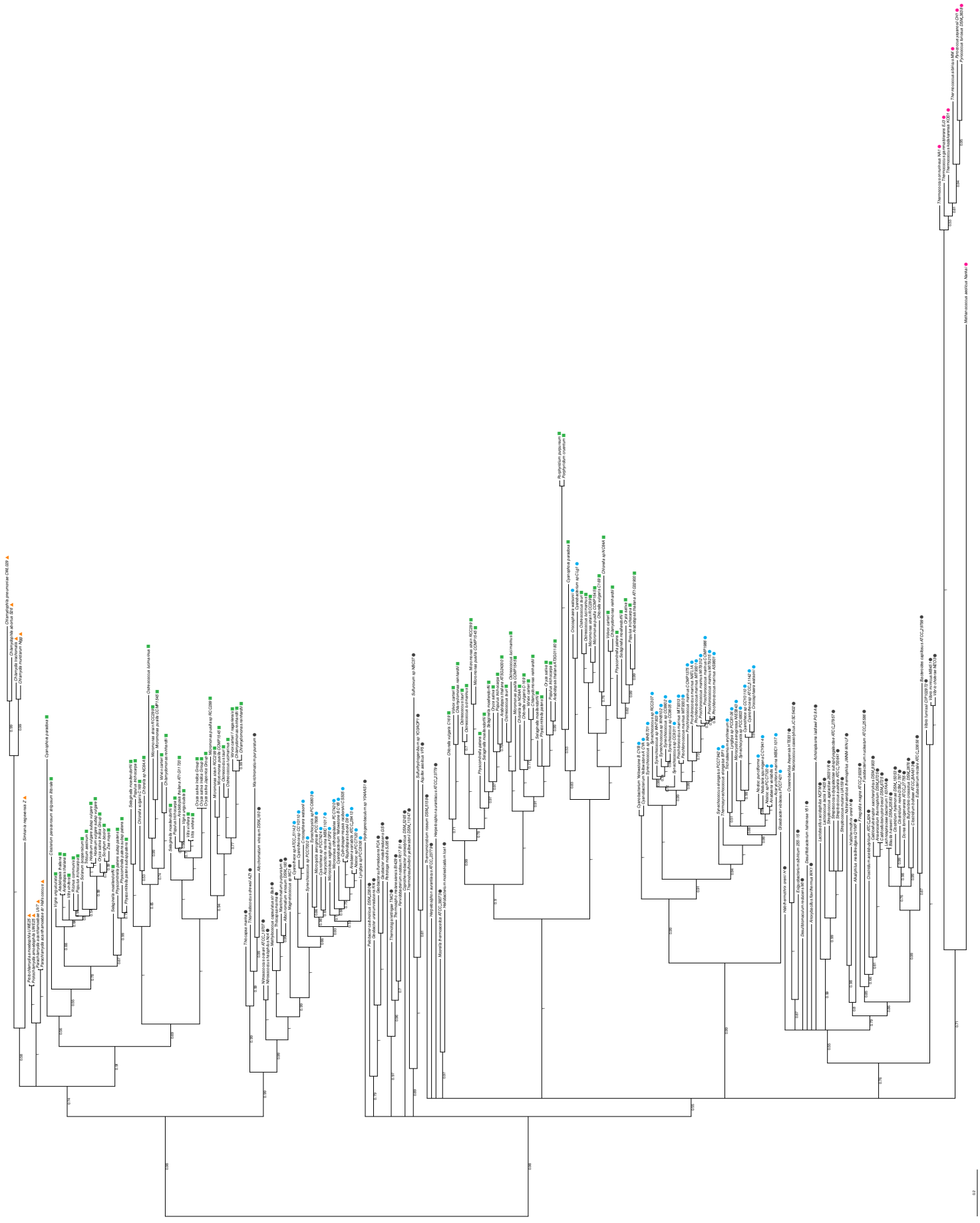
Supplementary Figure 2: **Phylogenetic analysis of the GlgP gene under the CAT+GTR model.** This is the complete tree upon which Fig. 3(b) is based. Archaeplastida sequences are denoted with green squares, other eukaryotes with purple circles, *Chlamydiae* with orange triangles, *Cyanobacteria* with cyan circles, and other bacterial groups with black circles. Branch supports are Bayesian posterior probabilities, and branch lengths are proportional to the expected number of substitutions per site, as indicated by the scale bar.


Chthoniobacter flavus
Candidatus Solibacter
Spirosoma linguale
Dyadobacter fermentans
Sphingobacterium sp
Dyadobacter fermentans
Nostoc punctiforme
Cyanothece sp
Actinosynnema mirum
Deinococcus maricopensis
Methylacidiphilum infernorum
Salinibacter ruber
Sphaerobacter thermophilus
Acidobacterium sp
Chloroflexus aurantiacus
Magnetococcus sp
Acidobacterium capsulatum
Terriglobus saanensis
Isosphaera pallida
Selaginella moellendorffii
Physcomitrella patens
Oryza sativa
Vitis vinifera
Arabidopsis thaliana
Ostreococcus tauri
Ostreococcus sp
Ostreococcus lucimarinus
Micromonas sp
Micromonas pusilla
Coccomyxa sp
Chlorella variabilis
Volvox carteri
Chlamydomonas reinhardtii
Physcomitrella patens
Selaginella moellendorffii
Oryza sativa
Vitis vinifera
Arabidopsis thaliana
Ostreococcus tauri
Ostreococcus sp
Ostreococcus lucimarinus
Micromonas sp
Micromonas pusilla
Coccomyxa sp
Chlorella variabilis
Volvox carteri
Chlamydomonas reinhardtii
Selaginella moellendorffii
Physcomitrella patens
Oryza sativa
Vitis vinifera
Arabidopsis thaliana
Ostreococcus tauri
Micromonas sp
Micromonas pusilla
Chlorella variabilis
Volvox carteri
Chlamydomonas reinhardtii
Ostreococcus tauri
Micromonas sp
Micromonas pusilla
Chlorella variabilis
Coccomyxa sp
Volvox carteri
Chlamydomonas reinhardtii
Galdieria sulphuraria
Chondrus crispus
Porphyridium purpureum
Cyanidioschyzon merolae
Cyanidioschyzon merolae
Cyanophora paradoxa
Chlamydophila pneumoniae
Chlamydophila pecorum
Chlamydophila caviae
Chlamydophila felis
Chlamydophila psittaci
Chlamydophila abortus
Chlamydia trachomatis
Chlamydia muridarum
Cyanophora paradoxa
Waddlia chondrophila
Parachlamydia acanthamoebae
Candidatus Protochlamydia
Abiotrophia defectiva
Halanaerobium ochraceum
Beggiatoa sp
Trichodesmium erythraeum
Synechocystis sp
Bacteroides salanitronis
Streptomyces sp
Streptomyces sp
Thermosipho melanesiensis
Ilyobacter polytropus
Roseiflexus castenholzii
Clostridium cellulovorans
Prochlorococcus marinus
Prochlorococcus marinus
Chloroherpeton thalassium
Anaerolinea thermophila
Candidatus Solibacter
Candidatus Koribacter
Terriglobus saanensis
Acidobacterium sp
0.2

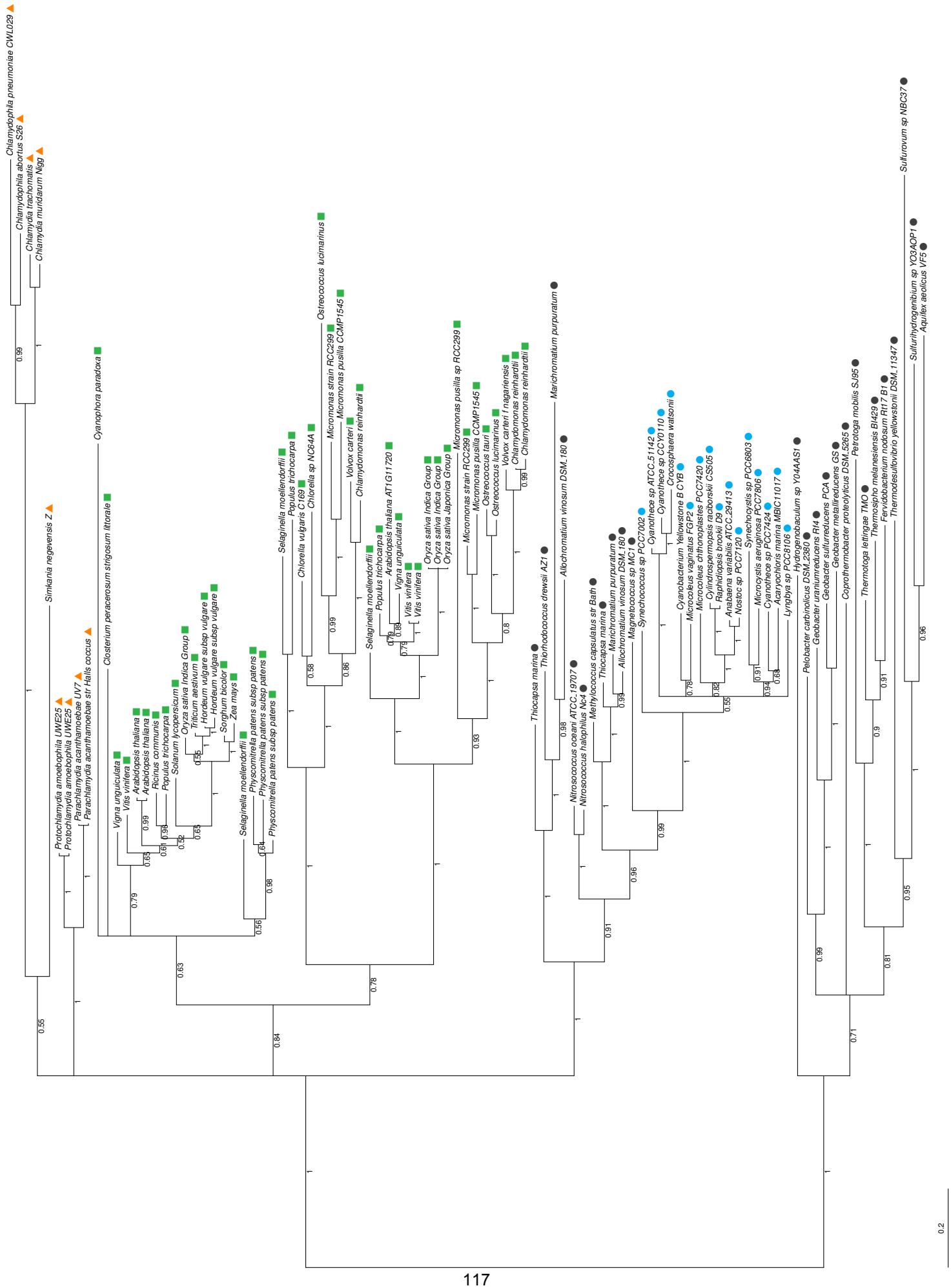
Supplementary Figure 3: **Phylogenetic analysis of the GlgX gene under the CAT+GTR model.** This is the complete tree upon which Fig. 3(c) is based. Archaeplastida sequences are denoted with green squares, *Chlamydiae* with orange triangles, *Cyanobacteria* with cyan circles, and other bacterial groups with black circles. Branch supports are Bayesian posterior probabilities, and branch lengths are proportional to the expected number of substitutions per site, as indicated by the scale bar.



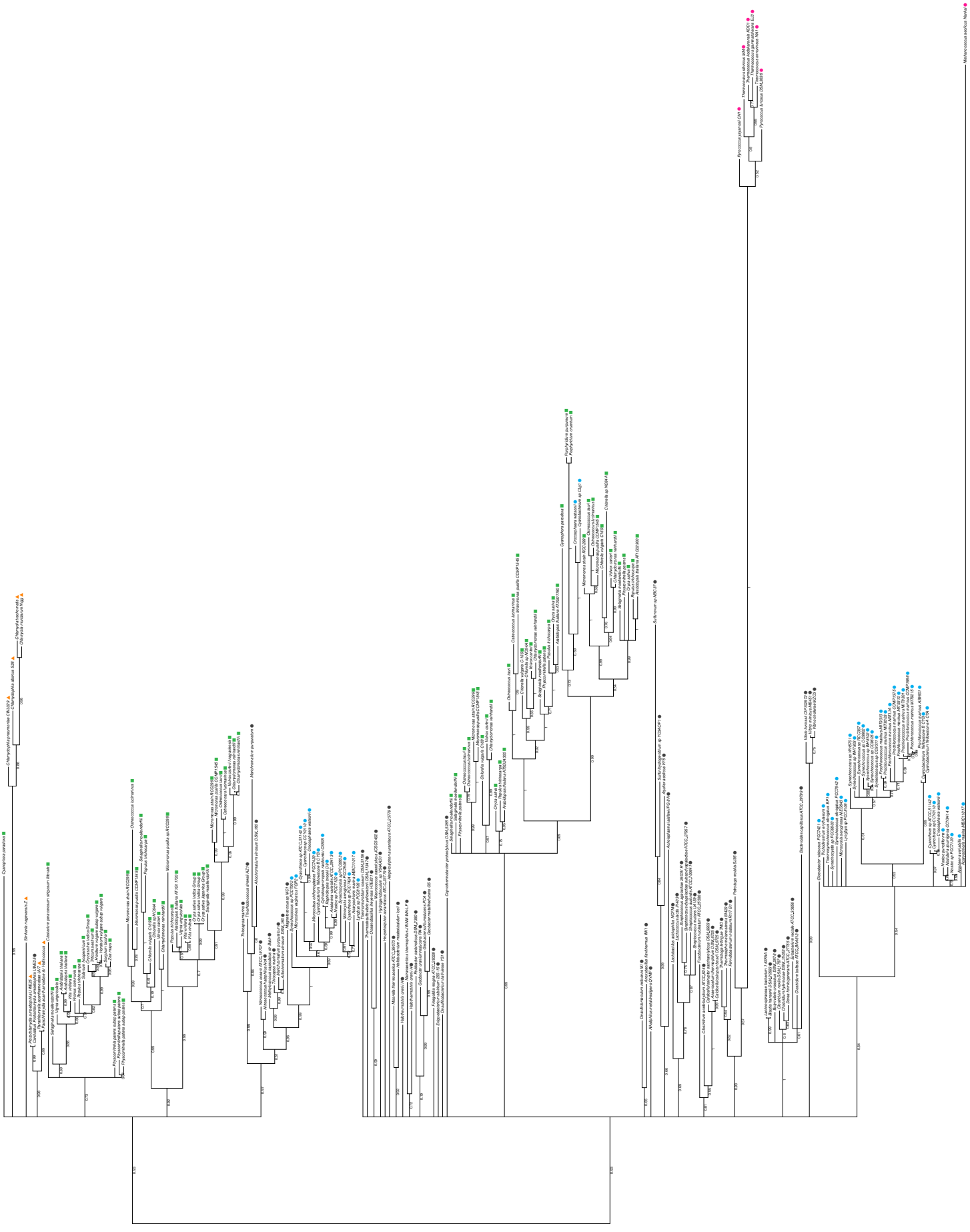
Supplementary Figure 4: **Phylogenetic analysis of the UhpC gene under the CAT+GTR model.** This is the complete tree upon which Fig. 3(d) is based. Archaeplastida sequences are denoted with green squares, *Chlamydiae* with orange triangles, and other bacterial groups with black circles. Branch supports are Bayesian posterior probabilities, and branch lengths are proportional to the expected number of substitutions per site, as indicated by the scale bar.



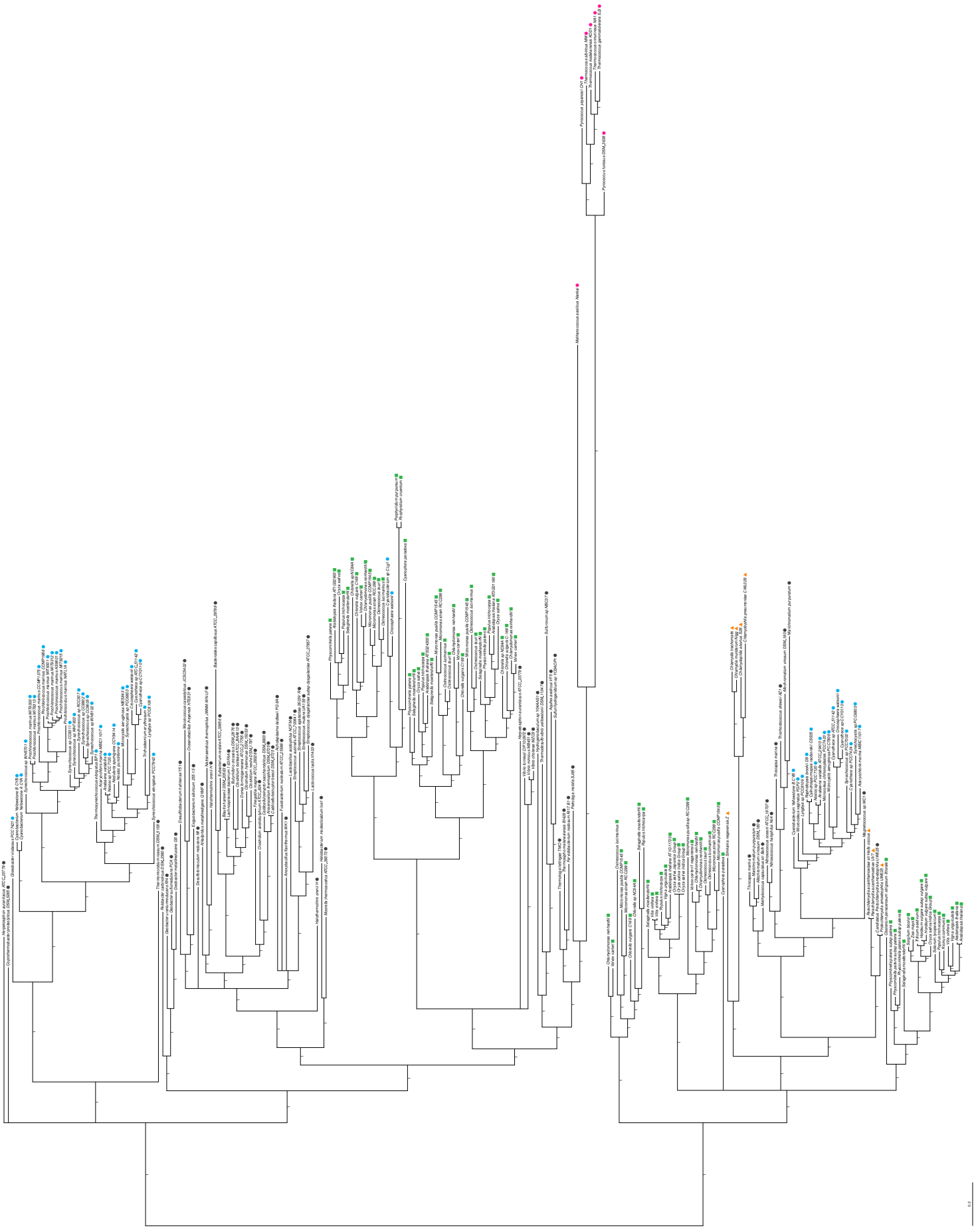
Supplementary Figure 5: **Phylogenetic analysis of the GlgA gene under the CAT+GTR model.** This is the complete tree upon which Fig. 4(a) is based. Archaeplastida sequences are denoted with green squares, *Chlamydiae* with orange triangles, *Cyanobacteria* with cyan circles, other bacterial groups with black circles, and Archaea with pink circles. Branch supports are Bayesian posterior probabilities, and branch lengths are proportional to the expected number of substitutions per site, as indicated by the scale bar.


Chlamydophila pneumoniae CWL029
Chlamydophila abortus S26
Chlamydia trachomatis
Chlamydia muridarum Nigg
Simkania negevensis Z
Protochlamydia amoebophila UWE25
Protochlamydia amoebophila UWE25
Parachlamydia acanthamoebae UV7
Parachlamydia acanthamoebae str Hall's coccus
Cyanophora paradoxa
Closterium peracerosum strigosum littorale
Vigna unguiculata
Vitis vinifera
Arabidopsis thaliana
Arabidopsis thaliana
Ricinus communis
Populus trichocarpa
Solanum lycopersicum
Oryza sativa Indica Group
Triticum aestivum
Hordeum vulgare subsp vulgare
Hordeum vulgare subsp vulgare
Sorghum bicolor
Zea mays
Selaginella moellendorffii
Physcomitrella patens subsp patens
Physcomitrella patens subsp patens
Physcomitrella patens subsp patens
Selaginella moellendorffii
Populus trichocarpa
Chlorella vulgaris C169
Chlorella sp NC64A
Ostreococcus lucimarinus
Micromonas strain RCC299
Micromonas pusilla CCMP1545
Volvox carteri
Chlamydomonas reinhardtii
Selaginella moellendorffii
Populus trichocarpa
Arabidopsis thaliana AT1G11720
Vigna unguiculata
Vitis vinifera
Vitis vinifera
Oryza sativa Indica Group
Oryza sativa Indica Group
Oryza sativa Japonica Group
Micromonas pusilla sp RCC299
Micromonas strain RCC299
Micromonas pusilla CCMP1545
Ostreococcus tauri
Ostreococcus lucimarinus
Volvox carteri f nagariensis
Chlamydomonas reinhardtii
Chlamydomonas reinhardtii
Thiocapsa marina
Thiorhodococcus drewsii AZ1
Marichromatium purpuratum
Allochromatium vinosum DSM 180
Nitrosococcus oceani ATCC 19707
Nitrosococcus halophilus Nc4
Methylococcus capsulatus str Bath
Thiocapsa marina
Marichromatium purpuratum
Allochromatium vinosum DSM 180
Magnetococcus sp MC1
Synechococcus sp PCC7002
Cyanothece sp ATCC 51142
Cyanothece sp CCY0110
Crocosphaera watsonii
Cyanobacterium Yellowstone B CYB
Microcoleus vaginatus FGP2
Microcoleus chthonoplastes PCC7420
Cylindrospermopsis raciborskii CS505
Raphidiopsis brookii D9
Anabaena variabilis ATCC 29413
Nostoc sp PCC7120
Synechocystis sp PCC6803
Microcystis aeruginosa PCC7806
Cyanothece sp PCC7424
Acaryochloris marina MBIC11017
Lyngbya sp PCC8106
Hydrogenobaculum sp Y04AAS1
Pelobacter carbinolicus DSM 2380
Geobacter uraniumreducens Rf4
Geobacter sulfurreducens PCA
Geobacter metallireducens GS
Coprothermobacter proteolyticus DSM 5265
Petrotoga mobilis SJ95
Thermotoga lettingae TMO
Thermosipho melanesiensis BI429
Fervidobacterium nodosum Rt17 B1
Thermodesulfovibrio yellowstonii DSM 11347
Sulfurovum sp NBC37
Sulfurihydrogenibium sp YO3AOP1
Aquifex aeolicus VF5
0.2

Supplementary Figure 6: **Phylogenetic analysis of the GlgA gene with only the closest outgroup clade under the CAT+GTR model.** This is the complete tree upon which Fig. 4(b) is based. Archaeplastida sequences are denoted with green squares, other eukaryotes with purple circles, *Chlamydiae* with orange triangles, *Cyanobacteria* with cyan circles, and other bacterial groups with black circles. Branch supports are Bayesian posterior probabilities, and branch lengths are proportional to the expected number of substitutions per site, as indicated by the scale bar.



Supplementary Figure 7: **Phylogenetic analysis of the Dayhoff-recoded GlgA alignment under the CAT+GTR model.** This is the complete tree upon which Fig. 4(c) is based. Archaeplastida sequences are denoted with green squares, *Chlamydiae* with orange triangles, *Cyanobacteria* with cyan circles, other bacterial groups with black circles, and Archaea with pink circles. Branch supports are Bayesian posterior probabilities, and branch lengths are proportional to the expected number of substitutions per site, as indicated by the scale bar.



Supplementary Figure 8: **Phylogenetic analysis of the GlgA gene under the nonstationary CAT+BP model.** This is the complete tree upon which Fig. 4(d) is based. Archaeplastida sequences are denoted with green squares, other eukaryotes with purple circles, *Chlamydiae* with orange triangles, *Cyanobacteria* with cyan circles, other bacterial groups with black circles, and Archaea with pink circles. Branch supports are Bayesian posterior probabilities, and branch lengths are proportional to the expected number of substitutions per site, as indicated by the scale bar.

Supplementary Table 1: Posterior predictive simulations of compositional heterogeneity for the single gene alignments analyzed in this study.

	Across-site composition			Across-branch composition		
	Observed	Predicted	*P*-value	Observed	Predicted	*P*-value
GlgC						
LG	7.78	9.71 +/- 0.1	**0**	0.0016	0.0014	0.069
CAT+GTR	7.78	7.86 +/- 0.19	0.33	0.004	0.0047 +/- 0.0009	0.63
GlgP						
LG	8.51	11.61 +/- 0.09	**0**	0.008	0.0026	**0**
CAT+GTR	8.51	8.91 +/- 0.09	**0**	0.008	0.0049 +/- 0.00057	**0**
GlgX						
LG	9.43	11.17 +/- 0.09	**0**	0.015	0.004	**0**
CAT+GTR	9.43	9.54 +/- 0.16697	0.24	0.015	0.0059 +/- 0.001	**0**
UhpC						
LG	8.465	9.84 +/- 0.14	**0**	0.009	0.0069 +/- 0.0007	**0.008**
CAT+GTR	8.465	8.41 +/- 0.09	0.71	0.009	0.0068 +/- 0.0008	**0.008**
GlgA						
LG	10.97	13.64 +/- 0.11	**0**	0.016	0.0083	**0.002**
CAT+GTR	10.97	11.49 +/- 0.16	**0**	0.016	0.009	**0.017**
CAT+GTR (closest outgroup)	8.253	8.414 +/- 0.2	0.21	0.017	0.008 +/- 0.001	**0**
CAT+GTR +Dayhoff	4.409	4.408 +/- 0.04	0.49	0.012	0.008	0.1

Chapter VI

Concluding discussion

Conclusion

The members of the *Chlamydiae* are truly fascinating and remarkable microbes. Few bacterial groups, indeed, could boast of such far-reaching implications as the *Chlamydiae*. From the causative agent of the most prevalent STI in the world, to the potential facilitator of all plant life (which then also would have contributed to human evolution), the *Chlamydiae* are an incredible group of bacteria. The prominent goal of this thesis was to explore the evolutionary history of this phylum. I accomplished this goal using a variety of contexts, such as exploring gene gains and losses, how the regulatory system is organized, and ancient lateral gene transfers.

Gene family evolution

My work on the evolution of gene families within the *Chlamydiae* has large implications, not only in the chlamydia field, but perhaps for how we understand the evolution of intracellular bacteria in general. Bacteria living inside eukaryotic host cells are of utmost importance both as pathogens and symbionts. A hallmark of this life style is a small genome size, typically devoid of redundancy. We discovered several surprisingly large gene families in members of the *Chlamydiae*, which corresponded to lineage specific expansion events. We showed that the gene families encode proteins that are targeted into the host cell and have the potential to subvert essential cellular pathways in eukaryotes. The evolution of these gene families mirrors processes observed in eukaryotes and might represent a previously undescribed way by which genetic diversity in intracellular organisms is generated.

This study opens the door for many follow up questions to how widespread this mode of evolution is among bacteria and the odd connection of this mode of evolution within plant genomes of these same gene families. In the study we didn't consider the initial origin of these expanded gene families, but clearly there was HGT into each of the lineages with expansions. Exploring the origins of these genes is an obvious next step for this study, however it is not without difficulty as these repeat proteins are notoriously difficult to align properly thus obfuscating phylogenetic signal. An alternative to a traditional phylogenetic analysis that may lend insight is the use of a sequence similarity network, in which one clusters the BLAST results in a network graph to show relationships (Mashiyama et al. 2014). The other obvious next step is to characterize the binding partners of the chlamydial F-box and BTB-box proteins experimentally. A possible direction is to develop a pull-down assay in which you use selected

F-box/BTB-box candidates as bait for amoeba lysate. The main challenge here is that, as we speculate in the paper, these expanded families may correspond to either a large number of targets within a narrow host range or a small number of targets within a large host range. If the situation is the latter, the choice of bait protein is fundamental to finding the interacting partner within the Acathamoeba system. For the BTB-box proteins in *Protochlamydia amoebophilus* UWE25, the current work on the transcriptome can guide the selection of candidates for those that are expressed.

Although our publication mainly described the evolutionary history of the selected expanded gene families, this should be expanded to model all gene family histories across the phylum, perhaps even the entire *Planctomycetes-Verrucomicrobium-Chlamydiae* superphylum. There are key questions, such as the magnitude of HGT and ancestral genome reconstruction of the last chlamydial ancestor, that are well within grasp given the data we already have produced. Application of "species-tree" aware gene-tree programs have great potential for answering both questions, as well as resolving some of the uncertainties still in the chlamydial species tree, such as the correct position for the *Simkaniacea* and *Chlamydia ibidis* (Szöllősi et al. 2015).

Gene regulatory networks

One of the major questions in chlamydial biology is how these organisms temporally regulate the developmental cycle. Aside from the basic research interest to better understand these organisms, there are also clinical aspects as halting the cycle may be of pharmacological interest. In this vein, the co-regulation networks we provide for the research community is a large step forward towards understanding at a global level how chlamydial regulatory schemes are organized within and between chlamydial organisms. This is an particularly exciting time for this avenue of research since work in the Tan group has recently published the first ChIP-Seq study for *Chlamydia trachomatis* examining the binding sites of the heat shock response regulator HrcA (Hanson and Tan 2015). While our network approach has been a huge leap forward, more ChIP-Seq studies are desperately needed to truly elucidate the regulatory networks within these organisms.

The loss of the alternative sigma factor, σ^{28}, in the environmental chlamydiae is a particular interesting question that warrants further investigation. Since several of the genes that are shown to be σ^{28}-regulated in the *Chlamydiaceae* are present in the genomes of environmental

chlamydia an obvious experiment is to determine what factor is now performing this role in these organisms. Although not discussed in the thesis, I had a foray into the lab in trying to set up a biotin mediated pull-down assay with the promoter region from the *Simkania negevensis* tail specific protease (*tsp*) promoter region. Given sufficient time to work out the kinks, I believe an approach like this will work at identifying proteins binding to these "σ^{28}-regulated" genes still present in the environmental chlamydia. An alternative approach to this method is to make a prediction about which regulator is performing the function, likely σ^{66} in this case, and perform either ChIP-Seq or DNAse protection experiments with the candidate promoter regions.

One of the major findings from our network analysis was the incorporation of hypothetical or uncharacterized proteins into functionally defined sub-clusters. This has obvious implications for those interested in the virulence gene cluster in which we incorporated hypotheticals, but really this can be extended to any functional category one is interested in. Another direction to take this research is looking at the co-evolution of the binding sites themselves and the transcription factor. This is particularly interesting for genes that have recently been acquired via HGT, where they must integrate into the regulatory network (Price et al. 2008).

***Chlamydiae* and the ancient gene transfer to the Archaeplastida**

The running comment for this study was that it was "negative results positively published." Despite our strong negation of a chlamydial role is establishing the plastid endosymbiosis, I do find this hypothesis very intriguing and genuinely commend Ball and colleagues (Ball et al. 2013) for putting it forth. Our word is certainly not the last on this topic, as already seen in a recent review (Soucy et al. 2015), and it shouldn't be for that matter. Open scientific dialog about these questions should be encouraged and welcome, and hopefully not bruise egos or polarize the community. There is no doubt that there is a certain affiliation of chlamydial sequences with those of members of the Archaeplastida. Whether or not these represent a phylogenetic artifact or ancient HGT events between cyanobacteria/eukaryotes is still the major open question. Our present study only considered those genes pertinent to the claims for the "ménage à trois", and not the other 60 or so genes. A Masters student and I are currently investigating a subset of these 60 genes to see how the overall trend is when we apply the better fitting CAT models to these data. This is rather arduous work due to the computational burdens and large amount of compositional heterogeneity exhibited by many of the datasets, but the results should be very interesting and additive to the discussion at hand. Although I don't

necessarily believe that ancient *Chlamydiae* facilitated the plastid establishment, I am completely open to there having been some HGT between the organisms involved. It very well may be that our new analysis supports HGT between Chlamydiae and Archaeplastida.

In order to robustly test this hypothesis I think we need to pare down the dataset to a reasonable representative number of taxa. With this smaller dataset we can really test differing hypotheses on a large number of genes within a reasonable amount of time. Additionally, in order to use the models implemented in the P4 package, such as the ability to model across species compositional heterogeneity, smaller data is a requisite. Furthermore, I have an odd feeling that large DNA-viruses may be playing an interesting role in HGT between chlamydia and plants. A recent study showed past infection of plants with these large giant viruses (Maumus et al. 2014), and given that these same type of viruses infect amoeba (Boyer et al. 2009) the link doesn't seem so far fetched. Perhaps the viruses are shuttling genes around the domains of life considered here. A way to test this is to create phylogenetically informed networks, in which one would be able to see the gene transfer "highways" between the amoeba associated organisms, eukaryotes (in this case plants in particular), and the giant viruses, as compared to an outgroup of non-amoeba associated bacteria (such as the other members of the PVC-superphylum).

General parting thoughts

More generally there is a real need to go after some of the more basic ecological questions involving the *Chlamydiae*. A friend and former PhD student in the Horn group made a nice stride in this direction in examining the diversity of chlamydial sequences in amplicon datasets (Lagkouvardos et al. 2014), but in order to interpret our genomic data correctly we must shed more light on the ecology of these organisms. Fundamental questions like what is the diversity within and between populations of *Chlamydiae* in an environmental sample and some attempt at estimating host range would be tremendous steps forward to fully understanding these organisms. One potential way forward that is particularly exciting is the use of "reverse ecology", in which population genomic data is used to make ecological predictions (Shapiro and Polz 2014), which can then be further experimentally investigated. Another ecological question of some interest is how amoeba-associated organisms affect the population structure of their hosts. For those organisms that lyse their host, the situation is akin to top-down selection via phages (Cordero and Polz 2014). As microbial eukaryote grazing is a major source of bacterial predation in natural environments it would be of note to investigate the dynamics of interaction

network in light of the amoeba-associated organisms. The topics and questions raised here are not easily answered, to be sure, but are perhaps quite vital in the future to understand all facets of chlamydial biology.

To end, I must say that it has been a tremendously enjoyable experience to delve deep into the evolutionary history of, in my humble opinion, one of the most fascinating groups of bacteria. The continual expansion of genetic tools for members of the *Chlamydiaceae* and their eventual application to environmental chlamydia is certainly opening new frontiers of research options that have never before been available to this community. Combine this with the ease and cost effectiveness of sequencing, a golden age of chlamydial biology is certainly on the horizon. I'm quite looking forward to see how the field of chlamydial biology unfolds and I am thankful that I have been able to contribute to the field in helping to understand the evolutionary history of these most remarkable *Chlamydiae*.

References

Ball SG, Subtil A, Bhattacharya D, Moustafa A, Weber APM, Gehre L, Colleoni C, Arias M-C, Cenci U, Dauvillée D. 2013. Metabolic Effectors Secreted by Bacterial Pathogens: Essential Facilitators of Plastid Endosymbiosis? Plant Cell Online 25:7–21.

Boyer M, Yutin N, Pagnier I, Barrassi L, Fournous G, Espinosa L, Robert C, Azza S, Sun S, Rossmann MG, et al. 2009. Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. Proc. Natl. Acad. Sci.:pnas.0911354106.

Cordero OX, Polz MF. 2014. Explaining microbial genomic diversity in light of evolutionary ecology. Nat. Rev. Microbiol. 12:263–273.

Hanson BR, Tan M. 2015. Transcriptional regulation of the Chlamydia heat shock stress response in an intracellular infection. Mol. Microbiol.:n/a – n/a.

Lagkouvardos I, Weinmaier T, Lauro FM, Cavicchioli R, Rattei T, Horn M. 2014. Integrating metagenomic and amplicon databases to resolve the phylogenetic and ecological diversity of the Chlamydiae. ISME J. 8:115–125.

Mashiyama ST, Malabanan MM, Akiva E, Bhosle R, Branch MC, Hillerich B, Jagessar K, Kim J, Patskovsky Y, Seidel RD, et al. 2014. Large-Scale Determination of Sequence, Structure, and Function Relationships in Cytosolic Glutathione Transferases across the Biosphere. PLoS Biol 12:e1001843.

Maumus F, Epert A, Nogué F, Blanc G. 2014. Plant genomes enclose footprints of past infections by giant virus relatives. Nat. Commun. [Internet] 5. Available from: http://www.nature.com/ncomms/2014/140627/ncomms5268/full/ncomms5268.html

Price M, Dehal P, Arkin A. 2008. Horizontal gene transfer and the evolution of transcriptional regulation in Escherichia coli. Genome Biol. 9:R4.

Shapiro BJ, Polz MF. 2014. Ordering microbial diversity into ecologically and genetically cohesive units. Trends Microbiol. 22:235–247.

Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: building the web of life. Nat. Rev. Genet. 16:472–482.

Szöllősi GJ, Tannier E, Daubin V, Boussau B. 2015. The Inference of Gene Trees with Species Trees. Syst. Biol. 64:e42–e62.

Chapter VII

Abstract / Zusammenfassung

Abstract

The bacterial phylum *Chlamydiae* is comprised of obligate intracellular parasites with direct relevance to human and animal health. The human pathogen *Chlamydia trachomatis* affects nearly 84 million people globally, and represents the leading cause of preventable blindness in the world. While much research focus has naturally been on these pathogens, they represent only a small fraction of the diversity of chlamydial organisms. Outside of the family *Chlamydiaceae*, which are the family pertaining to the known pathogens, lies a tremendous diversity of chlamydial organisms and they are associated with a fantastic array of eukaryotic hosts, ranging from free-living protists, enigmatic marine worms, fish, and insects. This work sought to use genomics to uncover the evolutionary history of these amazing microbes. Firstly, we studied the evolution of gene content throughout evolutionary time, discovering that certain chlamydial lineages have had tremendous genomic expansions as a result of gene duplications. The specific mode of evolution is rather unique for host-associated microbes and may represent a novel way in which these organisms generate genomic diversity. Secondly, we used comparative genomics to elucidate predicted regulatory networks for all fully sequenced chlamydial organisms. This work provides the first phylum wide examination as to how chlamydial organisms regulate gene expression and shed much needed insight on the conservation and differences in gene regulation between chlamydial organisms. Thirdly, we robustly tested a leading evolutionary hypothesis concerning the role of ancient chlamydiae in the establishment of the plastid endosymbiosis. Using more accurate complex phylogenetic models, we showed that there is little evidence to support the role of chlamydiae in the plastid symbiotic capture. The scope and breadth of evolutionary genomics analyses presented in this work have far reaching implications within the field of chlamydial biology and more generally furthers our understanding of evolutionary forces acting upon host-associated microbes.

Zusammenfassung

Das bakterielle Phylum *Chlamydiae* setzt sich aus intrazellulären Parasiten zusammen, die eine direkte Relevanz für die menschliche und tierische Gesundheit haben. Annähernd 84 Millionen Menschen leiden unter den Folgen einer Infektion mit *Chlamydia trachomatis*, der führenden Ursache für vermeidbare Blindheit. Während sich selbstverständlich viel Forschung auf diese Pathogene konzentriert, repräsentieren sie eigentlich nur einen kleinen Teil der Chlamydien-artigen Organismen. Ausserhalb der Familie der *Chlamydiaceae*, welche die wohlbekannten Pathogene beinhält, liegt eine gigantische Vielfalt von Chlamydien-artigen Organismen, die wiederum eine fantastische Auswahl an eukaryotischen Wirten ihr eigen nennen, die sich von frei lebenden Protisten über kaum studierte marine Würmer, Fische und Insekten erstreckt. Diese Arbeit versucht mit Hilfe von Genomik die Entstehungsgeschichte dieser erstaunlichen Mikroben freizulegen. Zuerst untersuchten wir die Evolution von genetischem Inhalt über evolutionäre Zeit und fanden dabei massive genomische Ausdehnung als ein Resultat von Genduplikation. Der Evolutionsmodus für wirtsassoziierte Mikroben ist einigermaßen einzigartig und könnte einen neuen Weg widerspiegeln in dem Organismen genomische Diversität generieren. Zum Zweiten wandten wir vergleichende Genomik an um die vorhergesagten regulatorischen Netzwerke aller zur Gänze sequenzierten chlamydien-artigen Organismen aufzuklären. Diese Arbeit stellt die erste Phylum-weite Untersuchung dar, die Einsicht in die Regulation der Genexpression gibt und notwendiges Licht auf die im Dunkeln liegende Konservierung und Differenzierung der Genregulation zwischen den chlamydien-artigen Organismen wirft. Zum Dritten testeten wir mit robusten Methoden eine führende evolutionäre Hypothese bezüglicher der Rolle von ursprünglichen Chlamydien in der Ausbildung der Plastidenendosymbiose. Mit Hilfe von genaueren, komplexen phylogenetischen Modellen konnten wir zeigen, dass es nur wenige unterstützende Hinweise für eine Rolle der Chlamydien in der symbiotischen Vereinnahmung der Plastiden gibt. Das Ausmaß und die Breite der angewandten evolutionären genomischen Analysen in dieser Arbeit haben weitgreifende Implikationen für das Feld der Chlamydienbiologie und erweitern im Allgemeinen unser Verständnis der evolutionären Kräfte, die auf wirtsassoziierte Mikroorganismen wirken.

Appendix

Acknowledgements
&
Curriculum Vitae

Acknowledgements

My graduate career seems to have followed the path of working with the rather obscure members of the PVC superphylum. I owe many thanks to many people for this particularly great journey. First and foremost I need to thank my wife for the willingness to move across the world for the sake of environmental chlamydia. If that isn't an act of love, I don't know what is. This has truly been a terrific adWIENture. Secondly, this PhD would not have been possible without the excellent friendship and guidance offered by my advisor Matthias Horn. Being under your tutelage has truly been a pleasure. At least we know that we are working with the real organisms of interest!

I would like to thank my collaboration partners; especially Tom Williams and Martin Embley for all of their help in moving my science forward both personally and professionally. It has also been a great pleasure to work alongside all of the SYMBIOMICS crew over the last few years. I enjoyed our 'stag-party' through Europe, and even in the US. I wish everyone all the best in the future!

I want to thank all of the members of the Symbiosis group past and present during my stay at DoME, especially Allen, Astrid, Frederick, Illias, Karen, Lena, Stephan, Nadia, and Paul. It has been an absolute pleasure being your friend and colleague. Thanks to Florian Moeller for being a good friend and letting me always steal coffee. Also thanks to everyone else who has been in the Oval Office during my time here. We have had quite a bit of fun… especially when wasps fly in the window. Thanks for saving me from these viscous beasts. Lastly, thanks to all of DOME for the great opportunity to work with such talented and driven individuals. Now its time to say "babatschi"!

DARYL DOMMAN

EMAIL ddomman@gmail.com

NATIONALITY USA

PHONE +43 1 4277 76621

Division of Microbial Ecology
University of Vienna
Althanstrasse 14
1090 Vienna
Austria

PROJECT INTERESTS

I enjoy using genomics to address fundamental questions about host-microbe interactions and the evolutionary mechanisms driving such processes.

PROGRAMMING SKILLS

PYTHON ●●●●●
R ●●●●●
PERL ●●●●●
UNIX ●●●●●
HTML/CSS ●●●○○
JAVA SCRIPT ●●●●○

GENOMICS

PHYLOGENETICS ●●●●●
GENOME ANALYSIS ●●●●●
METAGENOMICS ●●●●●
MOLECULAR EVOLUTION ●●●●●

MOLECULAR BIOLOGY

CLONING/SEQUENCING ●●●●●
GENETICS ●●●●○
PROTEIN ANALYSIS ●●●○○
CELL CULTURE ●●○○○
MICROSCOPY ●●○○○

EDUCATION

PhD in Biology (expected October 2015)
University of Vienna, Vienna, Austria
Matthias Horn, Advisor
2011- 2015

Masters in Molecular Biology
University of Wyoming, Laramie, Wyoming, USA
Naomi Ward, Advisor
2009-2011

Bachelors in Molecular Biology and Microbiology
University of Wyoming, Laramie, Wyoming, USA
2005-2009

FELLOWSHIPS

Marie Curie PhD Fellowship: ITN Symbiomics
Research stay with:
T. Martin Embley
Newcastle University, Newcastle, United Kingdom
July 2014

PUBLICATIONS

Domman, D., Horn, M., Embley, TM., and Williams, TA. (2015). Plastid establishment did not require a chlamydial partner. Nature Communications 6. doi:10.1038/ncomms7421.

Domman, D., Collingro, A., Lagkouvardos, I., Gehre, L., Weinmaier, T., Rattei, T., Subtil, A., and Horn, M. (2014). Massive Expansion of Ubiquitination-Related Gene Families within the Chlamydiae. Molecular Biology and Evolution 31, 2890–2904.

Domman, D., Steven, B., and Ward, N.L. (2010) Random transposon mutagenesis of Verrucomicrobium spinosum DSM 4136T. Archives of Microbiology 193, 307-312.

Domman, D., Horn, M. (In review at MBE). Following the footprints of Chlamydial gene regulation.

Domman, D., Weinmaier, T., Turaev, D., Rattei, T. Horn, M. & The Xenacoelomorpha genome consortium. (In prep). Bacterial symbionts of the Xenacoelomorpha.

SELECTED CONFERENCES

Society for Molecular Biology and Evolution General Meeting
(2015). Vienna, Austria. (Poster)
Plastid establishment did not require a chlamydial partner.
Domman, D., Horn, M., Embley, TM., and Williams, TA

Gordon Confrence on Animal-Microbe Symbiosis
(2015). Waterville Valley, NH, USA (Poster)
Comings and Goings: Evolution of gene content in the Chlamydiae.
Domman, D., Collingro, A.,T., Subtil, A., and Horn, M.

Chlamydia Basic Research Society Meeting
(2015). New Orleans, LA, USA. (Talk)
Insights into gene regulatory networks in the Chlamydiae.
Domman, D., and Horn, M.

DARYL DOMMAN

EMAIL ddomman@gmail.com

NATIONALITY USA

PHONE +43 1 4277 76621

Division of Microbial Ecology
University of Vienna
Althanstrasse 14
1090 Vienna
Austria

REFERENCES

Matthias Horn
PhD Advisor
University of Vienna
Division of Microbial Ecology
Vienna, Austria

Phone: +43 1 4277 76608
Email: horn@microbial-ecology.net

Michael Wagner
Department Head
University of Vienna
Division of Microbial Ecology
Vienna, Austria

Phone: +43 1 4277 76600
Email: wagner@microbial-ecology.net

Naomi Ward
Masters Advisor
University of Wyoming
Department of Molecular Biology
Laramie, Wyoming, USA

Phone: +1 307 766 3527
Email: nlward@uwyo.edu

T. Martin Embley
Host for Research Stay
Newcastle University
Institute for Cell and Molecular Biosciences
Newcastle, United Kingdom

Phone: +44 191 222 7702
Email: martin.embley@ncl.ac.uk