universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Ancient gene transfer from Chlamydiae to plants?"

verfasst von / submitted by

Sabine Felkel, BSc

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

Master of Science (MSc)

Wien, 2015 / Vienna 2015

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 833
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Ecology and Ecosystems
Betreut von / Supervisor:	Univ.-Prof. Dr. Matthias Horn

Index of abbreviations

(B)MCMC	(Bayesian) Markov chain Monte Carlo
C | P	Chlamydiae are in a sister relationship to plants
C to P	Donation from Chlamydiae to plants observed
EGT	Endosymbiont gene transfer
GTR	General time reversible
HGT	Horizontal gene transfer
HKY	Hasegawa, Kishino and Yano model
JC	Jukes & Cantor model
LBA	Long branch attraction
LG	Le & Gascuel model
ML	Maximum Likelihood
NR	Tree shows no clear resolution
NR, grouping	Tree shows no clear resolution, Chlamydiae grouping closely to plants
OTU	Operational taxonomic unit
P to C	Donation from plants to Chlamydiae observed
PVC	Planctomycetes, Verrucomicrobia and Chlamydiae
REC	Dayhoff recoding
T3SS	Type III secretion system

Table of contents

1 Abstract

In the course of this thesis, a selection of proteins which are apparently shared between Chlamydiae and plants have been observed to decipher if horizontal gene transfers happened between these two groups of organisms in ancient times. The phylogenetic trees, obtained with the best fitting methods and models, show far less support for the extensive gene transfer from Chlamydiae to plants than assumed in previous studies. Nevertheless, individual gene transfer events in general, including Chlamydiae and plants, as well as organisms from other groups of the tree of life, cannot be excluded. Rather, an enormous number of horizontal gene transfers, reported by several previously published studies, gathered further support by the results obtained in this thesis. These horizontal gene transfer events might have taken place not only between bacteria and eukaryotes, but might also have involved viruses and multiple exchanges for single genes. The additional exposure to a billion years of evolutionary change explains the difficulties in reconstructing the phylogenetic histories for these proteins, which resulted in some trees showing weird topologies, unexpected Bayesian posterior probabilities or partly unresolved branches. Nevertheless, the superiority of complex mixture models, taking into account across-site heterogeneity, over the simple homogeneity-assuming LG model, has been shown in this study. Especially the combination of the CAT + GTR mixture model with the Dayhoff recoding strategy to mitigate the effects of across-branch variation in substitution rates emerged as a very useful approach for the reconstruction of deep and complex phylogenies.

However, the models able to imitate the natural sequence evolution of genes that possibly underwent ancient inter-domain transfer events and the methods needed to address such deep-ranging phylogenetic questions about the evolution of life are still in their infancy.

Keywords: Ancient HGT, Archaeplastida, Bayesian methods, Chlamydiae donating genes, deep phylogeny, EGT, endosymbiont hypothesis, large datasets, LBA, ménage-à-trois, model comparison, PhyloBayes, plant evolution, primary plastid endosymbiosis, systematic errors

2 Introduction

2.1 Facts about symbiosis

Symbiosis is the Greek word for living together. It refers to a close and long-lasting relationship between two or more organisms of sometimes even phylogenetically very separated species (Dimijian, 2000). The term symbiosis includes far more associations with varying degree of intimacy and of different types than the „symbiosis“ used in vernacular language. If a symbiosis benefits all partners it is called mutualism. If there is no harm or benefit for one partner it is commensalism and in the antagonistic case of symbiosis one partner is parasitic to the other, that is the symbiont has a benefit at the cost of the host (Dimijian, 2000). As mentioned above, symbioses can be of varying degrees of intimacy (Horn et al., 2005). A symbiosis can be obligate for one or all partners, meaning that the hosts or rather the symbionts reproduction cycle is dependent on this particular relationship, or it can be facultative. In that case the association is useful for one or all partners but not necessary to be able to reproduce (Koga et al., 2003). There are even more categories one can assign a symbiosis to. Symbionts can have an extra-, epi- or intercellular lifestyle or an intracellular lifestyle like all known Chlamydiae, for instance (Margulis & Fester, 1991).

2.2 The endosymbiont hypothesis

More than a century passed by since Mereschkowksy came up with the idea that plastids derived from engulfed ancient cyanobacteria (Mereschkowsky, 1905). In 1970 the endosymbiont hypothesis experienced a revival through Lynn Margulis (Margulis & Fester, 1991; Margulis, 1970).

The hypothesis assumes that about 1.5 billion years ago (Facchinelli et al., 2013; Reyes-Prieto et al., 2006; Rodríguez-Ezpeleta et al., 2005) an ancient eukaryotic cell engulfed another organism, a bacterium belonging to the cyanobacteria, by phagocytosis and subsequently a symbiosis developed (Reyes-Prieto et al., 2007).

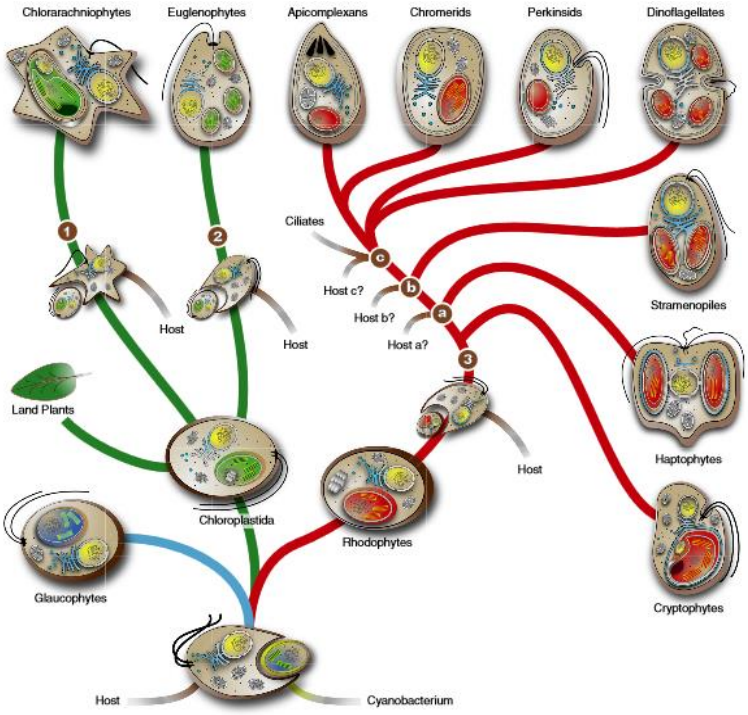

Figure 1: Primary endosymbiosis leading to plastid establishment and the three lineages of Archaeplastida: the Glaucophytes, Chloroplastida and Rhodophytes. After the initial uptake of a cyanobacterium by a heterotrophic host, two individual secondary endosymbiotic events lead to the Chlorarachniophytes (see 1 in figure) and Euglenophytes (see 2 in Figure 1). The radiation of secondary red plastids is not fully resolved, but the initial step (see 3 in Figure 1) was monophyletic too. (Zimorski et al., 2014)

Ancient cyanobacteria were the pioneers who started to enrich the earths athmosphere with oxygen by perfomring photosynthesis about 3.6 billion years ago (Dagan et al., 2013; Gould et al., 2008). Thus the symbiont, capable of photosynthesis (Gould et al., 2008), was able to provide photosynthetic products to the eukaryotic host cell and in return was sheltered (Ball et al., 2011; Facchinelli et al., 2013; Price et al., 2012). After a long period of time the symbiosis became stable so that the bacterium further developed into an organelle of this early eukaryote (Gould et al., 2008). Finally, the chloroplast, and therefore the ancestor of the plastid containing Archaeplastida (Facchinelli et al., 2013), was born (Zimorski et al., 2014). Subsequent in-depth phylogenetic analyses of plastid genomes revealed that they indeed represent a particular group of the cyanobacteria. Even though there is no reason to doubt the origin of the plastid (Keeling & Palmer, 2008), many aspects of this unique event still remain obscure. The plastid did undergo a billion years of evolutionary change slowly drawing a curtain over its origin (Delsuc et al., 2005). Mutations and changes in substitution rates in the course of adaption to its “host” make it very difficult to trace back the true evolutionary history.

2.3 The plant kingdom

The point in time when the symbiosis became stable was the hallmark of the development of Archaeplastida and subsequently the three lineages of the plant kingdom - the Glaucophyta, the Rhodophyta and the Viridiplantae/Chloroplastida (Figure 1) evolved (Deschamps et al., 2008; Facchinelli et al., 2013; Qiu et al., 2013). Due to several subsequent secondary and tertiary endosymbioses (Bhattacharya et al., 2004; Deschamps, 2014; Li et al., 2006; Moustafa et al., 2008; Ohta et al., 2003; Petersen et al., 2006; Reyes-Prieto et al., 2007; Rockwell et al., 2014; Yoon et al., 2004) that obviously took place, the chloroplast spread into different eukaryotes like diatoms, dinoflagellates or euglenids, for instance (Bhattacharya et al., 2004; Petersen et al., 2006; Price et al., 2012; Qiu et al., 2013). As a result of metabolic integration, adaption to the different hosts and their corresponding environments (Ball et al., 2011; Facchinelli et al., 2013; Gross & Bhattacharya, 2009), three different types of plastids developed: the ones containing blue, green or red pigments (Facchinelli et al., 2013; Gould et al., 2008). Although opinions differ, the most prominent and best supported theory is that all of them share the same ancestor, meaning that Archaeplastida are monophyletic (Deschamps et al., 2008; Domman et al., 2015; Marin et al., 2005; Petersen et al., 2006; Price et al., 2012; Rodríguez-Ezpeleta et al., 2005), as shown in Figure 1. There are also studies showing Glaucophytes being basal to all other Archaeplastida, e.g. they still have peptidoglycan from the bacterial endosymbiont in their plastid membrane (Stiller & Hall, 1997; Martin et al., 1998; Reyes-Prieto & Bhattacharya, 2007) and the pigment of *Cyanophora sp.* is the one that most closely resembles its cyanobacterial ancestor (McFadden & van Dooren, 2004). In addition to that, the green algae *Mesostigma* is the closest living relative to unicellular algae (Mcfadden, 2001) from which green land plants descended, leading to the conclusion that the Chloroplastida lineage was not the first that split. Furthermore, given the fact that red plastids have higher similarities to green plastids than the Glaucophytes plastid, there is much evidence for Glaucophytes being the first archaeplastidal lineage that evolved (Stiller & Hall, 1997; Rodríguez-Ezpeleta et al., 2005).

2.4 Uniqueness of the primary endosymbiosis?

Primary plastid endosymbiosis seems to have happened just once in evolutionary history and therefore the question why this is the case comes up frequently (Ball et al., 2013; Becker et al., 2008). However, recently something that can be considered a second primary

endosymbiosis has also been observed in the filose amoeba *Paulinella (*Bodył et al., 2007; Deschamps et al., 2008; Marin et al., 2005). *Paulinella chromatophora* took up strains of *Synechococcus* and *Prochlorococcus*, which belong to the cyanobacteria (Ball et al., 2013; Facchinelli et al., 2013; Qiu et al., 2013). It was shown that the two kidney-shaped endosymbionts cannot be cultivated outside the host (Marin et al., 2005; Keeling, 2004) and divide synchronously with host cell division (Hirt & Horner, 2004). Concluding from this, it simply seems that a lot of complex conditions need to be fulfilled for such a rare event in evolutionary history to occur. What exactly these conditions are still has to be explored.

2.5 Horizontal gene transfer

Chloroplast DNA is circular (Reyes-Prieto et al., 2007) with genes arranged in operons (Mcfadden, 2001). The number of genes varies from 70 genes in non-photosynthetic plastids up to 200 in the plastids of red algae - typically there are about 100 genes in the chloroplast of plants (Mcfadden, 2001). A lot of genes are not necessary in an intracellular lifestyle, leading to the loss of the majority of the original bacterial genome (Suzuki & Miyagishima, 2010) so that only about 200 kb are left from the original 2 – 4 mb genome (Mcfadden, 2001). However, observation of the proteome of the host nucleus revealed that genes of putative cyanobacterial origin are also present (Martin & Herrmann, 1998). This lead to the assumption that for some of the plastidal genes endosymbiont gene transfer (EGT) events (Qiu et al., 2013) probably took place as a corollary of metabolic integration (Facchinelli et al., 2013) and evolution. Furthermore, these proteins often have an amino-terminal extension that functions as plastid-targeted signal, a transit peptide (Li et al., 2006). This transit peptide ensures that the protein goes back to the plastid where it finally operates (Qiu et al., 2013, Mcfadden, 2001). The almost universal presence of this targeting peptide at the N-terminus of the proteins targeting plastid compartments is very useful for the identification of new plastid-destined proteins with bioinformatics (Qiu et al., 2013). The finding that there is something like a core set of horizontally transferred genes still present in the nucleus of all or at least most of Archaeplastida from all three lineages gave further support to their classification as a monophyletic group (Qiu et al., 2013).

2.6 The phylum Chlamydiae

All members of the phylum Chlamydiae are obligate intracellular bacteria covering a broad host range. They are among the most successful pathogens to humans (Haider et al., 2008b; Horn, 2008) and play an important role as symbionts of free-living amoebae. There is evidence that this phylum is tremendously diverse and displays wide distribution in nature. However, cultured and sequenced isolates, especially within environmental chlamydiae, are limited to few organisms. In the 1990s, the first members of chlamydia-like bacteria were identified mainly as symbionts of free-living amoebae. Since then, eight new families were described apart from the *Chlamydiaceae* (Figure 2).

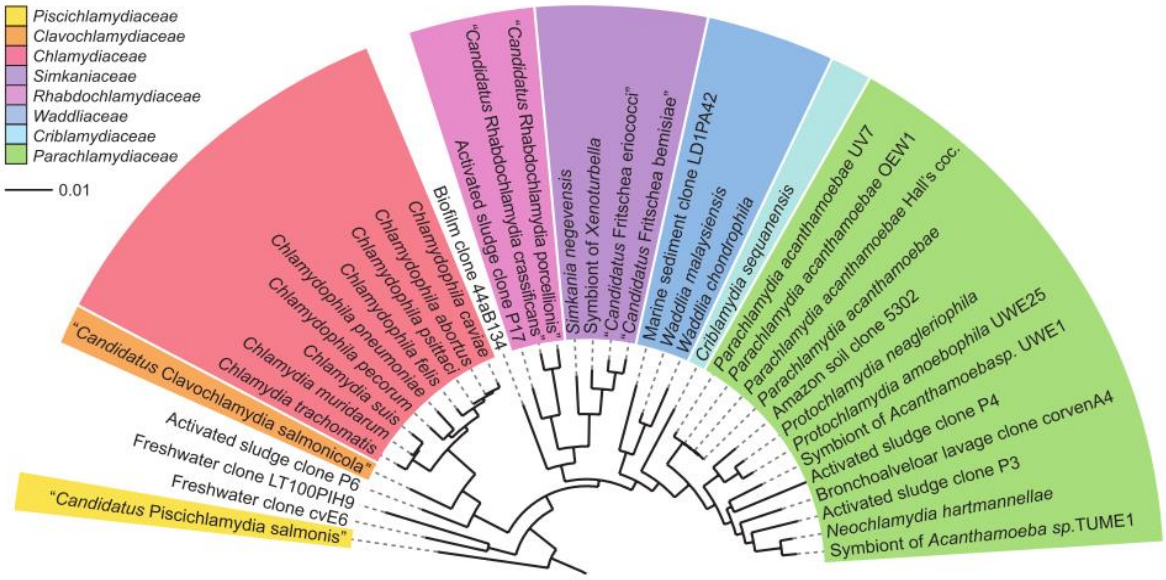

Figure 2: Overview of known and already described families of Chlamydiae, the pathogenic Chlamydiaceae and the chlamydia-like bacteria families. The recently discovered family Parilichlamydiaceae (Stride et al., 2013) is not shown in this figure. (Horn, 2008)

All members share a characteristic bi-phasic developmental cycle with morphologically distinct, extracellular, infectious elementary bodies (EB) and reproductive reticulate bodies (RB) as explained in Figure 3. The chlamydial developmental cycle takes place within a vacuole, referred to as an inclusion, which appears to circumvent the lysosomal function of the host cell (Eissenberg & Wyrick, 1981; Heinzen et al., 1996).

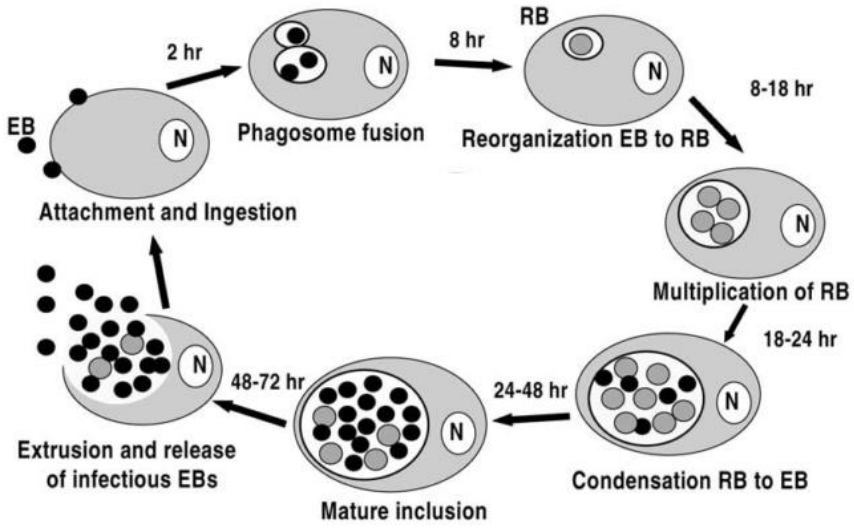

Figure 3: Life cycle of Chlamydiae in epithelial cells. The elementary body (EB) enters the host and transforms into a reticulate body (RB) to replicate inside the inclusion. When enough replication cycles took place, RBs condense again to EBs. After 72 hours the EBs are released and ready for a new infection cycle. (Hammerschlag, 2002)

Current phylogenetic trees demonstrate the wide distribution of the Chlamydiae in nature (Corsaro et al., 2010; Corsaro & Venditti, 2009; Horn, 2008; Schmitz-Esser et al., 2008). Previous studies revealed a huge, hidden family-level diversity of Chlamydiae, present in a variety of environments, particularly in marine habitats (Lagkouvardos et al., 2014). Even when applying the most conservative approach, clustering of chlamydial 16S rRNA gene sequences into operational taxonomic units (OTUs) revealed an unexpectedly high species, genus and family-level diversity within the Chlamydiae, including 181 putative families (Lagkouvardos, Weinmeier et al., 2014). As many other gram-negative bacteria, the Chlamydiae also possess a type III secretion system (T3SS) to colonize and parasitize susceptible hosts through the activity of translocated virulence effectors (Bretz & Hutcheson, 2004; Tseng et al., 2009). The T3SS protein complex crosses the membranes of both the symbiont and the host and allows the bacterium to directly inject effectors or toxins into the host which are necessary to manipulate host cell function (Coburn et al., 2007). Although controversial, some of the chlamydia-like bacteria have been implicated as emerging human pathogens (Corsaro & Greub, 2006; Faires et al., 2009). Previous studies suggest an association of chlamydia-like bacteria with respiratory disease, but more systematic studies are needed to confirm that there indeed is a causal association with disease or if their presence rather reflects the ubiquity of these bacteria in our environment (Collingro et al., 2005; Corsaro et al., 2002; Haider et al., 2008b). It is also known that some of the environmental chlamydiae have genomes twice as large as any of the pathogenic

Chlamydiaceae with few signs of recent lateral gene acquisition (Collingro et al., 2005; Horn et al., 2004). Probably chlamydia-like bacteria have to deal more often with fluctuating environmental conditions (Horn et al., 2004) and therefore might need a bigger assortment of genes which are part of defense and/or survival strategies. Although Chlamydiae are not found in plants, an unexpectedly high number of chlamydial genes are most similar to plant homologs (Huang & Gogarten, 2007) leading to the controversial ménage-à-trois hypothesis that an early representative of the Chlamydiae took part in the ancient events that led to the formation of the plant lineages (Brinkman et al., 2002; Collingro et al., 2011; Everett et al., 1999; Horn et al., 2004; Huang & Gogarten, 2007). The discovery of the so-called signature protein, unique to the PVC superphylum the Chlamydiae are part of (Wagner & Horn, 2006), might be helpful for tests of this hypothesis as well as for the identification of new species. The PVC superphylum, named for its main member phyla Planctomycetes, Verrucomicrobia and Chlamydiae, is a group within the bacteria domain whose members form distinct clades in phylogenetic trees, to some degree indicating common ancestry (Fuerst, 2013). There are even more open questions with respect to evolution within the Chlamydiae, e.g. finding the last common ancestor of chlamydia-like bacteria and their pathogenic counterparts (Horn et al., 2004). Thus, this well-separated phylum deserves attention because new insights into the biology and evolution of these bacteria will accelerate the progress in medicine and lead to insights into general evolutionary topics. Furthermore, this phylum is also of general evolutionary interest due to a proposal that Chlamydiae took part in the ancient events leading to the formation of the plant lineages (Ball et al., 2013; Becker et al., 2008; Facchinelli et al., 2013; Huang & Gogarten, 2007; Moustafa et al., 2008).

2.7 Was there a ménage-à-trois?

In several phylogenetic studies it was shown that other bacteria besides cyanobacteria possibly played a role as gene donors during the primary endosymbiosis event, e.g. Chlamydiae and Proteobacteria (Facchinelli et al., 2013; Qiu et al., 2013). First hints for the theory that Chlamydiae possibly facilitated the establishment of the primary plastid were found in 1998 by Stephens et al., who showed that a lot of *Chlamydia trachomatis* genes have closely related plant genes. Although even more different bacterial phyla were identified that could have been putative gene donors to plants or even facilitators of the primary endosymbiosis, scientists focussed on Chlamydiae. This is due to the fact that this phylum is

the only one whose members are all obligate intracellular bacteria which makes them more suspect (Suzuki & Miyagishima, 2010) in addition to the fact that Chlamydiae have a higher proportion of shared genes with plants compared to other taxa (Brinkman et al., 2002). What further supports the idea of investigating the Chlamydiae in more detail is the fact that this phylum is about 1.6 billion years old (Horn et al., 2004). This fits quite well to the hypothesis that ancient chlamydia facilitated primary endosymbiosis which took place around 1.5 billion years ago (Subtil et al., 2014). The authors of several studies tried to show that Chlamydiae were involved in primary plastid establishment and facilitated the connection between the cyanobacterium and the early eukaryotic host via e.g. the glycogen pathway, which is fully present in the chlamydial genome (Ball et al., 2011). More than 100 proteins seeming to have a chlamydial origin have been observed in previous studies (Ball et al., 2013; Becker et al., 2008; Facchinelli et al., 2013; Huang & Gogarten, 2007; Moustafa et al., 2008).

Although some alternative explanations (Deschamps et al., 2008) are in discussion, the most prominent model is that the chlamydial pathogen primed plastid establishment by secreting effectors via the T3SS into the host cytosol and the cyanobiont rescued the host by providing photosynthetic product to the chlamydial controlled pathway (Ball et al., 2013). An other alternative explanation is that both symbionts were located in a vesicle and first the chlamydia was on top of the ménage-à-trois and controlled the glycogen pathway. Later on, the cyanobacterium escaped from the vesicle and subsequently the chlamydial cell and its inclusion membrane was maintained as long as it provided useful effectors for the establishment of the plastid (Facchinelli et al., 2013). However, a recently published study raised doubts about the origin of chlamydia-like sequences in plant genomes by applying more sophisticated phylogenetics (Domman et al., 2015). Nevertheless, the idea that three partners were involved in the plastid establishment would be a good explanation for the primary endosymbiosis being so unique in evolutionary history (Becker et al., 2008; Facchinelli et al., 2013).

2.8 Introduction to phylogenetics

Phylogenetic methods play an important role in modern biology because they allow for visualization of evolution as descent from common ancestors (Baum & Smith, 2013). Since evolution seems to be treelike (Gogarten et al., 1999), a phylogenetic tree is a good approximation to reality. It is important to know is that branch lengths in a phylogenetic tree

do not directly represent time between species, but rather show distances as the number of substitutions a sequence needs to undergo to get to the state of the other sequence (Baum & Smith, 2013). Replacement matrices are used to compute substitution probabilities along branches, and thus the likelihood of the data (Le & Gascuel, 2008). Evolutionary models are based on a set of simplifying assumptions about the evolutionary process (Williams et al., 2011) and more or less simulate the random process of evolution of sequences using a reasonably small number of parameters, including the topology itself. If the values of these parameters are fixed at one possible set of values they might take the model attaches probabilities to each site of the alignment observed (Goodfellow et al., 2014). These probabilities vary between the different sets of parameters and the set leading to the observation of the highest overall probability is considered the best. However, high probabilities do not necessarily mean that the tree is accurate or that the model fits the data under investigation (Lartillo et al., 2007).

2.8.1 Evolutionary models

Empirical matrices like the general time-reversible model (GTR), used to analyse nucleotide sequences differ in the number of parameters and therefore in complexity. The most complex one, the GTR model, has 3 free parameters referring to different base frequencies in the data plus 6 free parameters representing the different substitution rates (Yang & Rannala, 2012). Examples for special cases of the GTR model, differing in assumed base frequencies and substitution rates and therefore numbers of parameters, are shown in Figure 4. Such standard phylogenetic models assume that the probabilities of the different nucleotide replacements are identical across sequences and thus across branches of a tree (Goodfellow et al., 2014).

Substitution models are also available for the analysis of protein sequences and are conveniently summarised in terms of a 20 x 20 rate-matrix, specifying the rate of substitution between each pair of amino-acids, whereas higher rates of substitution between biochemically similar amino-acids are assumed (Baum & Smith, 2013).

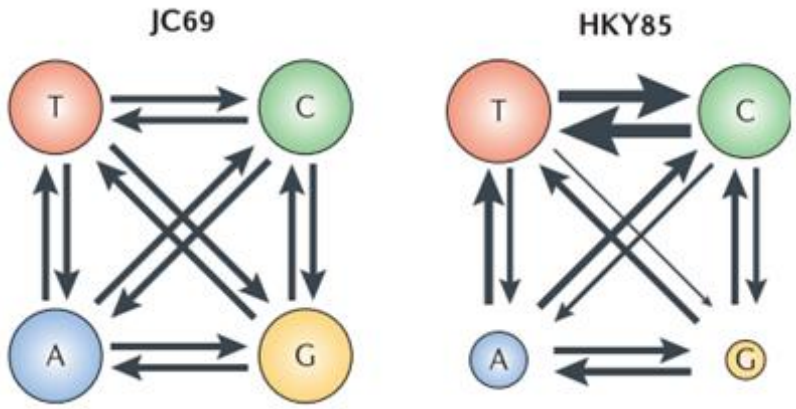

Figure 4: Markov models of nucleotide substitution. The thicker the arrows the higher the probability of substitution. The sizes of the circles around each base represent the nucleotide frequencies when the substitution process in in equilibrium. In contrast to the more complex Hasegawa, Kishino and Yano model (HKY85) or the Felsenstein model (F81, not shown here), the Jukes Cantor model (JC69) predicts equal proportions of the four nucleotides as well as homogeneous transition and transversion rates. Transitions are purine to purine (A and G) or pyrimidine to pyrimidine (C and T) substitutions. Transversions are changes from purines to pyrimidines or rather vice versa. (Yang & Rannala, 2012)

There are two main attitudes to assign rates to the different substitutions. First, in case of the GTR approach all parameters of the matrix are estimated directly from the data along with the other parameters of the model like tree topology or branch lengths (Lartillot et al., 2013). Given the number of parameters entailed by a 4 x 4 nucleotide matrix, this works well only if the dataset is big enough (Quang et al., 2008). Even more information is needed to get reliable estimations for protein analysis, depending on a GTR 20 x 20 amino acid matrix (Lartillot et al., 2013) and more than 200 parameters (Huelsenbeck et al., 2008).
Secondly, for empirical approaches like the WAG (Whelan & Goldman, 2001), JTT (Jones et al., 1992) or LG (Le & Gascuel, 2008) model matrix parameters with realistic properties have been learnt based on a separate database (Paterson & Lima, 2015). Such databases are based on several dozens of hundreds of single-gene alignments (Paterson & Lima, 2015).
Another alternative approach is represented by the empirical profile mixture models, e.g. the CAT model (Lartillot, 2004). Such mixture models also take into account that different sites in an alignment are under distinct evolutionary pressure (Lartillot, 2004). The among-site rate variation is accommodated through the assignment of different biochemical profiles, which are probability vectors over the 20 amino-acids, to each class of site (Lartillot et al., 2007). These classes differ in several features of sequences (Le et al., 2008), like for instance GC-content and different biochemical constraints (hydrophobic, polar, positively charged, etc.), which sites with similar features, and therefore substitution rates, are assigned to (Blanquart

& Lartillot, 2008). Whenever a substitution event occurs, an amino-acid is chosen at random, according to the probabilities defined by the profile (Lartillot et al., 2007). This is called a Poisson or a F81 amino-acid replacement process (Felsenstein, 1981). The likelihood at each site of an alignment is a weighted average over all available Poisson processes defined by the mixture (Quang et al., 2008). Mixture models perform particularly well on saturated data (Smith & Smith, 1996), that is data with reduced sequence divergence rates at single sites resulting from e.g. reverse mutations or homoplasies. These models are more robust to classical systematic errors and phylogenetic artefacts like long branch attraction (LBA) (Brinkmann et al., 2005) and therefore well suited for especially deep phylogenetic reconstruction. LBA (Bollback, 2002) is the grouping of very fast-evolving sequences (Lartillot et al., 2007) even if they are not related at all, just because they have similar high evolutionary rates (and therefore long branches). Furthermore, profile mixture models are available only in a Bayesian framework and not in a Maximum Likelihood (ML) context (Quang et al., 2008). By prior use of a Dirichlet process (Nguyen et al., 2013), the total number of classes and the respective amino-acid profiles, as well as the assignment of the sites to a given class, are free parameters of the model (Lartillot, 2004). Therefore, the model is able to adapt to the complexity present or not present in the data (Lartillot, 2004). The relative rates, also global exchangeabilities, are either taken from existing empirical models (e.g., CAT-JTT, WAG or LG model - classic empirical single-matrix models, i.e. with only one component) or estimated from the data (e.g. CAT-GTR model) (Lartillot et al., 2013). However, even the CAT model assumes that evolution is constant across branches (Goodfellow et al., 2014). Evolutionary rates show high site heterogeneity, depending on biochemical factors like genetic code, solvent exposure, protein folding or function (Le et al., 2008). Depending on these factors some sites are subject to high selective pressure and evolve rapidly, while others are highly conserved (Le et al., 2008).

As nucleotide heterogeneity may affect amino acid composition (Hervé Philippe & Roure, 2011) and therefore the topology of trees calculated by models which are not able to account for this variation, Dayhoff recoding was invented (CAT-GTR-REC model). Dayhoff recoding is a primitive way of mitigating the effects of compositional heterogeneity by binning amino acids which tend to replace each other into 6 groups based on biochemical properties (Hrdy et al., 2004). The substitution rates among amino acids in the same bin are higher than between bins, leading to a reduction of saturation and compositional heterogeneity, nevertheless, at

the cost of information (Hervé Philippe & Roure, 2011). However, in total a substantial improvement in phylogenetic accuracy and resolution is attained often, especially in very heterogeneous datasets, by mitigating the effects of systematic errors and poor model fit (Susko & Roger, 2007).

The complexity of the model of choice should fit to the complexity of the data observed and the question asked. Complex models are required for the investigation of ancient gene transfer events or large datasets that consist of sequences of all domains of life, for instance. A good way is to run chains for each model in parallel (Lartillot et al., 2013) and to compare the results of different models to find the best fitting model.

2.8.2 Methods for tree calculation

The evolutionary models introduced above can be applied in the context of Maximum Likelihood and/or in a Bayesian framework. Unlike alternative methods, e.g. Maximum Parsimony or Distance Methods (Yang & Rannala, 2012), both are characterbased statistical methods used to infer phylogenetic relationships. Maximum Likelihood finds the tree with the highest probability to give the observed data (Arora et al., 2006). A tree and its parameters are set up and a Markovian chain visits every site of the alignment and calculates the probability for each possible state (parameter). This is called the site likelihood. The overall tree likelihood is then the product of the sum of the site likelihoods (Baum & Smith, 2013). It is possible to maximize the likelihood for each tree by changing branch lengths and restarting the run. In the end, the computer searches through the treespace and finds the best tree, the tree with the highest likelihood. Given that in general $4^{(n-2)}$ is the number of histories when n is the number of sequences in a tree (Baum & Smith, 2013), a Maximum Likelihood tree search soon becomes not feasible anymore when increasing the number of sequences in the observed dataset (Yang & Rannala, 2012). Even for simple evolution models, Maximum Likelihood is very demanding computationally and it fully depends on the dataset for the calculation of the parameter values (Baum & Smith, 2013).

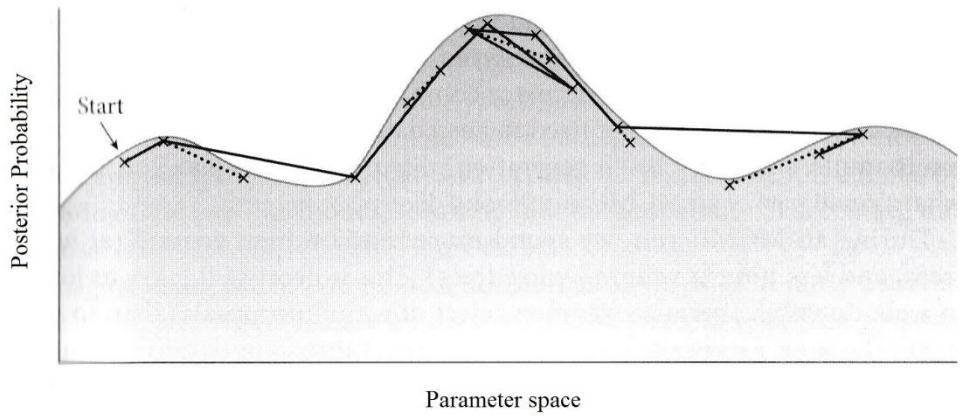

Figure 5: Schematic overview of the MCMC method. Each x represents a value that was proposed during the chain. Either the proposal has been accepted (solid line), or not (dotted line). Contrary to heuristic algorithms, sometimes even downhill proposals are accepted. (Baum & Smith, 2013)

The Bayesian method judges trees on their posterior probability – the best tree is the one that fits best considering data, model and prior beliefs (Arora et al., 2006). A Markov chain Monte Carlo (MCMC, Neal, 1998) analysis is applied. Here, the search wanders through a multidimensional parameter space containing all possible trees and branch lengths and free parameters for the model of evolution. It searches for the maximum posterior probability of a given profile (a topology) and then proposes a new profile and calculates the posterior probability again (Figure 5). Such a proposal and calculation of a new profile is called a MCMC generation (Baum & Smith, 2013). Unlike heuristic algorithms, where values that take the chain downwards are objected, in MCMC sometimes you also pick values that go downwards (Baum & Smith, 2013). In practice, to get more significant results and be able to identify chains that got stuck in a local maximum, one runs more chains in parallel, each starting from a different point, and compares the results (Lartillot et al., 2013). To ensure good mixing, that is that parameter space is well explored when the maximum is reached, multiple chains are run to ensure that the same topology is observed with good support, meaning that the chains converged (Baum & Smith, 2013).

A chain needs to reach stationarity before its values can be considered to come from the posterior distribution, i.e. the chain has to be at a higher point in the landscape (Baum & Smith, 2013). Therefore, the first trees created at the beginning of a MCMC chain, which have low probabilities, are removed (Figure 6). Basically, the cutoff value for this so-called burn-in is about 1/5 of the length of the MCMC chain, whereas length means number of generations (Lartillot et al., 2009).

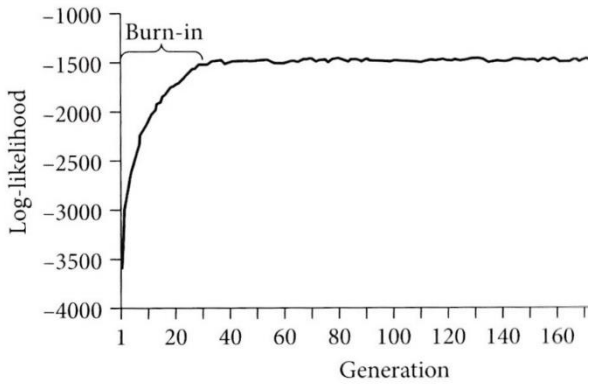

Figure 6: How the likelood changes during a Bayesian MCMC. The samples at the beginning, having bad likelihood values, should be discarded (burn-in) for posterior calculations. (Baum & Smith, 2013)

In the end, there is a consensus tree and the posterior probability of this tree is based on how often the same topology occurs in the post burn-in (Baum & Smith, 2013).

2.9 Project Aims

For this thesis a selection of plant proteins, for which homologues in Chlamydiae and therefore hints for the donation of these genes from Chlamydiae to plants have been found (Ball et al., 2013; Becker et al., 2008; Deschamps, 2014; Facchinelli et al., 2013; Huang & Gogarten, 2007; Moustafa et al., 2008), were reinvestigated. The previous findings support the idea that Chlamydiae donated a repertoire of genes to plants and that they probably even facilitated the establishment of the primary endosymbiosis, therefore leading to the emergence of plants. This study is a reanalysis of these proteins, whereat the sampling has been improved by adding new sequences to the datasets. In addition, modern phylogenetic methods and appropriate substitution models were applied to find out more about evolution, horizontal gene transfer that probably took place from any direction and how HGT makes it difficult and challenging even for sophisticated statistics of nowadays to address questions implying deep phylogeny. We did single-gene analysis of more than 20 proteins shared by plants and Chlamydiae, previously shown to have originated in Chlamydiae, and calculated trees with different phylogenetic models in a Bayesian context. We compared the results in hope to get new insights into model performances and HGT from any direction, the quantity of HGT events and therefore their role in phylogenetic studies, and as means for accelerating the evolutionary process of life.

3 Materials and Methods

3.1 Data sampling and modification

As this thesis is a reanalysis of already published studies, datasets were collected from five selected papers that previously worked on this topic (Ball et al., 2013; Becker et al., 2008; Collingro et al., 2011; Huang & Gogarten, 2007; Moustafa et al., 2008). Due to time and computational resource issues only a subset of 28 protein families was picked for this study. Nevertheless, to cover all papers as good as possible, datasets were chosen that have been investigated in at least three of these five papers.

The gene families were downloaded from the HOGENOM database (Penel et al., 2009).

Given that the number of sequences shows high variety in the original datasets (see column no. seq. in Table 1), an automated masking program was applied to minimize and filter datasets with more than 550 sequences to reduce computational efforts. For this purpose the trim option in T-Coffee 6.18 (Notredame et al., 2000) was used (for details and filter settings see Table S1).

Table 1: Overview and reference details of the 28 protein families. All 28 datasets have been previously observed in at least three of the five selected studies this thesis is based on. The numbers in the last five columns, named after the first authors, show in which papers they have been investigated before. First analysis of the raw HOGENOM files showed that the observed numbers of sequences show high variations. In most cases the chlamydial are in fact basal to plants sequences, indicating donation from Chlamydiae to plants, represented by the blue colored fields in the seventh column. For those five datasets showing no donor relationship, as indicated by the red color, either plants (no. 18, 20 and 28) or other bacterial sequences (no. 14 and 15) are missing in the original files. Therefore, no specific relationship prior to analysis can be supposed in these cases.

no.	gene name	protein name	gi:	ref	HOGENOM	C to P	no. seq.	Moustafa	Ball	Huang	Collingro	Becker
1	ntt_3	ATP/ADP translocase	46445874	YP_007239.1	HOG000238123		166	1		1	1	1
2	mdh	malate dehydrogenase	46447406	YP_008771.1	HOG000220953		469	1	1	1	1	1
3	ispE	probable isopentenyl monophosphate kinase	46447223	YP_008588.1	HOG000019600		697	1	1	1	1	1
4	ispD	2-C-Methyl-d-erythritol 4-phosphate cytidylyltransferase	46445961	YP_007326.1	HOG000218563		587	1	1	1	1	1
5	fabF	3-Oxoacyl-(acyl-carrier-protein) synthase	298537950	YP_008237.1	HOG000060166		1660	1	1	1	1	1
6	ispG	4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase	76788770	YP_007739.1	HOG000261018		99	1	1	1	1	1
7	gutQ	FOG: CBS domain	46401057	YP_008781.1	HOG000264729		881	1	1	1	1	1
8	ygo4	Phosphate Permease	15618590	NP_224876.1	HOG000231892		744	1	1	1	1	1
9	gpmA	phosphoglyceromutase	46445795	YP_007160.1	HOG000221682		1292	1	1	1	1	1
10	pc1106	probable isoamylase	46446740	YP_008105.1	HOG000239197		781	1	1	1	1	1
11	tyrS	Tyrosyl-tRNA synthetase	76788775	YP_008168.1	HOG000242790		918	1	1	1	1	1
12	CAB867	Cation transport ATPase	62185469	YP_220254.1	HOG000250399		1366	1	1	1	1	
13	kdsB	cytidylyltransferase	89898215	YP_515325.1	HOG000007602		753	1		1	1	1
14	plsB	glycerol-3-phosphate acyltransferase	76789550	YP_001654267	HOG000034930		27	1	1	1		1
15	pc0324	hypothetical protein pc0324	46445958	YP_007323.1	HOG000084053		14	1	1		1	1
16	rlmH	hypothetical protein pc1708	46447342	YP_008707.1	HOG000218433		584	1	1	1	1	
17	dapL	L,L-diaminopimelate aminotransferase	46446319	YP_007684.1	HOG000223061		162	1		1	1	1
18	pepF	Oligoendopeptidase F	46400453	YP_008177.1	HOG000059490		346	1	1	1		1
19	pnp	polynucleotide phosphorylase/polyadenylase	46446277	YP_007642.1	HOG000218326		893	1		1	1	1
20	pc0378	Predicted sulfur transferase	46399653	YP_007377.1	HOG000034811		194	1	1		1	1
21	rsmH	probable S-adenosyl-methyltransferase	46445945	YP_007310.1	HOG000049778		884	1	1			1
22	trpD	anthranilate phosphoribosyltransferase	89898246	YP_515356.1	HOG000230451		1142	1	1		1	
23	yfhC	cytosine/adenosine deaminases	89898061	YP_515171.1	HOG000085050		1395	1	1		1	
24	nhaD	Na+/H+ antiporter NhaD and related arsenite permeases	297620457	YP_003708594.1	HOG000251774		74	1		1	1	
25	tgt	queuine tRNA-ribosyltransferase	46446428	YP_007793.1	HOG000223473		1315	1			1	1
26	pc0141	rRNA methylases	46445775	YP_007140.1	HOG000218799		797	1	1		1	
27	miaA	tRNA delta(2)-isopentenylpyrophosphate transferase	46446877	YP_008242.1	HOG000039995		795	1	1			1
28	yohl	tRNA-dihydrouridine synthase	46400854	YP_008578.1	HOG000217854		249	1	1		1	

3.2 Alignment and modification

The datasets were augmented with new chlamydial as well as plant sequences found via BLAST (Altschup et al., 1990) against the NCBI database (Wheeler et al., 2004). Subsequently, alignments were created with Clustal Omega 1.2.1 (Sievers et al., 2011). To keep reproducibility, manual editing of the alignments was avoided. However, to detect and remove poorly aligned positions and divergent regions, alignments were modified with Gblocks 0.91b (Castresana, 2000; Talavera & Castresana, 2007) to become more suitable for subsequent phylogenetic analyses. Gblocks parameters were set so that the minimum length of a block was three and all gap positions were allowed. In case of the datasets which were edited with T-Coffee, relaxed Gblocks was applied: parameters for the minimum number of sequences for a conserved position or rather a flank position, were set to 50% of the number of sequences plus 1 or rather 85% of the number of sequences of the alignment, as recommended by the manual. All alignments used for tree and subtree calculations can be found in the Supplementary information 8.1.2.

3.3 Subtrees

Whenever the resulting trees of a dataset did not converge or were just weakly supported, but still looked promising in some way, subtrees were created in hope to either find higher support for any conclusion or not. Therefore, either real subsamples were selected with Figtree 1.4.2 (tree.bio.ed.ac.uk/software/figtree/) or Archaeopteryx (Han et al., 2009) and more sequences found via BLAST against the NCBI database were added, or, if the dataset was small enough, BLAST results were added directly to the original datasets (final number of sequences shown in Table S1). The datasets were supplemented mainly with plants, chlamydial, but also other sequences from the PVC superphylum (Gupta et al., 2012; Lagkouvardos et al., 2014; Wagner & Horn, 2006), in hope to stabilize the position of Chlamydiae in the topologies of the trees. As for the raw datasets, alignments were created using Clustal Omega and subsequently an entropy-based trimming with default options performed with BMGE 1.1 (Criscuolo & Gribaldo, 2010).

3.4 Analysis prior to phylogenetic analaysis

In the original studies the authors mainly worked with Maximum Likelihood methods (Yang & Rannala, 2012). As a corollary of this, simple single-matrix substitution models were applied which do not adequately account for heterogeneity between sites within a sequence, not to mention heterogeneity across sequences. It is necessary to mitigate the effects of heterogeneity and to apply appropriate substitution models when working with big datasets consisting of sequences from different domains of life and trying to address questions like if there were ancient horizontal gene transfer events.

Nevertheless, ProtTest 3.4 (Abascal et al., 2005; Guindon et al., 2010) was used to evaluate which single-matrix model is the best fitting simple model for each dataset, which was then tested against the performance of more complex substitution models.

Furthermore, results of previous studies indicate that tree calculations may be biased by compositional heterogeneity of sequences. Given that sequences with similar GC content, or amino acid usage for instance (Behura & Severson, 2013; Inagaki & Roger, 2006; Jørgensen et al., 2007; Suzuki, 2003) are more likely to be grouped together, no matter if they are closely related or not, wrong relationships may be predicted. Processes like convergent evolution at single sites along two lineages can lead to homoplasies and genetic saturation (Liu et al., 2014), meaning that over time the appearance of the sequence divergence rate is reduced which makes it harder to resolve historical relationships. Beforehand, to get an idea of a probable impact of the amino acid usages of the sequences on the resulting topologies of the trees, quick correspondence analyses were done with Jalview 2.7 (Waterhouse et al., 2009).

3.5 Phylogenetic methods and models

Tree calculations were perfomed in a Bayesian context (Huelsenbeck et al., n.d.). For each dataset the same five different models, varying in complexity, were used to create phylogenetic trees. The resulting trees were compared to find out which model is superior to the others or simply how they performed in general. The following models were applied, in ascending order with respect to complexity: the best fitting simple model identified by ProtTest for each dataset was always the LG model, more complex models used were CAT, CAT60, CAT + GTR and the most complex was the CAT + GTR + Dayhoff model. In case of the subtrees only the most important LG, CAT + GTR and CAT + GTR + Dayhoff models were

tested. Calculations were performed with the phylogenetic programs PhyloBayes 3.3 (Lartillot et al., 2009) and PhyloBayes-MPI 1.5a (Lartillot et al., 2013).

For each analysis two chains were run in parallel and the bpcomp and tracecomp programmes were used to assess convergence. Every tree stopped after 10.000 cycles if the maximum discrepancies in bipartition frequencies (bpcomp) still were 1, meaning they would never converge according to the manual. Otherwise, trees were judged as converged when the bpcomp and summary statistics (tracecomp) between the two chains had all dropped below 0.3, and the effective sample size of each parameter was at least 100, or rather 0.1 and 50 in case of the subtrees, as recommended in the PhyloBayes manual. For all trees that did not converge and had maximum discrepancies in bipartition frequencies between 0.31 and 0.99, consensus trees for each individual chain were calculated. The topologies of the two resulting trees were tested for similarities. Whenever topologies beared resemblance to each other subtrees were prepared to gain further support for putative conclusions. Furthermore, five independent MCMC chains with NH-PhyloBayes 0.2.1 (Blanquart & Lartillot, 2006) using the site-heterogeneous CAT model and the BP option (Blanquart & Lartillot, 2008), which allows for changes of model settings at breakpoints along the tree (empirical optimisation), were run for the ADP/ATPase-translocase. The chains were run for at least 2.000 cycles and, disregarding convergence, the resulting topologies tested for similarity.

3.6 Posterior predictive tests

In a final step, posterior predictive simulations were performed to evaluate model fit. The programmes ppred (PhyloBayes 3.3) and readpb_mpi (PhyloBayes-MPI 1.5a) were used to perform tests of across-site (site-specific biochemical diversity) and across-branch (compositional homogeneity) heterogeneity for each tree calculated. A model was judged to have failed a particular test if the test statistic calculated on the real data fell outside the central 95% of the simulated distribution.

4 Results

In all previous studies that served as data origin for this study simple substitution models were applied with Maximum Likelihood methods to address the very complex question of an ancient relationship between Chlamydiae and plants. Here, after improving and augmenting the original datasets and applying more sophisticated models using Bayesian interference instead of Maximum Likelihood, a reanalysis was done to find evidence for single ancient horizontal gene transfer events. Nevertheless, a representative of simple models was also applied for reasons of comparability. The goal was to show that appliance of a complex substitution model is necessary and superior to single-matrix models when trying to address questions implying deep phylogeny and ancient gene transfers. For this purpose, Prottest was used to identify the best fitting simple model for each dataset. The results showed that the LG model is the best fitting simple model for all datasets used in this study. Therefore, the LG model was the model of choice to represent single-matrix models.

4.1 Overview

A total number of 174 trees and subtrees were created using five different substitution models for each dataset. All trees and subtrees can be found in the Supplementary information 8.2.1. Although the biggest datasets were reduced by filtering steps, the computing time for some trees was still too tremendous to get highly supported and/or converged trees. Nevertheless, they were included in the analysis part albeit they were treated as less trustworthy.

4.1.1 Summary of topologies observed

Each tree was rooted within clearly separated bacterial outgroup sequences to be able to investigate a probable connection of the topologies of chlamydial and plant sequences. The focus was on assigning each tree to one of three main topologies of interest: gene donation from Chlamydiae to plants, vice versa or a sister topology (Figure 7). Trees that showed no clear resolution, but the tendency of plants and chlamydial sequences grouping together in some way, were assigned to a fourth category. Given the proportion of trees that did not converge, there is of course another category which cases were no clear resolution at all could be observed are assigned to.

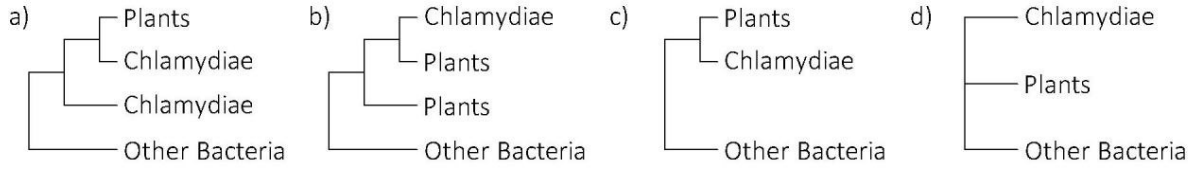

Figure 7: Schematic overview of the four main topologies observed in the data. Donation from Chlamydiae to plants is shown in a) and the opposite case is shown in b). A sister topology is demonstrated in c) and in d) no resolution could be observed.

Interestingly, the case of Chlamydiae and plant sequences being sister to each other has been the most observed throughout all models (Table 2). The sum of the number of trees showing any donation is actually lower than the number of sister topologies, for each model in particular as well as in total. However, Table 2 also shows that gene donation from Chlamydiae to plants has been predicted just as often in case of the representative for simple models, consistently with what has been shown in previous studies. The most complex model on the other hand delivers the fewest cases supporting the idea of Chlamydiae as gene donor. In fact, comparatively many CAT + GTR + REC trees showed the opposite case, a donation from plants to Chlamydiae. Nevertheless, another interesting point shown in Table 2 is that many of the trees with no clear resolution are grouping chlamydial and plant sequences at least closely to each other.

Table 2: Overview of the results of all 174 trees and subtrees. Each tree, analyzed here by the respective model applied, has been assigned to one of the five topology categories. C | P indicates sister topology, C to P and P to C stand for donation of the gene from C to P and vice versa, NR, grouping indicates no resolution but Chlamydiae and plants are grouping closely and NR means no resolution at all. For further details see Table S2.

	C | P	C to P	P to C	NR, grouping	NR
LG	15	12	5	9	1
CAT	11	3	3	6	1
CAT60	11	6	1	6	0
CAT + GTR	17	8	4	11	2
CAT + GTR + REC	17	3	7	10	5
	71	**32**	**20**	**42**	**9**

4.1.2 Topologies and model comparison of converged trees

After first insights into the overall results of this study, it is time to have a closer look at the more reliable and significant results. The pattern that emerged when analyzing all trees calculated more or less stays the same when looking at the predicted topologies of only the 71 trees that converged (Table 3). However, at the expense of observed donations from

plants to Chlamydiae, the proportion of cases that show gene donation from Chlamydiae to plants seems rather high.

Table 3: Overview of the results of all converged trees and subtrees. Each tree, analyzed here by the respective model applied, has been assigned to one of the five topology categories. C | P indicates sister topology, C to P and P to C stand for donation of the gene from C to P and vice versa, NR, grouping indicates no resolution but Chlamydiae and plants are grouping closely and NR means no resolution at all. For further details see Table S3.

	C | P	C to P	P to C	NR, grouping	NR
LG	6	10	3	4	0
CAT	4	1	0	1	0
CAT60	1	0	0	1	0
CAT + GTR	8	5	2	5	1
CAT + GTR + REC	6	3	3	4	3
	25	**20**	**7**	**15**	**4**

4.1.3 Having a closer look: Taking into account Bayesian posterior probabilities

It is very important to also consider Bayesian posterior probabilities when it comes to analyzing the significance and reliability of predicted branches in phylogenetic trees. The prediction of branches was considered as stable and highly reproducible when the Bayesian posterior probabilities of the node was at least 0.8. As a consequence of this selection process a lot of trees were filtered out and just 40 remained for analysis. Despite the small number of trees left, the previously observed pattern could be found again, that is sister topologies have been observed twice as often as donation from C to P or vice versa (Table 4).

Table 4: Overview of the results of all converged trees and subtrees with Bayesian posterior probabilities equal or higher than 0.8 for the nodes leading to conclusions about the relationship between plants and Chlamydiae. Each tree, analyzed here by the respective model applied, has been assigned to one of the five topology categories. C | P indicates sister topology, C to P and P to C stand for donation of the gene from C to P and vice versa, NR, grouping indicates no resolution but Chlamydiae and plants are grouping closely. Results showing no resolution at all are not shown here. For further details see Table S4.

	C | P	C to P	P to C	NR, grouping
LG	4	3	2	3
CAT	3	1	0	1
CAT60	1	0	0	1
CAT + GTR	4	2	2	4
CAT + GTR + REC	2	1	2	4
	14	**7**	**6**	**13**

The proportion of topologies monitored stayed the same when analyzing the 103 leftover trees that almost converged or had maximum discrepancies in bipartition frequencies (bpcomp) of still 1, even after 10.000 cycles (Table S6). Furthermore, the proportion of

observed sister topologies increased with higher maximum discrepancies in bipartition frequencies. However, after consideration of Bayesian posterior probabilities and filtering, far less trees predicting a sister topology of plants and Chlamydiae were discarded (Table S8). Although many trees did not perfectly converge and/or possess highly supported nodes, all trees were used to create Table 2, Table 3 and Table 4 since the same trend has been observed throughout all data produced.

4.2 Selection of trees with special topologies

The result of one tree that looked peculiar in some way was the CAT + GTR tree of the Cytidyltransferase family (no. 13, kdsB). In this case, the organisms of interest are predicted to be sister to each other with a predicted eukaryotic origin and almost sufficient Bayesian posterior probability support (Figure 8). Some sequences belonging to the Oomycetes are arranged basal to the clade containing plants and Chlamydiae. Furthermore, a similar result has been observed for the CAT + GTR subtree of the Probable Isoamylase family (no. 10, pc1106), where *Trichoplax adhaerens*, the only extant representative of the phylum Placozoa, is grouping closely with the chlamydial and plant sequences present in the dataset.

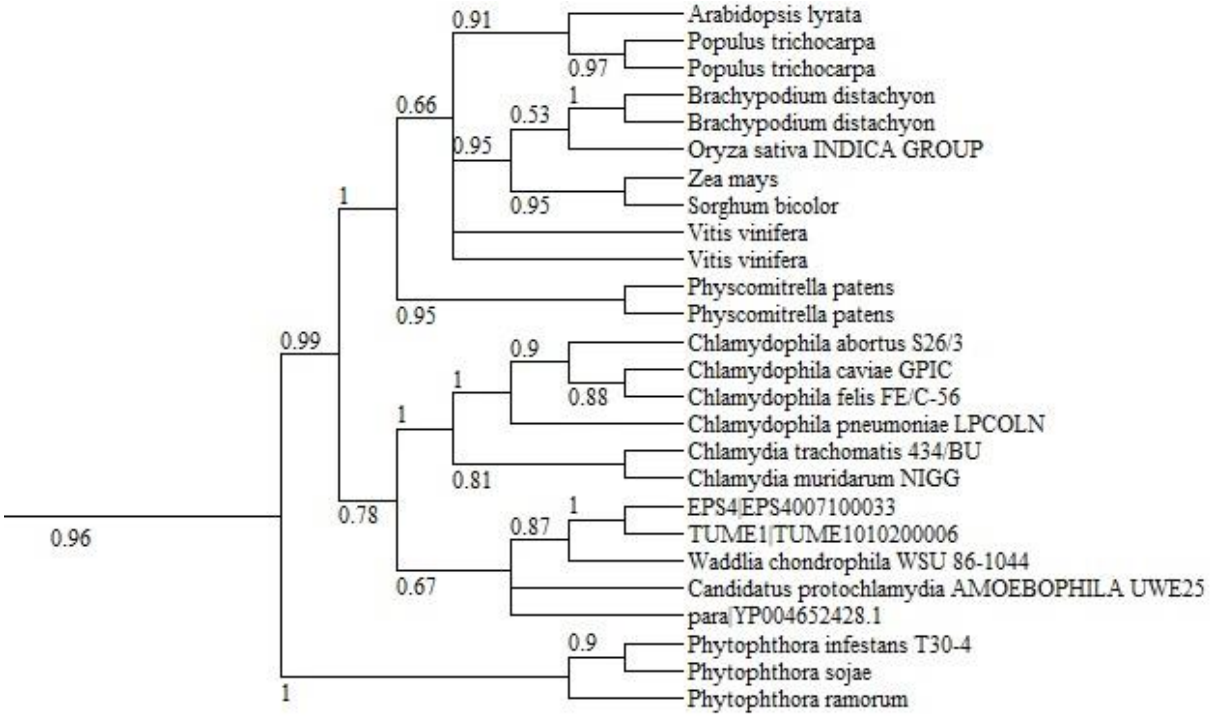

Figure 8: A section of the cladogram of the CAT + GTR tree of the Cytidyltransferase family (no. 13, kdsB). Plants and Chlamydiae are actually sister to each other, whereas some Oomycetes are arranged basal to the clade containing the sequences of interest. The tree is rooted within bacterial sequences, which are not shown here. Full tree can be found in the Supplementary information 8.2.1.

An interesting feature is also shown by the LG subtree for the Phosphate permease protein family (no. 8, ygo4). Here, the actually monophyletic plants are divided up into two clades (Figure 9). In one clade the plants are grouping with bacterial sequences, showing the Chlamydiae as weakly supported sister phylum to plants. Similar results have been also observed in case of the CAT + GTR + REC trees of the protein families Cation transport ATPase (no. 12, CAB867) and Queuine tRNA-ribosyltransferase (no. 25, tgt), which can be found in the Supplementary information 8.2.1. Additionally, a viral sequence from the genus

Coccolithovirus, infecting the calcareous alga *Emiliania huxleyi*, is grouping within the second clade primarly consisting of sequences from lower plants.

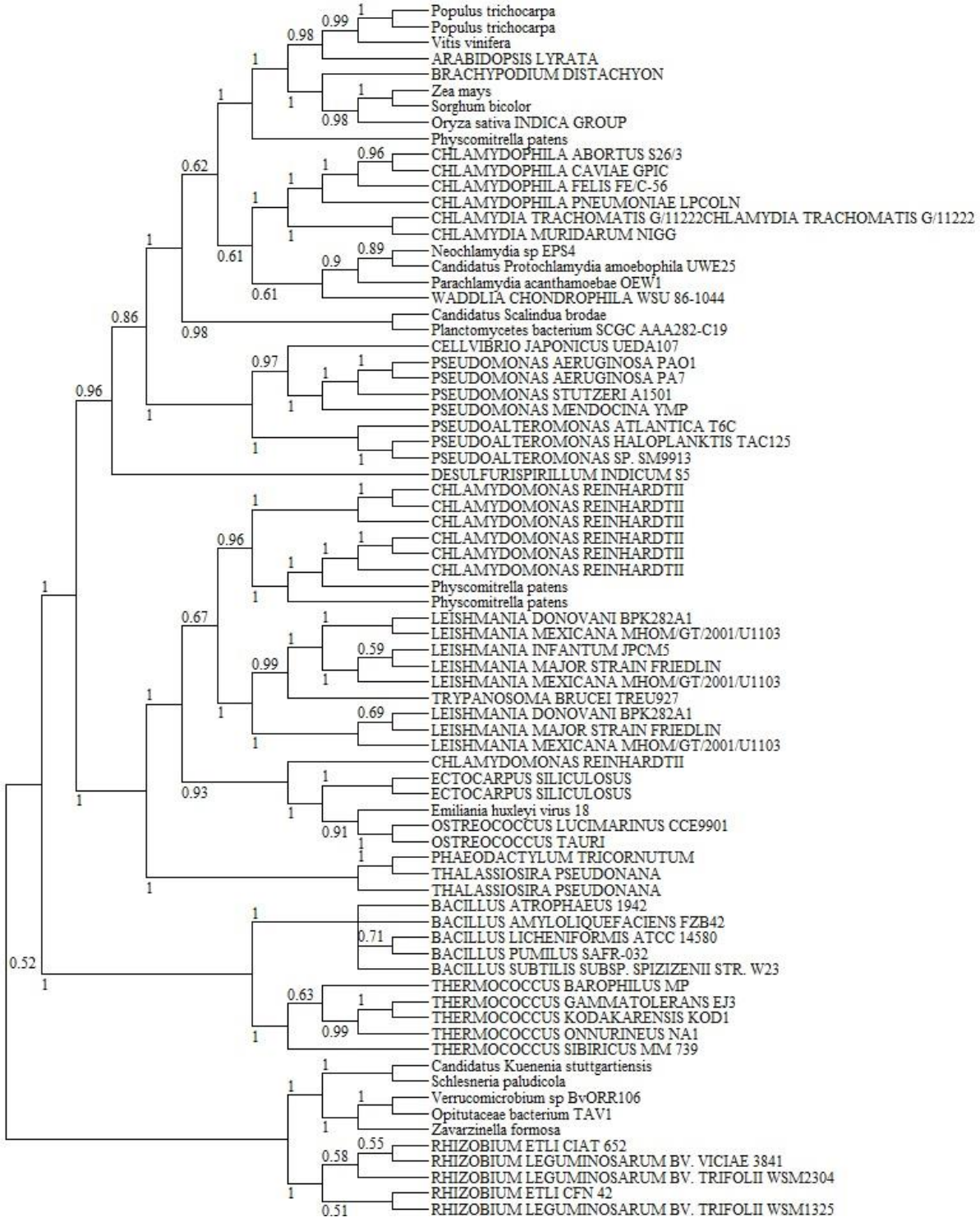

Figure 9: Cladogram of the LG subtree for protein family no. 8 (ygo4), Phosphate permease. A sister topology of plants and Chlamydiae is shown, weakly supported by Bayesian posterior probabilities. Plant sequences divide up into two clades. One clade is grouping with bacterial sequences and in the second clade a viral sequence is grouping among the plants. The tree is rooted within bacterial sequences.

Another interesting tree is the CAT + GTR subtree of the Polynucleotide phosphorylase protein family (no. 19, pnp). Here, at first sight the phylum Chlamydiae and plants seem to be sister to each other, however, a group consisting of three sequences from lower plants is arranged basal to the chlamydial sequences (Figure 10) giving rise to the assumption that gene donation occured from plants to Chlamydiae. The peculiarity of especially *Cyanidioschyzon merolae* grouping with chlamydial sequences could also be observed in the case of the original trees as well as subtrees of dataset no. 18 (pepF), the Oligopeptidase F family, which can be found in the Supplementary information 8.2.1.

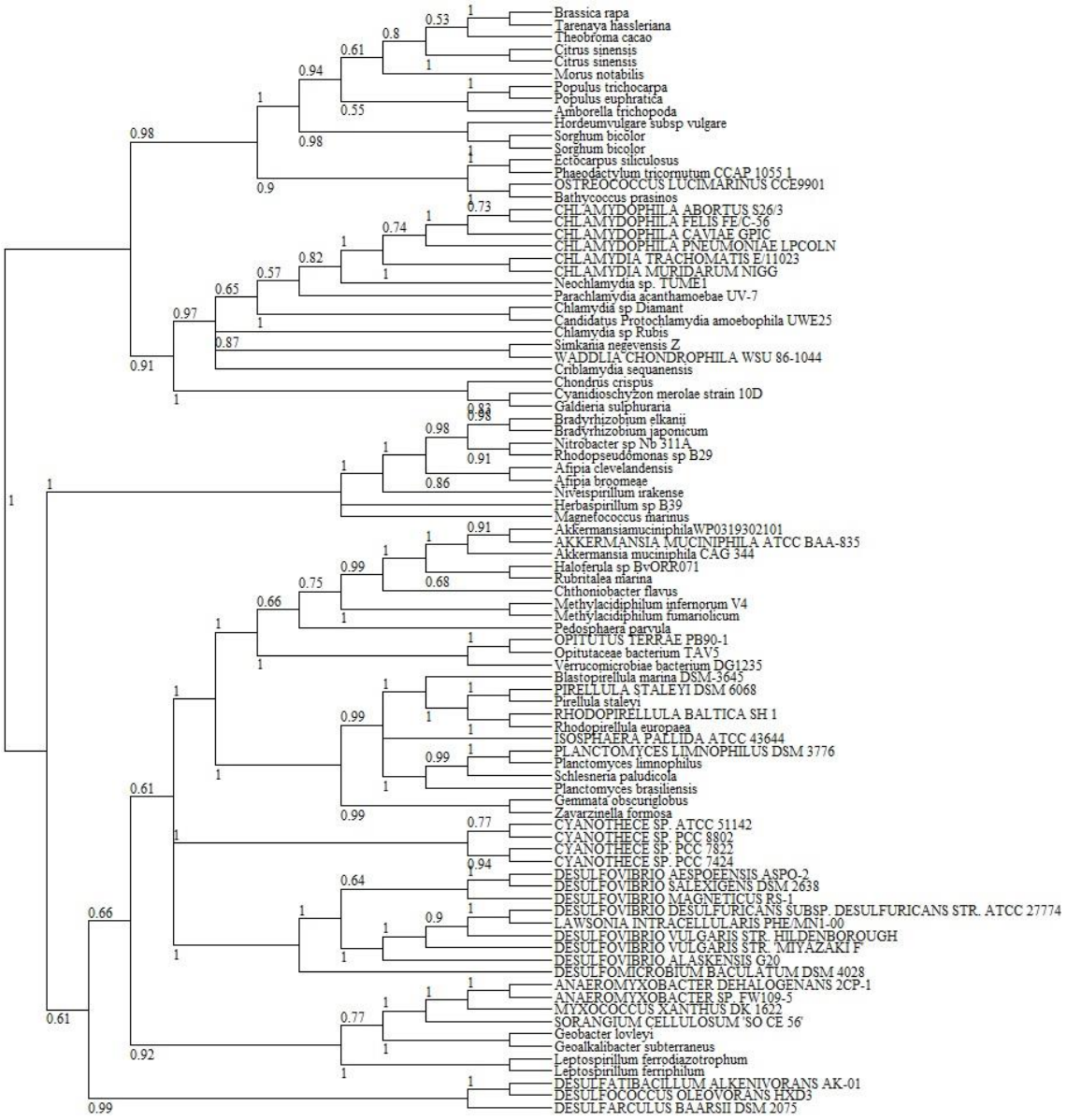

Figure 10: Cladogram of the CAT + GTR subtree for protein family no. 19 (pnp), Polynucleotide phosphorylase. Plants and Chlamydiae are almost completely and significantly separated. Three lower plant sequences are grouping basal to the Chlamydiae. The tree is rooted within bacterial sequences.

4.3 Model comparison: Complex vs. Simple

Apart from reinvestigating ancient gene transfer events, especially from Chlamydiae to plants, this study had a second major aim, namely to address the more systematic question of superiority of one model over another. For this purpose, the topologies of the calculated trees under the different models applied were directly compared to each other, for each dataset. In Table 5 a summary of the comparison of the three models considered most important is shown. It is clearly obvious that the majority of cases where conformity has been observed are related to Chlamydiae and plants being sister to each other. The simple model LG more often predicts any gene donation event than the more complex models. In total, a sister topology has been calculated more frequently under the CAT + GTR than the CAT + GTR + REC model.

Table 5: Direct comparison of the number of observed topologies for the most important models: the representative for single-matix models, LG, the most complex model not taking into account across-branch heterogeneity, CAT + GTR, and the CAT + GTR + REC model as a primitive method to model heterogeneity across branches. Fields indicating consistent observed topologies for the respective pair of models under investigation are colored in green. Results showing no clear resolution are not shown here. See Table S9 for details.

	LG C | P	LG C to P	LG P to C	GTR C | P	GTR C to P	GTR P to C
REC C | P	7	8	1	9	3	0
REC C to P	1	1	0	0	5	0
REC P to C	3	2	2	4	1	2
GTR C | P	12	3	2			
GTR C to P	2	7	0			
GTR P to C	0	1	1			

4.4 Comparing trees calculated on original data with subtrees

Whenever the resulting topology of a tree calculated with the original (but improved) dataset was not supported very well with high Bayesian posterior probabilities or if it did not even converge, but the result still looked promising in some way (e.g. almost significant convergence or Bayesian posterior probability values and/or close grouping of plants and Chlamydiae) and/or PCA of amino acid usage reinforced this assumption, subtrees were prepared. In total, a number of subtrees were created for 18 of the 25 protein families observed in this study, whereas only the three most important models were tested: LG, CAT + GTR and CAT + GTR + REC. To evaluate if a subtree either confirmed or refused a topology proposed by the original tree, i.e. either increased or decreased the significance of the first result, direct model comparisons of the original trees and subtrees were made. The most

overlaps were detected when a model predicted Chlamydiae and plants being sister to each other (Table 6). Donation from plants to Chlamydiae has been observed just twice in both, the original tree and the subtree. There are less cases where a subtree refused the sister topology proposed by the appropriate original tree, than vice versa.

Table 6: Direct model comparison of all original trees and subtrees, for each dataset subtrees were calculated for. Green fields indicate an overlap between the topology predicted by a particular model, observed in the original tree as well as in the subtree. See Table S10 for details.

		---------- Subtrees ----------		
		C | P	C to P	P to C
- Originals -	C | P	13	3	3
	C to P	4	5	2
	P to C	1	0	2

4.5 Analyzing each protein family separately

After evaluation of the overall data produced and the development of a general idea of the results of this study, it was also of interest to analyze the topologies observed for each individual protein family. By comparison of the topologies calculated through the use of the different models it is possible to say whether the case of ancient gene donation from Chlamydiae to plants can be excluded or not, partly depending on an altogether accordance of the predicted topologies.

Table 7: Summary of the predicted topologies for original trees and subtrees and the three main models of interest. Red color indicates that it was not possible to infer a relationship between Chlamyidae and plants prior to analysis, because either plants (no. 18, 20 and 28) or other bacterial sequences (no. 14 and 15) were missing in the raw data. The blue color represents cases where donation from Chlamydiae to plants has been observed and the opposite case is indicated by grey fields. Observation of a sister topology of plants and Chlamydiae is shown in green. Trees and subtrees that did not result in a clear resolution but the tendency of chlamydial and plant sequences grouping together are shown in yellow. Converged trees are marked with a black star (*) and cases with additionally good support of the nodes of interest are denoted with red stars (*). Empty fields represent trees and subtrees that either showed no resolution or were not produced at all (no. 5, 7, 14 and 15). For details see Table S11, Table S12 and Table S13.

			raw data	originals			subtrees		
no.	chlamydiae homolog	gene name	C to P	LG	CAT + GTR	CAT + GTR +REC	LG	CAT + GTR	CAT + GTR + REC
1	ATP/ADP translocase	ntt_3		*	*	*	*	*	*
2	malate dehydrogenase	mdh					*	*	
3	probable isopentenyl monophosphate kinase	ispE					*	*	*
4	2-C-Methyl-d-erythritol 4-phosphate cytidylyltransferase	ispD					*	*	*
5	3-Oxoacyl-(acyl-carrier-protein) synthase	fabF							
6	4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase	ispG		*	*	*	*	*	*
7	FOG: CBS domain	gutQ							
8	Phosphate Permease	ygo4					*	*	*
9	phosphoglyceromutase	gpmA							
10	probable isoamylase	pc1106					*		
11	Tyrosyl-tRNA synthetase	tyrS					*	*	*
12	Cation transport ATPase	CAB867					*	*	*
13	cytidylyltransferase	kdsB							
14	glycerol-3-phosphate acyltransferase	plsB							
15	hypothetical protein pc0324	pc0324							
16	hypothetical protein pc1708	rlmH					*	*	*
17	L,L-diaminopimelate aminotransferase	dapL		*	*	*		*	*
18	Oligoendopeptidase F	pepF					*	*	*
19	polynucleotide phosphorylase/polyadenylase	pnp					*	*	*
20	Predicted sulfur transferase	pc0378		*	*	*	*		
21	probable S-adenosyl-methyltransferase	rsmH					*	*	*
22	anthranilate phosphoribosyltransferase	trpD							
23	cytosine/adenosine deaminases	yfhC							
24	Na+/H+ antiporter NhaD and related arsenite permeases	nhaD		*	*	*		*	
25	queuine tRNA-ribosyltransferase	tgt							
26	rRNA methylases	pc0141							
27	tRNA delta(2)-isopentenylpyrophosphate transferase	miaA					*	*	*
28	tRNA-dihydrouridine synthase	yohl		*			*	*	*

As shown in Table 7, the gene donations from Chlamydiae to plants shown in previous studies are not strongly supported for almost all samples. Only in case of the L,L-diaminopimelate aminotransferase (no. 17), results from original trees as well as subtrees do all agree on an ancient horizontal gene transfer event. On the other hand, there is also only a single protein family, the Polynucleotide phosphorylase (no. 19), for which it was possible to refute the theory of gene donation with original trees as well as subtrees. However, for the Cytidyltransferase (no. 13) dataset, for which no subtrees were prepared, also all models overlapped in their predictions of Chlamydiae and plants being sister to each other. Additionally, in case of the datasets no. 12 and no. 24 almost all models agreed on a sister topology. Interestingly, even strong support for a gene donation from the eukaryotic plants to Chlamydiae was found for the 2-C-Methyl-d-erythritol 4-phosphate cytidyltransferase protein family (no. 4). Another interesting aspect is the fact that even the most simple model, the LG model, predicted a sister topology more often than a donation from Chlamydiae to plants.

4.6 PCA: Role of amino acid usage and heterogeneity

The attempt to get an idea of a probable impact of the amino acid usages of sequences on the topologies of the calculated phylogenetic trees via a quick principal component analysis (PCA analysis, Ringner, 2008) with Jalview failed. The PCA blots showed no specific patterns with respect to the arrangement of Chlamydiae and plants relative to other sequenes. In addition, no positive correlations of the amino acid usages of the single sequences of the alignments used for tree calculations and the topologies observed could be detected (data not shown).

4.7 Posterior predictive tests

Data replicates from a subset of the created data are produced with BMCMC methods and simulated from the posterior predictive distribution for posterior predictive tests (Gelman et al., 2004). Posterior predictive distributions represent the probability that the data is observed with respect to the model used and its prior assumptions (Baum & Smith, 2013). The comparison of the probability values calculated on the basis of observed and simulated data is giving a clue about how good a model fits the data (Gelman et al., 1996). In Bayesian statistics for phylogenetics posterior predictive tests are very important to evaluate model fit and therefore significance of the results. A summary of all statistics, convergence values and posterior predictive tests, for both across-site composition (saturation/diversity test) and across-branch composition (global homogeneity test), is shown in Table S11, Table S12 and Table S13. The best fitting representative for the simple single-matrix models, the LG model, was outcompeted by the more complex models throughout the whole analysis. The saturation/diversity test revealed that in most of the cases either CAT + GTR or CAT + GTR + REC were the best fitting models. As expected, all models failed the global homogeneity test. However, the CAT + GTR + REC model, mitigating the effects of across-branch heterogeneity in a primitive way, performed best in modelling across-branch compositions in the majority of cases.

5 Discussion

The conclusion that ancient Chlamydiae facilitated primary endosymbiosis is not universally accepted. Some people believe these genes have much more complex histories. Proposals are that gene transfers took place among different prokaryotic lineages, including cyanobacteria, eventually combined with subsequent transfers into Archaeplastida and/or further into Chlamydiae (Brinkman et al., 2002; Dagan & Martin, 2009; Martin & Roettger, 2012). Since the improvement of phylogenetic methods never stops, many people started to reanalyze previously published studies. The goal is to either find further support for a particular hypothesis or to question it with newer results calculated on improved datasets using more sophisticated methods. For instance, it had long been assumed that the three domains hypothesis best describes the origin of eukaryotes. The three domains hypothesis assumes that eukaryotes emerged as the sister group to the monophyletic Archaea. However, a reanalysis by Rinke et al. in 2013, including methods and models that do account for the compositional heterogeneity found in a big dataset, lead to results supporting the competing eocyte hypothesis, in which core genes of eukaryotic cells originated from within the Archaea (Williams & Embley, 2014). This is just one case showing that the usage of a method or substitution model not fitting to the data investigated leads to wrong topologies in phylogenetic trees. In the course of this thesis, a reanalysis of previous studies supporting the ménage-à-trois was done to find out more about the role of ancient HGT. The datasets were updated and augmented with new sequences, if available, and the best-fitting methods and substitution models were used to increase reliability of the results.

5.1 The Chlamydiae and plants conundrum

In Table 2 it is shown that most of the trees created in the course of this study predicted a sister topology for Chlamydiae and plants. Regarding converged trees only, the pattern stays the same, although the proportion of apparent support for Chlamydiae as gene donor seems higher, as shown in Table 3. However, consideration of Bayesian posterior probabilities for the nodes of interest reveals that a lot of the trees predicting either a sister topology or donation from Chlamydiae to plants are discarded in case of all three models of main interest (Table 4).

Furthermore, the results in Table 2 show that most of the LG trees tend to show a sister topology instead of donation, as would be expected according to the previous findings. These results, together with the fact that so many LG trees predicting C to P were filtered out after considering Bayesian posterior probabilities (Table 4) and despite the higher proportion of C to P in Table 3, are very conflicting with the hypothesis questioned. All this shows outmost contradiction to what has been concluded in previous studies which also used simple substitution models, but in a maximum likelihood environment.
However, the complex models CAT + GTR and CAT + GTR + REC, which do take into account across-site heterogeneity, did not perform very differently to the LG model. The issue probably was that, due to time and resource problems, only few trees converged and had good support values in general, making it hard to compare in numbers, as shown in Table 4. Nevertheless, relaxing the conditions and also taking into account converged trees with weak Bayesian posterior probabilities for the nodes of interest, it is at least clearly visible that gene donation from Chlamydiae to plants has been observed in far less cases, where complex models have been applied, compared to the simple model (Table 3).

Besides gaining further insight into evolution and horizontal gene transfer, there is of course another, more systematic, but interesting point: The direct comparison of models applied for the different datasets let one conclude about their performance and how they fit to the question asked and data observed. Therefore, the LG model was compared with the CAT + GTR and CAT + GTR + REC model, separately for each dataset. In Table 5 it is shown that the majority of cases with an observation of identical topologies for different models are related to Chlamydiae and plants being sister to each other, with most overlap between LG and CAT + GTR. The CAT + GTR + REC model on the other side was able to refute the donation events proposed by the LG model more often. As expected, the LG model predicted gene transfers more frequently than the complex models for the different protein families, in general. However, to get a better idea of which model fits best to the data investigated in this study and which results, whether supporting ancient horizontal gene transfer events or not, are reliable or not, posterior predictive test results need to be taken into account.

Since for some protein families a second set of trees was prepared calculated on an augmented subset of the data used for first tree calculations it is also worth to compare their

topologies observed with those predicted by their forerunners. By doing so it is possible to find out if the smaller but augmented datasets used for subtree calculation helped to either gain resolution or confirm previously predicted results. It is of special interest if the direct augmentation and improvement of the datasets no. 6 (IspG), no. 17 (dapL), no. 20 (pc0378) and no. 24 (nhaD) lead to more significant results compared to the appropriate first tree. The subtrees which are most conform to their forerunners for the different models applied are the ones predicting a sister topology of plants and Chlamydiae (Table 6). Additionally, the proportion of the models that fall into the stable C | P group are relatively balanced between all three models, whereas the stable C to P topologies are mainly predicted by the use of the LG model (Table S10). However, proper analysis of these results, to conclude about significantly supported gene transfers or sister topologies for individual protein families shared by Chlamydiae and plants, is only possible when also interpreting the results shown in Table 7. In total, further support for 12 trees showing sister topologies and five donation predicting trees could be provided by subtree results. A look at Table S10 reveals that only the donation in the L,L-diaminopimelate aminotransferase family (no. 17, dapL) and the sister topology in the Polynucleotide phosphorylase family (no. 19, pnp) and the Na+/H+ antiporter NhaD family (no. 24, nhaD) were sufficiently supported by the overlap of not only the LG model, but rather the complex model results. Given that, it is shown that the sampling was already good enough in the first run. Nevertheless, the augmentation with new sequences indeed lead to further support for the firstly calculated trees in case of no. 17 (dapL) and no. 19 (pnp), representing further evidence of how important it is to do robust phylogenetic analyses by comparison of models and improvement of datasets to avoid doubtful conclusions. While for no. 17 (dapL) and no. 24 (nhaD) previous results could be confirmed, the direct augmentation of the data for no. 6 (ispG) and no. 20 (pc0378) in sum lead to trees that were not able to clearly resolve the relationship between plants and Chlamydiae anymore (Table 7).

Additionally, the fact that some trees from the first set refute a sister topology for a protein family proposed by the appropriate subtree and vice versa is worth mentioning (Table 6). Especially for no. 8 (ygo4) and no. 28 (yohl) the subtree approach lead to results showing C | P instead of the previously shown donations (Table 7). Nevertheless, a closer look at Table S10 reveals that not one of the cases with subtrees predicting C to P and first trees predicting C | P, or the reverse case, was calculated on recoded data that are assumed to provide the

most reliable results. Also considering that some of the particular recoded trees in Table 7 are struggling with P to C and C | P topologies (no. 28, yohl and no. 8, ygo4) or having problems to gain resolution (no. 18, pepF), this leads to the conclusion that either the simpler models are far more sensitive to the sampling (it is possible that too much information got lost by the subtree approach) and/or the sampling still needs to be improved (i.e. more data sequenced) to win resolution. The finding, that mainly complex models were applied in cases where the resulting topology was insensitive to sampling, represents further support for the assumed superiority of complex over simple substitution models when working in this context. Furthermore, the results for no. 13 (kdsB) all show a sister relationship, however, with weak support, also shown in Table 7. For no. 12 (CAB867) and no. 24 (nhaD) almost all models agreed on a sister topology, although with weak support of convergence values and Bayesian posterior probabilities. One case for relatively high support of the more unusual P to C is no. 4 (ispD).

5.2 The role of HGT in life

Dagan et al. showed in 2013 that plants actually share more genes with other bacteria than with Chlamydiae. It might be that the findings which the ménage-à-trois hypothesis is based on are simply a product of cumulative effects of systematic errors and several ancient HGT events among different kinds of cellular groups (Alsmark et al., 2013) having an impact on phylogenetic reconstruction. Some of the trees calculated in this study show suspicious positions of particular species for which extensive gene exchanges have been reported earlier.

As shown in Figure 8, some species belonging to the Oomycetes show basal grouping to the plant and Chlamydiae clades which are shown as sister to each other. Here, an eukaryotic origin is predicted strengthening observations from Richards et al. in 2006 showing that some Oomycetes have gained genes from fungi and even red algae. A similar observation has been made for the CAT + GTR subtree of no. 10 (pc1106), where the only representative of the Placozoa, *Trichoplax adhaerens*, is closely grouping with plant and chlamydial sequeces (the tree can be found in the Supplementary information 8.2.1). In a previous study the authors suggested a *Rickettsiae* endosymbiont for *T. adhaerens*, leading to the hypothesis that horizontal gene transfer between different domains also took place here (Driscoll et al.,

2013). Several other occasions have been reported where gene exchange lead to the acquirement of foreign genes, for instance some *Physcomitrella patens* genes probably originate in fungi, plants and/or even viruses (Yue et al., 2012). Other studies reporting HGT include several diatom species like *Thalassiosira pseudonana*, *Phaeodactylum tricornutum* or the green algae *Ostreococcus tauri* (Keeling & Palmer, 2008), which also have odd topologies in some of the results in this study.

Furthermore, evidence for gene transfers across all possible combinations of groups of organisms has been found. Moliner et al. stated in 2009, that Amoebae serve as "genetic melting pot" and that some HGT events have also been described between bacteria and viruses (Moreira & Brochier-Armanet, 2008; Thomas et al., 2011) - especially the giant *Acanthamoeba polyphaga mimivirus* whose genome consists of genes acquired from several bacteria. Additionally, gene exchanges between members of the *Legionellales* and *Chlamydiales,* intracellular amoebae-resistant endosymbionts living within common inclusions in their amoebal hosts, have been reported (Gimenez et al., 2011). Based on this knowledge, it is only compreshensible to conclude that possibly subsequent HGT between one of the endosymbionts and the amoebael host took place, thereby further obscuring the phylogenetic signal (Bertelli & Greub, 2012). The common strategies for horizontal gene transfer are conjugation, transformation and transduction (Baron, 1996). An extremely rare case would be a fourth form of horizontal gene transfer where eukaryote-to-eukaryote exchange happens (Archibald et al., 2003; Andersson, 2009). During the evolution of diatoms and dinoflagellates such events occured, enabled by secondary and tertiary endosymbioses.

An interesting idea is that viruses could act as mediators for gene transport. Figure 9 represents a case where a viral sequence grouped within a plant clade, provoking the idea that probably even viruses took part in the ancient events that are now causing the mysterious grouping of plants and Chlamydiae in so many single-gene analyses. An interesting study by Redrejo-Rodriguez et al. in 2012 reported on the dubious existence of nuclear localization signals in bacteriophages which are involved in HGT between the appropriate prokaryotic hosts and eukaryotic organisms associated. Due to the fact that bacteriophages only infect prokaryotes and the virus would never get the chance to directly get in touch with eukaryotes, except for Transformation (Keeling & Palmer, 2008), the suspicion comes up that the viruses somehow facilitate probable inter-domain gene transfers. Of course, the results

shown in this study do not support the idea that the virus in Figure 9 served as a means for gene transport from plants to Chlamydiae, nevertheless gene transfer from plants to the virus cannot be excluded.

In addition to that, the tree shown in Figure 9 is an example for plants that split up into two different clades, the one including the viral sequence and the second grouping with bacteria. Similar trees have been calculated on the data of no. 12 and no. 25 which can be found in the Supplementary information 8.2.1. However, there is a lot of evidence that plants are monophyletic. Therefore, these results further underline the extent of systematic errors these tree calculations suffered from.

The third special case shown in Figure 10 represents a tree where Chlamydiae could be interpreted as sister to plants if there were not those few, relatively long-branched, plants basal to the chlamydial sequences. However, lower plants are known to have GC-contents different to those of higher plants (Costantini et al., 2013) which are known to be AT-rich. Therefore, this observation can also be attributed to systematic errors, especially LBA, as these species are living in extremely acidic environments, leading to increased GC-contents. Additionally, there is also support for the theory that the red algae *Cyanidioschyzon merolae*, which is among the strange grouping plants in Figure 10, did undergo HGT events with non-organellar sources prior to the divergence of red algae and green plants (Huang & Gogarten, 2008).

5.3 Phylogenetic artifacts and systematic errors

Besides multiple HGT events it is postulated that systematic errors also lead to biased results (Brinkmann et al., 2005; Lartillot et al., 2007; Liu et al., 2014) in this thesis or even to wrong conclusions about a relationship between Chlamydiae and plants in previous studies.

5.3.1 Size matters: Effect of large datasets and short alignments

A reason for the models to have failed in resolving the phylogenies could be the enormous datasets (Hervé Philippe et al., 2011). The use of large datasets generally increases the resolution of phylogenetic trees; however, a strongly supported tree does not necessarily mean that it is true (Lartillot et al., 2007). Especially large datasets including sequences from the different domains of life are affected by systematic errors which can often lead to biased

results and conclusions (Nishihara et al., 2007). Features of large datasets causing these errors are, for instance, big evolutionary distances, eventually entailing functional shifts and shifts across-taxa heterogeneity (Philippe et al., 2003). Even the most complex CAT + GTR + REC model applied in this study does not take into account such high variation in substitution rates leading to systematic errors (Hirt et al., 1999). This could have effected and biased the positions of our plant and Chlamydiae species of interest. Smaller subtrees including our sequences of interest and some outgroup sequences were prepared for a reasonable selection of the investigated datasets to find out if the tree calculations in this study were also affected by these errors. It was attempted to ameliorate the effects of LBA and to stabilize the position of the Chlamydiae by adding even more chlamydial and plant sequences, especially sequences from the PVC superphylum, (Lagkouvardos et al., 2014; Wagner & Horn, 2006) to the subsets. In phylogenetics it is debated whether the greatest improvement in accuracy results from an increased number of characters (alignment length) or species (number of sequences aligned). As it is likely to produce more accurate results, assembling data rich in both characters and species is necessary (Delsuc et al., 2005). However, the combination of single-gene analysis with investigation of ancient horizontal gene transfers requiring datasets that consist of highly diverse sequences, does not allow much regulation considering that. No. 4 (ispD), no. 16 (rlmH), no. 22 (trpD), no. 23 (yfhC) and no. 26 (pc0141) are representatives for very short alignments observed in this study. A look at their trees reveals that indeed calculations on these data did not perform very well, as especially these cases no converged trees were obtained when calculating trees on the large original datasets (Table 7). In a study published in 2014 by Knie et al. the authors wrote about ancient horizontal gene transfer events of even several tRNA molecules from Chlamydiae to plants. However, according to the findings in this thesis the results shown by Knie et al. are most likely biased by the usage of inappropriate methods and models and the alignments observed contain not enough information to reconstruct ancient transfer events with today's methods.
In addition to that, the authors of a recently published study reported that filtering and trimming of alignments does not always lead to an improvement of subsequent phylogenetic reconstructions (Tan et al., 2015). Besides the higher risk of getting unresolved branches, this process also often causes an increased proportion of well-supported, but wrong branches (Talavera & Castresana, 2007).

5.3.2 Heterogeneity

Simple models assume homogeneity across sites of the alignment and thus branches of a tree (Goodfellow et al., 2014). Previously published studies on the ménage-à-trois hypothesis have made use of only simple single-matrix models like JTT (Huang & Gogarten, 2007), WAG (Becker et al., 2008; Moustafa et al., 2008) or LG (Ball et al., 2013), which all share the assumption of homogeneous evolutionary rates across sites of sequences throughout an alignment as well as between the individual sequences (Goodfellow et al., 2014). However, natural molecular sequences often violate several of these simplifying assumptions because of existing compositional variations across sites of an alignment and branches of a tree (Philippe & Roure, 2011; Philippe et al., 2011; Williams et al., 2011). More sophisticated models are needed when trying to investigate ancient gene transfer events, especially inter-domain transfers which require the observation of large datasets and are assumed in the ménage-à-trois hypothesis. If a model does not adequately account for the heterogeneity in the data observed it has a poor model fit (Roure & Philippe, 2011). Especially variation in base composition across sites of sequences is a very common feature of real molecular sequence data that arises from site-specific selective constraints (Le et al., 2008). Violation of the model assumptions by the data can lead to wrong, but highly supported phylogenetic predictions (Philippe et al., 2011) and conclusions caused by artifacts like the well-known and widespread long-branch attraction (LBA). Other biases can result in the artefactual grouping of species with similar GC-content, nucleotide or amino-acid composition, all having an impact on the codon usage of sequences (Behura & Severson, 2013; Inagaki & Roger, 2006; Suzuki, 2003). It has also been shown previously that there are major differences in the codon usage pattern between pathogenic and environmental chlamydiae (Zhou et al., 2006). Heterotachy (Lopez et al., 2002), a feature especially present in protein evolution, has been recently confirmed as an important process which can lead to phylogenetic reconstruction artefacts if underestimated by the model. Statistical tests like principal component analyses (PCA) of the amino acid usages or similar could help to evaluate which properties in particular might impact phylogenetic reconstruction, and to what extent (Su et al., 2009; Hsieh & Yang, 2008). One could conclude that the more similar the topologies observed in the PCA blots are to the tree topologies, the higher the probability that biased results have been obtained. Posterior predictive test statistics are useful to check for biochemical properties of the data and model fit (Baum & Smith, 2013; Gelman et al., 1996).

5.4 Testing model fit

Assessment of the Bayesian posterior probabilities is not very computationally demanding, as they are directly calculated on the original data (Huelsenbeck et al., 2001). Through the assessment of sampling effects these statistical indices represent the probability that the calculated tree is correct. The reliability of a tree is conditional on the data observed and the method applied, which should be able to handle compositional properties of the data. As a consequence, if the method does not correctly model the features of the data, a wrong and misleading tree can often receive strong statistical support (Delsuc et al., 2005). A method is statistically consistent, i.e. the model applied fits the data, if it converges towards the true value as more data is added (Delsuc et al., 2005; Ziheng, 2014). Simulation studies showed that when the assumptions of a substitution model are violated by the data it fails to capture the evolutionary complexity of the data (Kolaczkowski & Thornton, 2004). To finally check whether the assumptions of the models applied in this study are met by the observed data, or rather, finally reveal which of the models fits the data best, posterior predictive simulations (Bollback, 2002) have been performed for all of them and for each alignment that was used for tree calculations. A model that fits the data should be able to generate the data (Boussau et al., 2014). Comparison of the summary statistics computed on the true and simulated data should give an idea of the model fit (Huelsenbeck et al., 2001).

The results of the statistical tests show that the data were highly heterogeneous in both across-site and especially across-branch composition. Irrespective of a few exceptions, all models are rejected by the across-branch test, as expected. However, the Dayhoff recoding obviously was able to mitigate the effects of heterogeneity at least a bit and performed best with the *P-values* that were the closest to 0.5 in most of the datasets (Table S11, Table S12 and Table S13). A bit more interesting is the fact that in many cases the models tended to be incapable of reproducing the site-wise diversity present in the data. According to the statistics, the probability of overestimating the site-wise diversity increased with model simplicity and vice versa, a trend for increased underestimation could be reported the more complex the models got. In total, these results indicate that the site-homogeneous LG models and, to a lesser extent the CAT + GTR model, may fall prey to LBA. On the other hand, the CAT + GTR + REC model may overcorrect against LBA, what is in full accordance with the results from Boussau et al., 2014. Nevertheless, keeping in mind that all single gene datasets observed in this study consist of sequences with partially huge phylogenetic distances, it is

reasonable to say that the CAT + GTR + REC model is still the method that modeled the data best. However, better methods including across-branch and across-site heterogeneity still need to be developed.

5.5 Conclusion and Outlook

Altogether, these results show how difficult it is to find an answer to complex large-scale phylogenetic questions. In cases where, in addition to that, ancient gene transfer events are assumed, doing proper phylogenetic reconstruction is even harder and the extrememly sophisticated methods needed to address such questions are still in their infancy.

Even most of the single-gene analyses results presented here, where the best methods and models available for phylogenetic reconstruction were applied, gave no clear and consistent answer to the question how plants and Chlamydiae are related. Therefore, the findings of this thesis show that compositional heterogeneity in the dataset has been even more extremely underestimated in previous studies by the use of improper methods and models that do not adequately fit the data. The findings suffered from systematic errors, leading to biased results and possibly wrong conclusions about plant evolution. Of course, this study still also does not take into account the heterogeneity of the data that is truly present. However, with the new results it was possible to show probable weaknesses of previous approaches. Additionally, the findings point out that it is necessary to pay attention when concluding from trees calculated with inappropriate models or unfitting methods. It is also motivating evidence that the methods and models are more and more approximate to real evolution.

Nevertheless, the observation of plant and chlamydial genes grouping closely together in so many cases is indeed suspicious. However, this extensive gene exchange that possibly took place (Alsmark et al., 2013) does not necessarily mean that there was a ménage-à-trois. Perhaps it is more a sign of an ancient long-period infection, symbiosis or co-habitation of the same ecological niche (Domman et al., 2015; Loftus et al., 2005). Even with the robust phylogenetics applied here, previously published results on HGT between several organisms just gained weak support due to the fact that no consistent results were obtained. With the results of this study not much evidence for particular cases of HGT could be provided, but similarity to previous findings has been shown. Nevertheless, it cannot be excluded that some of the predicted HGTs are simply a product of systematic artifacts. It is very difficult to resolve especially single-gene phylogenies which are assumed to have undergone ancient gene

transfers, because so many different factors and processes have to be modelled properly to represent what really happened. For further, more reliable in-depth investigations on this topic I recommend the non-stationary CAT + BP model (Blanquart & Lartillot, 2008) which makes it possible to jointly model across-branch as well as across-site compositional variations in sequence evolution. This model accounts for across-site variation, a feature provided by the CAT model, and works in combination with a process in which composition can change at specific breakpoints (Blanquart & Lartillot, 2006) across the phylogenetic tree, therefore including variations in substitution rates across branches. Alternatively, P4 (p4.nhm.ac.uk/), a Python package for analysis of molecular sequences which can also use heterogeneous models for tree calculations (Foster, 2004), could be useful for further studies.

6 Zusammenfassung

Im Zuge dieser Masterarbeit wurde eine Auswahl an Proteinfamilien untersucht für welche Homologien in Pflanzen und Chlamydien gefunden wurden. Die Ergebnisse vorangehender Studien haben zur Vermutung veranlasst Chlamydien hätten die jeweiligen Gene zur Entstehungszeit der Pflanzen an diese durch horizontalen Gentransfer übertragen. Ziel war es durch Neuanalyse zu entschlüsseln ob es tatsächlich Interaktionen zwischen antiken Chlamydien und den Ureltern des Pflanzenreichs gab. Es wurden zahlreiche phylogenetische Bäume erstellt für deren Berechnung die bestmöglichen verfügbaren Methoden und Modelle angewandt wurden. Die neuen Ergebnisse liefern weit weniger Hinweise für diesen debattierten, urzeitlichen, exzessiven Genaustausch zwischen Pflanzen und Chlamydien. Individueller Genaustausch im Allgemeinen kann jedoch nicht ausgeschlossen werden. Es konnten sogar Ergebnisse gewonnen werden die bereits aufgestellte Hypothesen über horizontalen Genaustausch zwischen den verschiedensten Organismen wieder ins Leben rufen. Transfers könnten demnach nicht nur innerhalb oder zwischen Bakterien und Eukaryoten stattgefunden haben. Viren könnten eine bedeutende Rolle gespielt haben oder multipler Genaustausch einzelner Gene stattgefunden haben. Das zusätzliche Einwirken von Milliarden von Jahren an Evolution würde schließlich die Schwierigkeit der Rekonstruktion der wahren Geschichte dieser Proteine erklären. Denn es wurden auch einige phylogenetische Bäume berechnet die unerwartete Topologien, Bayesian posterior probabilities oder starke Polytomien zeigen. Im Rahmen der Fragestellung dieser Arbeit konnte weiters die Überlegenheit von komplexen mixture models gegenüber dem einfachen LG Modell gezeigt werden. Während das simple Modell von homogenen Substitutionsraten ausgeht beziehen komplexere Alternativen realitätsnahere, heterogene Muster entlang eines Alignments mit ein. Insbesondere die Kombination des CAT + GTR Modells mit der Dayhoff recoding Strategie war nützlich um selbst dem verzerrenden Effekt der Heterogenität zwischen Sequenzen entgegenzuwirken. Das umkodierte Modell hat sich als nützlichstes erwiesen um an eine derart komplexe, weit zurückblickende und domänenübergreifende Fragestellung heranzugehen. Die Notwendigkeit weiterer, besserer Methoden und Modelle für noch reellere Simulation von Evolution ist allerdings unabstreitbar und deren Entwicklung wird zukünftig sicher für weitere, vielleicht überraschende Aufschlüsse über die Entstehung alles Lebens sorgen.

7 References

7.1. Links

tree.bio.ed.ac.uk/software/figtree/

p4.nhm.ac.uk/

7.2. Literature

Abascal, F., Zardoya, R., & Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics (Oxford, England)*, *21*(9), 2104–5. http://doi.org/10.1093/bioinformatics/bti263

All, B. E. D. H. (1997). The origin of red algae : Implications for plastid evolution, *94*(April), 4520–4525.

Alsmark, C., Foster, P. G., Sicheritz-Ponten, T., Nakjang, S., Martin Embley, T., & Hirt, R. P. (2013). Patterns of prokaryotic lateral gene transfers affecting parasitic microbial eukaryotes. *Genome Biology*, *14*(2), R19. http://doi.org/10.1186/gb-2013-14-2-r19

Altschup, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic Local Alignment Search Tool, 403–410. http://doi.org/10.1016/S0022-2836(05)80360-2

Archibald, J. M., Rogers, M. B., Toop, M., Ishida, K.-I., & Keeling, P. J. (2003). Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga Bigelowiella natans. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 7678–7683. http://doi.org/10.1073/pnas.1230951100

Ball, S., Colleoni, C., Cenci, U., Raj, J. N., & Tirtiaux, C. (2011). The evolution of glycogen and starch metabolism in eukaryotes gives molecular clues to understand the establishment of plastid endosymbiosis. *Journal of Experimental Botany*, *62*(6), 1775–801. http://doi.org/10.1093/jxb/erq411

Ball, S. G., Subtil, A., Bhattacharya, D., Moustafa, A., Weber, A. P. M., Gehre, L., … Dauvillée, D. (2013). Metabolic effectors secreted by bacterial pathogens: essential facilitators of plastid endosymbiosis? *The Plant Cell*, *25*(1), 7–21. http://doi.org/10.1105/tpc.112.101329

Baum, D. (2013). The origin of primary plastids: a pas de deux or a ménage à trois? *The Plant Cell*, *25*, 4–6. http://doi.org/10.1105/tpc.113.109496

Becker, B., Hoef-Emden, K., & Melkonian, M. (2008). Chlamydial genes shed light on the evolution of photoautotrophic eukaryotes. *BMC Evolutionary Biology*, *8*, 203. http://doi.org/10.1186/1471-2148-8-203

Behura, S. K., & Severson, D. W. (2013). Codon usage bias: Causative factors, quantification methods and genome-wide patterns: With emphasis on insect genomes. *Biological Reviews*, *88*, 49–61. http://doi.org/10.1111/j.1469-185X.2012.00242.x

Bertelli, C., & Greub, G. (2012). Lateral gene exchanges shape the genomes of amoeba-resisting microorganisms. *Frontiers in Cellular and Infection Microbiology*, *2*(August), 1–15. http://doi.org/10.3389/fcimb.2012.00110

Bhattacharya, D., Yoon, H. S., & Hackett, J. D. (2004). Photosynthetic eukaryotes unite: endosymbiosis connects the dots. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, *26*(1), 50–60. http://doi.org/10.1002/bies.10376

Blanquart, S., & Lartillot, N. (2006). A Bayesian compound stochastic process for modeling

nonstationary and nonhomogeneous sequence evolution. *Molecular Biology and Evolution*, *23*(Foster), 2058–2071. http://doi.org/10.1093/molbev/msl091

Blanquart, S., & Lartillot, N. (2008). A site- and time-heterogeneous model of amino acid replacement. *Molecular Biology and Evolution*, *25*, 842–858. http://doi.org/10.1093/molbev/msn018

Bodył, A., Mackiewicz, P., & Stiller, J. W. (2007). The intracellular cyanobacteria of Paulinella chromatophora: endosymbionts or organelles? *Trends in Microbiology*, *15*(7), 295–296. http://doi.org/10.1016/j.tim.2007.05.002

Bollback, J. P. (2002). Bayesian model adequacy and choice in phylogenetics. *Molecular Biology and Evolution*, *19*(Wilks 1938), 1171–1180. http://doi.org/10.1093/oxfordjournals.molbev.a004175

Boussau, B., Walton, Z., Delgado, J. a., Collantes, F., Beani, L., Stewart, I. J., … Huelsenbeck, J. P. (2014). Strepsiptera, Phylogenomics and the Long Branch Attraction Problem. *PLoS ONE*, *9*(10), e107709. http://doi.org/10.1371/journal.pone.0107709

Bretz, J. R., & Hutcheson, S. W. (2004). MINIREVIEW Role of Type III Effector Secretion during Bacterial Pathogenesis in Another Kingdom, *72*(7), 3697–3705. http://doi.org/10.1128/IAI.72.7.3697

Brinkman, F. S. L., Blanchard, J. L., Cherkasov, A., Av-Gay, Y., Brunham, R. C., Fernandez, R. C., … Greberg, H. (2002). Evidence that plant-like genes in Chlamydia species reflect an ancestral relationship between Chlamydiaceae, cyanobacteria, and the chloroplast. *Genome Research*, *12*(8), 1159–67. http://doi.org/10.1101/gr.341802

Brinkmann, H., van der Giezen, M., Zhou, Y., Poncelin de Raucourt, G., & Philippe, H. (2005). An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Systematic Biology*, *54*(5), 743–757. http://doi.org/10.1080/10635150500234609

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, *17*, 540–552. http://doi.org/10.1093/oxfordjournals.molbev.a026334

Coburn, B., Sekirov, I., & Finlay, B. B. (2007). Type III Secretion Systems and Disease, *20*(4), 535–549. http://doi.org/10.1128/CMR.00013-07

Collingro, A., Tischler, P., Weinmaier, T., Penz, T., Heinz, E., Brunham, R. C., … Horn, M. (2011). Unity in variety--the pan-genome of the Chlamydiae. *Molecular Biology and Evolution*, *28*(12), 3253–70. http://doi.org/10.1093/molbev/msr161

Collingro, A., Toenshoff, E. R., Taylor, M. W., Fritsche, T. R., Wagner, M., & Horn, M. (2005). “Candidatus Protochlamydia amoebophila”, an endosymbiont of Acanthamoeba spp. *International Journal of Systematic and Evolutionary Microbiology*, *55*(Pt 5), 1863–1866. http://doi.org/10.1099/ijs.0.63572-0

Corsaro, D., & Greub, G. (2006). Pathogenic Potential of Novel Chlamydiae and Diagnostic Approaches to Infections Due to These Obligate Intracellular Bacteria Pathogenic Potential of Novel Chlamydiae and Diagnostic Approaches to Infections Due to These Obligate Intracellular Bacteria, *19*(2). http://doi.org/10.1128/CMR.19.2.283

Corsaro, D., Pages, G. S., Catalan, V., Loret, J.-F., & Greub, G. (2010). Biodiversity of amoebae and amoeba-associated bacteria in water treatment plants. *International Journal of Hygiene and Environmental Health*, *213*(3), 158–66. http://doi.org/10.1016/j.ijheh.2010.03.002

Corsaro, D., & Venditti, D. (2009). Detection of Chlamydiae from freshwater environments by PCR, amoeba coculture and mixed coculture. *Research in Microbiology*, *160*(8), 547–52. http://doi.org/10.1016/j.resmic.2009.08.001

Corsaro, D., Venditti, D., & Valassina, M. (2002). New parachlamydial 16S rDNA phylotypes detected in human clinical samples. *Research in Microbiology*, *153*(9), 563–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12455703

Costantini, M., Alvarez-Valin, F., Costantini, S., Cammarano, R., & Bernardi, G. (2013). Compositional patterns in the genomes of unicellular eukaryotes. *BMC Genomics*, *14*(1), 755. http://doi.org/10.1186/1471-2164-14-755

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 Evolutionary Biology*, *10*, 210. http://doi.org/10.1186/1471-2148-10-210

Dagan, T., & Martin, W. (2009). Getting a better picture of microbial evolution en route to a network of genomes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*, 2187–2196. http://doi.org/10.1098/rstb.2009.0040

Dagan, T., Roettger, M., Stucken, K., Landan, G., Koch, R., Major, P., … Martin, W. F. (2013). Genomes of Stigonematalean cyanobacteria (subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids. *Genome Biology and Evolution*, *5*(1), 31–44. http://doi.org/10.1093/gbe/evs117

Delsuc, F., Brinkmann, H., & Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Reviews. Genetics*, *6*(May), 361–375. http://doi.org/10.1038/nrg1603

Deschamps, P. (2014). Primary endosymbiosis: have cyanobacteria and Chlamydiae ever been roommates? *Acta Societatis Botanicorum Poloniae*, *83*(4), 291–302. http://doi.org/10.5586/asbp.2014.048

Deschamps, P., Colleoni, C., Nakamura, Y., Suzuki, E., Putaux, J.-L., Buléon, A., … Ball, S. (2008). Metabolic symbiosis and the birth of the plant kingdom. *Molecular Biology and Evolution*, *25*(3), 536–48. http://doi.org/10.1093/molbev/msm280

Dimijian, G. G. (2000). Evolving together: the biology of symbiosis, part 1. *Proceedings (Baylor University. Medical Center)*, *13*, 217–226.

Domman, D., Horn, M., Embley, T. M., & Williams, T. a. (2015). Plastid establishment did not require a chlamydial partner. *Nature Communications*, *6*, 1–8. http://doi.org/10.1038/ncomms7421

Driscoll, T., Gillespie, J. J., Nordberg, E. K., Azad, A. F., & Sobral, B. W. (2013). Bacterial DNA sifted from the Trichoplax adhaerens (Animalia: Placozoa) genome project reveals a putative rickettsial endosymbiont. *Genome Biology and Evolution*, *5*(4), 621–645. http://doi.org/10.1093/gbe/evt036

Eissenberg, L. G., & Wyrick, P. B. (1981). Inhibition of phagolysosome fusion is localized to Chlamydia psittaci-laden vacuoles. *Infection and Immunity*, *32*(2), 889–96. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=351526&tool=pmcentrez&rendertype=abstract

Everett, K. D., Bush, R. M., & Andersen, a a. (1999). Emended description of the order Chlamydiales, proposal of Parachlamydiaceae fam. nov. and Simkaniaceae fam. nov., each containing one monotypic genus, revised taxonomy of the family Chlamydiaceae, including a new genus and five new species, and standards. *International Journal of Systematic Bacteriology*, *49 Pt 2*(1 999), 415–40. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10319462

Facchinelli, F., Colleoni, C., Ball, S. G., & Weber, A. P. M. (2013). Chlamydia, cyanobiont, or host: who was on top in the ménage à trois? *Trends in Plant Science*, *18*(12), 673–9. http://doi.org/10.1016/j.tplants.2013.09.006

Faires, M. C., Gehring, E., Mergl, J., & Weese, J. S. (2009). Methicillin-resistant Staphylococcus aureus in marine mammals. *Emerging Infectious Diseases*, *15*(12), 2071–2. http://doi.org/10.3201/eid1512.090220

Foster, P. G. (2004). Modeling compositional heterogeneity. *Systematic Biology*, *53*(3), 485–495. http://doi.org/10.1080/10635150490445779

Fuerst, J. A. (2013). The PVC superphylum : exceptions to the bacterial definition ?, 451–466. http://doi.org/10.1007/s10482-013-9986-1

Gelman, a, Gelman, a, Meng, X.-L., Meng, X.-L., Stern, H., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. Vol.6, No.4. *Statistica Sinica*, *6*(4), 733–807. http://doi.org/10.1.1.142.9951

Gimenez, G., Bertelli, C., Moliner, C., Robert, C., Raoult, D., Fournier, P.-E., & Greub, G. (2011). Insight into cross-talk between intra-amoebal pathogens. *BMC Genomics*, *12*(1), 542. http://doi.org/10.1186/1471-2164-12-542

Gogarten, J. P., Doolittle, W. F., & Lawrence, J. G. (1999). Prokaryotic Evolution in Light of Gene Transfer, 2226–2238.

Gould, S. B., Waller, R. F., & McFadden, G. I. (2008). Plastid evolution. *Annual Review of Plant Biology*, *59*, 491–517. http://doi.org/10.1146/annurev.arplant.59.032607.092915

Gross, J., & Bhattacharya, D. (2009). Mitochondrial and plastid evolution in eukaryotes: an outsiders' perspective. *Nature Reviews. Genetics*, *10*(7), 495–505. http://doi.org/10.1038/nrg2610

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307–321. http://doi.org/10.1093/sysbio/syq010

Gupta, R. S., Bhandari, V., & Naushad, H. S. (2012). Molecular signatures for the pvc clade (planctomycetes, verrucomicrobia, chlamydiae, and lentisphaerae) of bacteria provide insights into their evolutionary relationships. *Frontiers in Microbiology*, *3*(September), 1–19. http://doi.org/10.3389/fmicb.2012.00327

Haider, S., Collingro, A., Walochnik, J., Wagner, M., & Horn, M. (2008a). Chlamydia -like bacteria in respiratory samples of community- acquired pneumonia patients, *281*, 198–202. http://doi.org/10.1111/j.1574-6968.2008.01099.x

Haider, S., Collingro, A., Walochnik, J., Wagner, M., & Horn, M. (2008b). Chlamydia-like bacteria in respiratory samples of community-acquired pneumonia patients. *FEMS Microbiology Letters*, *281*(2), 198–202. http://doi.org/10.1111/j.1574-6968.2008.01099.x

Hammerschlag, M. R. (2002). The intracellular life of chlamydiae. *Seminars in Pediatric Infectious Diseases*, *13*(4), 239–48. http://doi.org/10.1053/spid.2002.127201

Heinzen, R. A., Scidmore, M. A., Rockey, D. D., & Hackstadt, T. (1996). Differential interaction with endocytic and exocytic pathways distinguish parasitophorous vacuoles of Coxiella burnetii and Chlamydia trachomatis. *Infection and Immunity*, *64*(3), 796–809. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=173840&tool=pmcentrez&rendertype=abstract

Hirt, R. P., Logsdon, J. M., Healy, B., Dorey, M. W., Doolittle, W. F., & Embley, T. M. (1999). Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(January), 580–585. http://doi.org/10.1073/pnas.96.2.580

Horn, M. (2008). Chlamydiae as symbionts in eukaryotes. *Annual Review of Microbiology*, *62*,

113–31. http://doi.org/10.1146/annurev.micro.62.081307.162818
Horn, M., Collingro, A., Schmitz-Esser, S., Beier, C. L., Purkhold, U., Fartmann, B., … Wagner, M. (2004). Illuminating the evolutionary history of chlamydiae. *Science (New York, N.Y.)*, *304*(5671), 728–30. http://doi.org/10.1126/science.1096330
Horn, M., Wagner, M., & Santic, M. (2005). MINIREVIEW Amoebae as Training Grounds for Intracellular Bacterial Pathogens, *71*(1), 20–28. http://doi.org/10.1128/AEM.71.1.20
Hrdy, I., Hirt, R. P., Dolezal, P., Bardonová, L., Foster, P. G., Tachezy, J., & Embley, T. M. (2004). Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. *Nature*, *432*(December), 618–622. http://doi.org/10.1038/nature03149
Huang, J., & Gogarten, J. P. (2007). Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? *Genome Biology*, *8*(6), R99. http://doi.org/10.1186/gb-2007-8-6-r99
Huang, J., & Gogarten, J. P. (2008). Concerted gene recruitment in early plant evolution. *Genome Biology*, *9*, R109. http://doi.org/10.1186/gb-2008-9-7-r109
Huelsenbeck, J. P., Joyce, P., Lakner, C., & Ronquist, F. (2008). Bayesian analysis of amino acid substitution models. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1512), 3941–3953. http://doi.org/10.1098/rstb.2008.0175
Huelsenbeck, J. P., Rannala, B., & Masly, J. P. (n.d.). An Introduction to Bayesian Inference of Phylogeny. *DNA Sequence*, 1–7.
Huelsenbeck, J. P., Ronquist, F., Nielsen, R., & Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science (New York, N.Y.)*, *294*(December), 2310–2314. http://doi.org/10.1126/science.1065889
Hsieh, K. L., Yang, I. C. (2008) Incorporating PCA and fuzzy-ART techniques into achieve organism classification based on codon usage consideration doi:http://dx.doi.org/10.1016/j.compbiomed.2008.05.007
Inagaki, Y., & Roger, A. J. (2006). Phylogenetic estimation under codon models can be biased by codon usage heterogeneity. *Molecular Phylogenetics and Evolution*, *40*, 428–434. http://doi.org/10.1016/j.ympev.2006.03.020
Jørgensen, F. G., Schierup, M. H., & Clark, A. G. (2007). Heterogeneity in regional GC content and differential usage of codons and amino acids in GC-poor and GC-rich regions of the genome of Apis mellifera. *Molecular Biology and Evolution*, *24*, 611–619. http://doi.org/10.1093/molbev/msl190
Keeling, P. J., & Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nature Reviews. Genetics*, *9*(august), 605–618. http://doi.org/10.1038/nrg2386
Knie, N., Polsakiewicz, M., & Knoop, V. (2014). Horizontal gene transfer of chlamydial-like tRNA genes into early vascular plant mitochondria, *49*(0), 1–21.
Koga, R., Tsuchida, T., & Fukatsu, T. (2003). Changing partners in an obligate symbiosis: a facultative endosymbiont can compensate for loss of the essential endosymbiont Buchnera in an aphid. *Proceedings. Biological Sciences / The Royal Society*, *270*(April), 2543–2550. http://doi.org/10.1098/rspb.2003.2537
Kolaczkowski, B., & Thornton, J. W. (2004). Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature*, *431*(October), 980–984. http://doi.org/10.1038/nature02917
Lagkouvardos, I., Jehl, M.-A., Rattei, T., & Horn, M. (2014). Signature protein of the PVC superphylum. *Applied and Environmental Microbiology*, *80*(2), 440–5. http://doi.org/10.1128/AEM.02655-13
Lagkouvardos, I., Weinmaier, T., Lauro, F. M., Cavicchioli, R., Rattei, T., & Horn, M. (2014).

Integrating metagenomic and amplicon databases to resolve the phylogenetic and ecological diversity of the Chlamydiae. *The ISME Journal*, *8*(1), 115–25. http://doi.org/10.1038/ismej.2013.142

Lartillot, N. (2004). A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. *Molecular Biology and Evolution*, *21*(6), 1095–1109. http://doi.org/10.1093/molbev/msh112

Lartillot, N., Brinkmann, H., & Philippe, H. (2007). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evolutionary Biology*, *7 Suppl 1*, S4. http://doi.org/10.1186/1471-2148-7-S1-S4

Lartillot, N., Lepage, T., & Blanquart, S. (2009). PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics (Oxford, England)*, *25*(17), 2286–8. http://doi.org/10.1093/bioinformatics/btp368

Lartillot, N., Rodrigue, N., Stubbs, D., & Richer, J. (2013). PhyloBayes-MPI A Bayesian software for phylogenetic reconstruction using mixture models MPI version, 1–21.

Le, S. Q., & Gascuel, O. (2008). An improved general amino acid replacement matrix. *Molecular Biology and Evolution*, *25*, 1307–1320. http://doi.org/10.1093/molbev/msn067

Le, S. Q., Lartillot, N., & Gascuel, O. (2008). Phylogenetic mixture models for proteins. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *363*(October), 3965–3976. http://doi.org/10.1098/rstb.2008.0180

Li, S., Nosenko, T., Hackett, J. D., & Bhattacharya, D. (2006). Phylogenomic analysis identifies red algal genes of endosymbiotic origin in the chromalveolates. *Molecular Biology and Evolution*, *23*(3), 663–74. http://doi.org/10.1093/molbev/msj075

Liu, Y., Cox, C. J., Wang, W., & Goffinet, B. (2014). Mitochondrial Phylogenomics of Early Land Plants: Mitigating the Effects of Saturation, Compositional Heterogeneity, and Codon-usage Bias. *Systematic Biology*, *63*(6), 862–878. http://doi.org/10.1093/sysbio/syu049

Loftus, B., Anderson, I., Davies, R., Alsmark, U. C. M., Samuelson, J., Amedeo, P., … Hall, N. (2005). The genome of the protist parasite Entamoeba histolytica. *Nature*, *433*, 865–868. http://doi.org/10.1038/nature03291

Lopez, P., Casane, D., & Philippe, H. (2002). Heterotachy, an important process of protein evolution. *Molecular Biology and Evolution*, *19*, 1–7. http://doi.org/10.1093/oxfordjournals.molbev.a003973

Marin, B., Nowack, E. C. M., & Melkonian, M. (2005). A plastid in the making: evidence for a second primary endosymbiosis. *Protist*, *156*(4), 425–32. http://doi.org/10.1016/j.protis.2005.09.001

Martin, W., Goremykin, V., & Hansmann, S. (1998). Gene transfer to the nucleus and the evolution of chloroplasts, *393*(MAY), 162–165.

Martin, W., Herrmann, R. G., Braunschweig, D.-, & Institut, B. (1998). Update on Gene Transfer from Organelles to the Nucleus Gene Transfer from Organelles to the Nucleus : How Much , What Happens , and Why ? 1, 9–17.

Martin, W., & Roettger, M. (2012). Modern endosymbiotic theory : Getting lateral gene transfer into the equation. *Journal of Endocytobiosis and Cell Research*, *23*, 1–5.

Mcfadden, G. I. (2001). Chloroplast Origin and Integration 1, *125*(January), 50–53.

McFadden, G. I., & van Dooren, G. G. (2004). Evolution: red algal genome affirms a common origin of all plastids. *Current Biology : CB*, *14*(13), R514–6. http://doi.org/10.1016/j.cub.2004.06.041

Moliner, C., Raoult, D., & Fournier, P. E. (2009). Evidence of horizontal gene transfer between amoeba and bacteria. *Clinical Microbiology and Infection*, *15*, 178–180.

http://doi.org/10.1111/j.1469-0691.2008.02216.x
Moreira, D., & Brochier-Armanet, C. (2008). Giant viruses, giant chimeras: the multiple evolutionary histories of Mimivirus genes. *BMC Evolutionary Biology*, *8*, 12. http://doi.org/10.1186/1471-2148-8-12
Moustafa, A., Reyes-Prieto, A., & Bhattacharya, D. (2008). Chlamydiae has contributed at least 55 genes to Plantae with predominantly plastid functions. *PloS One*, *3*(5), e2205. http://doi.org/10.1371/journal.pone.0002205
Neal, R. M. (1998). Probabilistic Inference Using Markov Chain Monte Carlo Methods. *Technical Report*, *1*, 1–144. http://doi.org/10.1021/np100920q
Nguyen, V.-A., Boyd-Graber, J., & Altschul, S. F. (2013). Dirichlet mixtures, the Dirichlet process, and the structure of protein space. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, *20*(1), 1–18. http://doi.org/10.1089/cmb.2012.0244
Nishihara, H., Okada, N., & Hasegawa, M. (2007). Rooting the eutherian tree: the power and pitfalls of phylogenomics. *Genome Biology*, *8*(9), R199. http://doi.org/10.1186/gb-2007-8-9-r199
Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, *302*, 205–217. http://doi.org/10.1006/jmbi.2000.4042
Ohta, N., Matsuzaki, M., Misumi, O., Miyagishima, S. Y., Nozaki, H., Tanaka, K., … Shin, I. T. (2003). Complete sequence and analysis of the plastid genome of the unicellular red alga Cyanidioschyzon merolae.[erratum appears in DNA Res. 2003 Jun 30;10(3):137]. *DNA Research*, *10*, 67–77. http://doi.org/10.1093/dnares/10.2.67
Penel, S., Arigon, A.-M., Dufayard, J.-F., Sertier, A.-S., Daubin, V., Duret, L., … Perrière, G. (2009). Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*, *10 Suppl 6*, S3. http://doi.org/10.1186/1471-2105-10-S6-S3
Petersen, J., Teich, R., Brinkmann, H., & Cerff, R. (2006). A “green” phosphoribulokinase in complex algae with red plastids: evidence for a single secondary endosymbiosis leading to haptophytes, cryptophytes, heterokonts, and dinoflagellates. *Journal of Molecular Evolution*, *62*(2), 143–57. http://doi.org/10.1007/s00239-004-0305-3
Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., & Baurain, D. (2011). Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS Biology*, *9*(3). http://doi.org/10.1371/journal.pbio.1000602
Philippe, H., Casane, D., Gribaldo, S., Lopez, P., & Meunier, J. (2003). Heterotachy and functional shift in protein evolution. *IUBMB Life*, *55*(4-5), 257–265. http://doi.org/10.1080/1521654031000123330
Philippe, H., & Roure, B. (2011). Difficult phylogenetic questions: more data, maybe; better methods, certainly. *BMC Biology*, *9*(1), 91. http://doi.org/10.1186/1741-7007-9-91
Price, D. C., Chan, C. X., Yoon, H. S., Yang, E. C., Qiu, H., Weber, A. P. M., … Bhattacharya, D. (2012). Cyanophora paradoxa genome elucidates origin of photosynthesis in algae and plants. *Science (New York, N.Y.)*, *335*(6070), 843–7. http://doi.org/10.1126/science.1213561
Qiu, H., Price, D. C., Weber, A. P. M., Facchinelli, F., Yoon, H. S., & Bhattacharya, D. (2013). Assessing the bacterial contribution to the plastid proteome. *Trends in Plant Science*, *18*(12), 680–7. http://doi.org/10.1016/j.tplants.2013.09.007
Quang, L. S., Gascuel, O., & Lartillot, N. (2008). Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics*, *24*(20), 2317–2323. http://doi.org/10.1093/bioinformatics/btn445

Redrejo-Rodriguez, M., Munoz-Espin, D., Holguera, I., Mencia, M., & Salas, M. (2012). Functional eukaryotic nuclear localization signals are widespread in terminal proteins of bacteriophages. *Proceedings of the National Academy of Sciences*, 3–8. http://doi.org/10.1073/pnas.1216635109

Reyes-Prieto, A., & Bhattacharya, D. (2007). Phylogeny of nuclear-encoded plastid-targeted proteins supports an early divergence of glaucophytes within Plantae. *Molecular Biology and Evolution*, *24*(11), 2358–61. http://doi.org/10.1093/molbev/msm186

Reyes-Prieto, A., Hackett, J. D., Soares, M. B., Bonaldo, M. F., & Bhattacharya, D. (2006). Cyanobacterial contribution to algal nuclear genomes is primarily limited to plastid functions. *Current Biology : CB*, *16*(23), 2320–5. http://doi.org/10.1016/j.cub.2006.09.063

Reyes-Prieto, A., Weber, A. P. M., & Bhattacharya, D. (2007). The origin and establishment of the plastid in algae and plants. *Annual Review of Genetics*, *41*, 147–68. http://doi.org/10.1146/annurev.genet.41.110306.130134

Ringner, M. (2008). What is principal component analysis? *Nat Biotechnol*, *26*(3), 303–304. http://doi.org/10.1038/nbt0308-303

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J.-F., … Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, *499*(7459), 431–7. http://doi.org/10.1038/nature12352

Rockwell, N. C., Lagarias, J. C., & Bhattacharya, D. (2014). Primary endosymbiosis and the evolution of light and oxygen sensing in photosynthetic eukaryotes. *Frontiers in Ecology and Evolution*, *2*(October), 1–13. http://doi.org/10.3389/fevo.2014.00066

Rodríguez-Ezpeleta, N., Brinkmann, H., Burey, S. C., Roure, B., Burger, G., Löffelhardt, W., … Lang, B. F. (2005). Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. *Current Biology : CB*, *15*(14), 1325–30. http://doi.org/10.1016/j.cub.2005.06.040

Roure, B. & Philippe, H. (2011). Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. *BMC Evolutionary Biology*, *11*(1), 17. http://doi.org/10.1186/1471-2148-11-17

Schmitz-Esser, S., Toenshoff, E. R., Haider, S., Heinz, E., Hoenninger, V. M., Wagner, M., & Horn, M. (2008). Diversity of bacterial endosymbionts of environmental acanthamoeba isolates. *Applied and Environmental Microbiology*, *74*(18), 5822–5831. http://doi.org/10.1128/AEM.01093-08

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., … Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*(539). http://doi.org/10.1038/msb.2011.75

Smith, J. M., & Smith, N. H. (1996). Synonymous nucleotide divergence: What is “saturation”? *Genetics*, *142*(3), 1033–1036.

Stride, M. C., Polkinghorne, a, Miller, T. L., Groff, J. M., Lapatra, S. E., & Nowak, B. F. (2013). Molecular characterization of “Candidatus Parilichlamydia carangidicola,” a novel Chlamydia-like epitheliocystis agent in yellowtail kingfish, Seriola lalandi (Valenciennes), and the proposal of a new family, “Candidatus Parilichlamydiaceae” fam. nov. (or. *Applied and Environmental Microbiology*, *79*(5), 1590–7. http://doi.org/10.1128/AEM.02899-12

Su, M., Lin, H., Yuan, H. S., & Chu, W. (2009). Categorizing Host-Dependent RNA Viruses by Principal Component Analysis of Their Codon Usage Preferences, *16*(11), 1539–1547.

Subtil, A., Collingro, A., & Horn, M. (2014). Tracing the primordial Chlamydiae: extinct parasites of plants? *Trends in Plant Science*, *19*(1), 36–43.

http://doi.org/10.1016/j.tplants.2013.10.005
Susko, E., & Roger, A. J. (2007). On reduced amino acid alphabets for phylogenetic inference. *Molecular Biology and Evolution*, *24*, 2139–2150. http://doi.org/10.1093/molbev/msm144
Suzuki, H. (2003). Heterogeneity in Synonymous Codon Usage among Genes of Diverse Bacterial Genomes. *Bioinformatics*, *451*, 450–451.
Suzuki, K., & Miyagishima, S. (2010). Eukaryotic and eubacterial contributions to the establishment of plastid proteome estimated by large-scale phylogenetic analyses. *Molecular Biology and Evolution*, *27*(3), 581–90. http://doi.org/10.1093/molbev/msp273
Talavera, G., & Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, *56*(4), 564–577. http://doi.org/10.1080/10635150701472164
Tan, G., Muffato, M., Ledergerber, C., Herrero, J., Goldman, N., Gil, M., & Dessimoz, C. (2015). Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference, 1–33.
Thomas, V., Bertelli, C., Collyn, F., Casson, N., Telenti, A., Goesmann, A., … Greub, G. (2011). Lausannevirus, a giant amoebal virus encoding histone doublets. *Environmental Microbiology*, *13*, 1454–1466. http://doi.org/10.1111/j.1462-2920.2011.02446.x
Tree, T. H. E., & Eukaryotes, O. F. (2004). D Iversity and Evolutionary History of Plastids and Their Hosts 1. *American Journal of Botany*, *91*(10), 1481–1493.
Tseng, T.-T., Tyler, B. M., & Setubal, J. C. (2009). Protein secretion systems in bacterial-host associations, and their description in the Gene Ontology. *BMC Microbiology*, *9*(Suppl 1), S2. http://doi.org/10.1186/1471-2180-9-S1-S2
Wagner, M., & Horn, M. (2006). The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance. *Current Opinion in Biotechnology*, *17*(3), 241–9. http://doi.org/10.1016/j.copbio.2006.05.005
Waterhouse, A. M., Procter, J. B., Martin, D. M. a, Clamp, M., & Barton, G. J. (2009). Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics*, *25*(9), 1189–1191. http://doi.org/10.1093/bioinformatics/btp033
Wheeler, D. L., Church, D. M., Edgar, R., Federhen, S., Helmberg, W., Madden, T. L., … Wagner, L. (2004). Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Research*, *32*(October 2008), D35–D40. http://doi.org/10.1093/nar/gkh073
Whelan, S., & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution*, *18*(5), 691–699. http://doi.org/10.1093/oxfordjournals.molbev.a003851
Williams, D., Fournier, G. P., Lapierre, P., Swithers, K. S., Green, A. G., Andam, C. P., & Gogarten, J. P. (2011). A Rooted Net of Life. *Biology Direct*, *6*(1), 45. http://doi.org/10.1186/1745-6150-6-45
Williams, T. a, & Embley, T. M. (2014). Archaeal "dark matter" and the origin of eukaryotes. *Genome Biology and Evolution*, *6*(3), 474–81. http://doi.org/10.1093/gbe/evu031
Yang, Z., & Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nature Reviews. Genetics*, *13*(5), 303–14. http://doi.org/10.1038/nrg3186
Yoon, H. S., Hackett, J. D., Ciniglia, C., Pinto, G., & Bhattacharya, D. (2004). A molecular timeline for the origin of photosynthetic eukaryotes. *Molecular Biology and Evolution*, *21*(5), 809–18. http://doi.org/10.1093/molbev/msh075
Yue, J., Hu, X., Sun, H., Yang, Y., & Huang, J. (2012). Widespread impact of horizontal gene transfer on plant colonization of land. *Nature Communications*, *3*, 1152.

http://doi.org/10.1038/ncomms2148
Zhou, T., Sun, X., & Lu, Z. (2006). Synonymous codon usage in environmental chlamydia UWE25 reflects an evolutional divergence from pathogenic chlamydiae. *Gene*, *368*, 117–125. http://doi.org/10.1016/j.gene.2005.10.035
Zimorski, V., Ku, C., Martin, W. F., & Gould, S. B. (2014). Endosymbiotic theory for organelle origins. *Current Opinion in Microbiology*, *22*, 38–48. http://doi.org/10.1016/j.mib.2014.09.008

8 Supplementary information

8.1 Supplementary Materials and Methods

8.1.1 Manipulation of datasets and subtrees

Table S1: Overview showing which datasets were filtered with T-Coffee. The trim option was used to remove redundant sequences. The filter was set to either 70, 80, 85, 90 or 95% sequence identity. The last three columns show the number of sequences before and after the filtering step as well as the number of sequences used for subtree calculations. Most of the subtrees were calculated on real subsets of the original data, augmented with some more plant, PVC and/or bacterial sequences. In case of the anyway small samples no. 6, 17, 20 and 24 new sequences were directly added to the dataset used for calculation of the first tree.

no.	gene name	protein name	gi:	ref	HOGENOM	C to P	T-Coffee filter	no. seq. orig.	no. seq. fil.	no. seq. sub.
1	ntt_3	ATP/ADP translocase	46445874	YP_007239.1	HOG000238123			166		138
2	mdh	malate dehydrogenase	46447406	YP_008771.1	HOG000220953			469		204
3	ispE	probable isopentenyl monophosphate kinase	46447223	YP_008588.1	HOG000019600		95%	697	510	92
4	ispD	2-C-Methyl-d-erythritol 4-phosphate cytidylyltransferase	46445961	YP_007326.1	HOG000218563		95%	587	405	97
5	fabF	3-Oxoacyl-(acyl-carrier-protein) synthase	298537950	YP_008237.1	HOG000060166			1660		
6	ispG	4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase	76788770	YP_007739.1	HOG000261018			99		132
7	gutQ	FOG: CBS domain	46401057	YP_008781.1	HOG000264729			881		
8	ygo4	Phosphate Permease	15618590	NP_224876.1	HOG000231892		95%	744	425	69
9	gpmA	phosphoglyceromutase	46445795	YP_007160.1	HOG000221682		90%	1292	502	
10	pc1106	probable isoamylase	46446740	YP_008105.1	HOG000239197		95%	781	528	177
11	tyrS	Tyrosyl-tRNA synthetase	76788775	YP_008168.1	HOG000242790		95%	918	547	148
12	CAB867	Cation transport ATPase	62185469	YP_220254.1	HOG000250399		70%	1366	612	75
13	kdsB	cytidylyltransferase	89898215	YP_515325.1	HOG000007602		95%	753	502	
14	plsB	glycerol-3-phosphate acyltransferase	76789550	YP_001654267	HOG000034930			27		
15	pc0324	hypothetical protein pc0324	46445958	YP_007323.1	HOG000084053			14		
16	rlmH	hypothetical protein pc1708	46447342	YP_008707.1	HOG000218433		95%	584	392	158
17	dapL	L,L-diaminopimelate aminotransferase	46446319	YP_007684.1	HOG000223061			162		188
18	pepF	Oligoendopeptidase F	46400453	YP_008177.1	HOG000059490			346		101
19	pnp	polynucleotide phosphorylase/polyadenylase	46446277	YP_007642.1	HOG000218326		95%	893	530	90
20	pc0378	Predicted sulfur transferase	46399653	YP_007377.1	HOG000034811			194		244
21	rsmH	probable S-adenosyl-methyltransferase	46445945	YP_007310.1	HOG000049778		95%	884	483	108
22	trpD	anthranilate phosphoribosyltransferase	89898246	YP_515356.1	HOG000230451		80%	1142	596	
23	yfhC	cytosine/adenosine deaminases	89898061	YP_515171.1	HOG000085050		70%	1395	573	
24	nhaD	Na+/H+ antiporter NhaD and related arsenite permeases	297620457	YP_003708594.1	HOG000251774			74		112
25	tgt	queuine tRNA-ribosyltransferase	46446428	YP_007793.1	HOG000223473		85%	1315	533	
26	pc0141	rRNA methylases	46445775	YP_007140.1	HOG000218799		95%	797	554	
27	miaA	tRNA delta(2)-isopentenylpyrophosphate transferase	46446877	YP_008242.1	HOG000039995		95%	795	573	73
28	yohl	tRNA-dihydrouridine synthase	46400854	YP_008578.1	HOG000217854			249		75

8.1.2 Alignments used for tree calculations

https://figshare.com/s/6f379e508fd9fbbb313b

8.2 Supplementary Results

8.2.1 Collection of all trees

https://figshare.com/s/6f379e508fd9fbbb313b

8.2.2 Tree analysis: Observed topologies and model comparison

Table S2: Overview of the results of all 174 trees and subtrees. Subtrees are indicated by .1, connecting the no. of the protein family and the model used. C | P indicates sister topology, C to P and P to C stand for donation of the gene from C to P and vice versa, NR, grouping indicates no resolution but Chlamydiae and plants are grouping closely and NR means no resolution at all.

C | P			C to P		P to C	NR, grouping		NR
10.1_gtr	24_catfix	21_gtr	11_cat	11.1_gtr	10.1_lg	1.1_gtr	25_lg	16.1_rec
10_catfix	24_gtr	21_lg	17.1_gtr	11.1_lg	10.1_rec	1.1_lg	4_cat	16_rec
10_gtr	24_rec	23_lg	17.1_lg	12.1_gtr	10_cat	1.1_rec	4_catfix	18.1_gtr
11_rec	26_cat	24.1_gtr	17.1_rec	16_catfix	10_lg	1_cat	6.1_gtr	18.1_rec
13_cat	26_catfix	24.1_rec	17_catfix	8_cat	16.1_lg	1_catfix	6.1_rec	23_cat
13_catfix	26_lg	24_cat	17_gtr	8_catfix	2_lg	1_gtr	9_cat	23_gtr
13_gtr	27.1_gtr	12_rec	17_lg	8_gtr	22_cat	1_lg	9_catfix	23_rec
13_lg	27.1_lg	8_rec	17_rec		22_gtr	1_rec	9_gtr	27.1_rec
13_rec	27_cat	20.1_rec	18.1_lg		23_catfix	10_rec	9_lg	
17_cat	27_catfix	8.1_gtr	20_cat		27_rec	11_catfix	9_rec	
18_cat	27_gtr	8.1_lg	20_catfix		28_rec	11_gtr	22_catfix	
18_catfix	27_lg	11.1_rec	20_gtr		3.1_gtr	11_lg	22_lg	
18_gtr	28.1_gtr		20_lg		3.1_lg	16.1_gtr	22_rec	
18_lg	28.1_lg		20_rec		3.1_rec	16_cat		
18_rec	28.1_rec		24.1_lg		4.1_gtr	16_gtr		
19.1_gtr	28_cat		24_lg		4.1_rec	16_lg		
19.1_lg	3_rec		26_gtr		4_rec	2.1_gtr		
19.1_rec	4_gtr		26_rec		8.1_rec	2.1_lg		
19_cat	4_lg		28_catfix			2.1_rec		
19_catfix	6_cat		28_gtr			2_gtr		
19_gtr	6_catfix		28_lg			20.1_gtr		
19_lg	6_gtr		3_cat			20.1_lg		
19_rec	6_rec		3_catfix			21.1_rec		
2_cat	12.1_lg		3_gtr			21_cat		
2_catfix	12.1_rec		3_lg			21_rec		
2_rec	12_catfix		4.1_lg			25_cat		
21.1_gtr	12_catfix		6.1_lg			25_catfix		
21.1_lg	12_gtr		6_lg			25_gtr		
21_catfix	12_lg		8_lg			25_rec		

Table S3: Overview of the results of all converged trees. Subtrees are indicated by .1, connecting the no. of the protein family and the model used. C | P indicates sister topology, C to P and P to C stand for donation of the gene from C to P and vice versa, NR, grouping indicates no resolution but Chlamydiae and plants are grouping closely and NR means no resolution at all.

C | P		C to P	P to C	NR, grouping	NR
6_rec	21.1_lg	4.1_lg	3.1_rec	1_lg	16.1_rec
6_cat	8.1_lg	6.1_lg	3.1_gtr	1_catfix	18.1_rec
6_gtr	8.1_gtr	6_lg	3.1_lg	1_cat	18.1_gtr
17_cat	11.1_rec	17.1_rec	4.1_rec	1_gtr	27.1_rec
19.1_rec	12.1_rec	17.1_gtr	4.1_gtr	1_rec	
19.1_gtr	12.1_lg	17_lg	10.1_lg	1.1_lg	
19.1_lg		17_rec	16.1_lg	1.1_gtr	
24_rec		17_gtr	8.1_rec	1.1_rec	
24_gtr		18.1_lg		2.1_gtr	
24_catfix		24_lg		2.1_lg	
24_cat		24.1_lg		6.1_rec	
24.1_gtr		28_lg		6.1_gtr	
28.1_lg		20_lg		16.1_gtr	
28.1_rec		20_cat		20.1_lg	
28.1_gtr		20_rec		21.1_rec	
28_cat		20_gtr			
27.1_gtr		11.1_gtr			
27.1_lg		12.1_gtr			
21.1_gtr		11.1_lg			

Table S4: Overview of the results of all converged trees that have Bayesian posterior probabilities of at least 0.8 for the nodes leading to conclusions about the relationship between plants and Chlamydiae. Subtrees are indicated by .1, connecting the no. of the protein family and the model used. C | P indicates sister topology, C to P and P to C stand for donation of the gene from C to P and vice versa, NR, grouping indicates no resolution but Chlamydiae and plants are grouping closely. Results showing no resolution at all are not listed.

C | P		C to P	P to C	NR, grouping	
6_rec	28_cat	4.1_lg	3.1_rec	6.1_rec	1_rec
6_cat	27.1_gtr	6.1_lg	3.1_gtr	6.1_gtr	1.1_lg
11.1_lg	27.1_lg	17.1_gtr	4.1_rec	20.1_lg	1.1_gtr
24_rec	21.1_lg	17_lg	4.1_gtr	16.1_gtr	1.1_rec
24_gtr	21.1_gtr	17_rec	10.1_lg	21.1_rec	
24_catfix		17_gtr	16.1_lg	1_lg	
24_cat		20_cat		1_catfix	
24.1_gtr				1_cat	
28.1_lg				1_gtr	

Table S5: Overview of the results of all trees that did not converge. Subtrees are indicated by .1, connecting the no. of the protein family and the model used. C | P indicates sister topology, C to P and P to C stand for donation of the gene from C to P and vice versa, NR, grouping indicates no resolution but Chlamydiae and plants are grouping closely and NR means no resolution at all.

C | P		C to P	P to C	NR, grouping		NR
19_rec	21_lg	17.1_lg	10.1_rec	16_cat	11_catfix	23_rec
6_catfix	11_rec	20_catfix	4_rec	22_rec	11_lg	16_rec
18_catfix	4_lg	28_gtr	22_cat	20.1_gtr	25_gtr	23_gtr
26_lg	21_gtr	11_cat	22_gtr	2.1_rec	25_lg	23_cat
27_gtr	13_cat	26_rec	28_rec	21_rec		
19_catfix	10_catfix	28_catfix	23_catfix	21_cat		
18_rec	13_lg	17_catfix	10_cat	9_rec		
27_lg	2_rec	3_gtr	27_rec	22_catfix		
2_catfix	10_gtr	8_lg	2_lg	4_cat		
2_cat	19_lg	3_cat	10_lg	9_cat		
10.1_gtr	19_cat	26_gtr		4_catfix		
24.1_rec	19_gtr	3_lg		25_rec		
18_gtr	18_lg	3_catfix		25_catfix		
13_rec	21_catfix	8_catfix		2_gtr		
3_rec	27_catfix	8_cat		22_lg		
27_cat	12_rec	16_catfix		25_cat		
4_gtr	12_catfix	8_gtr		9_catfix		
13_gtr	12_gtr			16_gtr		
13_catfix	12_lg			16_lg		
26_cat	12_cat			9_lg		
18_cat	8_rec			9_gtr		
26_catfix	20.1_rec			10_rec		
23_lg				11_gtr		

Table S6: Summary of the observed topologies of all trees and subtrees that did not converge. Each tree, analyzed here by the respective model applied, has been assigned to one of the five topology categories. C | P indicates sister topology, C to P and P to C stand for donation of the gene from C to P and vice versa, NR, grouping indicates no resolution but Chlamydiae and plants are grouping closely and NR means no resolution at all.

	C | P	C to P	P to C	NR, grouping	NR
LG	9	3	2	5	0
CAT	7	3	2	5	1
CAT60	10	6	1	5	0
CAT + GTR	10	3	1	6	1
CAT + GTR + REC	10	1	4	6	2
	46	**16**	**10**	**27**	**4**

Table S7: Overview of the results of all trees that did not converge but have high Bayesian posterior probabilities of at least 0.8 for the nodes leading to conclusions about the relationship between plants and Chlamydiae. Subtrees are indicated by .1, connecting the no. of the protein family and the model used. C | P indicates sister topology, C to P and P to C stand for donation of the gene from C to P and vice versa, NR, grouping indicates no resolution but Chlamydiae and plants are grouping closely. Results showing no resolution at all are not listed.

C | P			C to P	P to C	NR, grouping	
19_rec	21_lg	12_rec	17.1_lg	10.1_rec	16_cat	25_catfix
6_catfix	11_rec	12_catfix	20_catfix	22_gtr	22_rec	2_gtr
18_catfix	4_lg	12_gtr	28_gtr	28_rec	20.1_gtr	22_lg
26_lg	21_gtr	12_lg	11_cat	23_catfix	2.1_rec	25_cat
27_gtr	13_lg	12_cat	3_gtr	10_cat	21_rec	9_catfix
19_catfix	2_rec	23_lg	8_lg	2_lg	21_cat	16_gtr
18_rec	10_gtr	27_cat	3_lg	10_lg	9_rec	16_lg
27_lg	19_lg	4_gtr	3_catfix		22_catfix	9_lg
2_catfix	19_cat	13_gtr	8_catfix		4_cat	9_gtr
2_cat	19_gtr	13_catfix	8_cat		9_cat	10_rec
10.1_gtr	18_lg	26_cat	8_gtr		4_catfix	11_gtr
24.1_rec	21_catfix	18_cat			25_rec	11_catfix
18_gtr	27_catfix	26_catfix			25_gtr	11_lg
					25_lg	

Table S8: Summary of the observed topologies of all trees and subtrees that did not converge converge but have high Bayesian posterior probabilities of at least 0.8 for the nodes leading to conclusions about the relationship between plants and Chlamydiae.. Each tree, analyzed here by the respective model applied, has been assigned to one of the five topology categories. C | P indicates sister topology, C to P and P to C stand for donation of the gene from C to P and vice versa, NR, grouping indicates no resolution but Chlamydiae and plants are grouping closely. Results showing no resolution at all are not listed.

	C | P	C to P	P to C	NR, grouping
LG	9	3	2	5
CAT	6	2	1	5
CAT60	9	3	1	5
CAT + GTR	10	2	1	6
CAT + GTR + REC	6	0	2	6
	40	**10**	**7**	**27**

Table S9: Direct comparison of the observed topologies for the most important models: the representative for single-matix models, LG, the most complex model not taking into account across-branch heterogeneity, CAT + GTR, and the CAT + GTR + REC model as a primitive method to model heterogeneity across branches. Subtrees are indicated by .1, connecting the no. of the protein family and the model used. C | P indicates sister topology, C to P and P to C stand for donation of the gene from C to P and vice versa, NR, grouping indicates no resolution but Chlamydiae and plants are grouping closely and NR means no resolution at all.

	LG C | P	LG C to P	LG P to C	LG NR, grouping	GTR C | P	GTR C to P	GTR P to C	GTR NR, grouping	GTR NR
REC C | P	28.1, 19, 19.1, 18, 13, 12, 12.1	6, 3, 24, 24.1, 20, 17, 17.1, 11.1	2	11, 20.1	6, 28.1, 24, 24.1, 19, 19.1, 18, 13, 12	3, 12.1, 11.1		2, 20.1	11
REC C to P	26	8				26, 8, 20, 17, 17.1			
REC P to C	4, 27, 8.1	4.1, 28,	3.1, 10.1		4, 27, 10.1, 8.1	28	4.1, 3.1		
REC NR, grouping	21, 21.1	6.1	10	25, 22, 2.1, 1, 1.1	21, 21.1, 10		22	6.1, 25, 2.1, 1, 1.1	
REC NR	27.1, 23	18.1,	16.1,	9, 16	27.1			9, 16, 16.1	23, 18.1
GTR C | P	4, 28.1, 27, 27.1, 21, 21.1, 19, 19.1, 18, 13, 12, 8.1	6, 24, 24.1	10.1, 10,						
GTR C to P	26, 12.1	3, 8, 28, 20, 17, 17.1, 11.1							
GTR P to C		4.1	3.1	22					
GTR NR, grouping		6.1	16.1, 2	16, 9, 25, 2.1, 1, 1.1, 11, 20.1					

Table S10: Direct model comparison of all original trees and subtrees. Subtrees are indicated by .1, connecting the no. of the protein family and the model used. C | P indicates sister topology, C to P and P to C stand for donation of the gene from C to P and vice versa, NR, grouping indicates no resolution but Chlamydiae and plants are grouping closely and NR means no resolution at all.

		---------- Subtrees ----------				
		C | P	C to P	P to C	NR, grouping	NR
---------- Originals ----------	C | P	11_rec, 19_rec, 24_rec, 12_rec, 10_gtr, 19_gtr, 21_gtr, 24_gtr, 27_gtr, 19_lg, 21_lg, 12_lg, 27_lg	12_gtr, 18_lg, 4_lg	3_rec, 8_rec, 4_gtr	2_rec, 6_rec, 6_gtr, 20_gtr	18_rec, 18_gtr
	C to P	28_gtr, 8_gtr, 28_lg, 8_lg	17_rec, 17_gtr, 17_lg, 24_lg, 6_lg	3_gtr, 3_lg	20_lg, 20_rec	
	P to C	28_rec		10_lg, 4_rec	2_lg	27_rec
	NR, grouping		11_gtr, 11_lg	10_rec, 16_lg	1_rec, 21_rec, 1_gtr, 16_gtr, 2_gtr, 1_lg	
	NR					16_rec

8.2.3 Statistics and posterior predictive tests

Table S11: Summary of the results of posterior predictive tests and bpcomp. The best fitting model for each original set and subset according to posterior predictive results is marked in red. Good convergence values are bold.

no	protein	dataset	model	saturation/diversity test:	across-site composition		homogeneity - global test:	across-branch composition		convergence
				observed diversity	posterior predicted	*P-value*	observed homogeneity	mean predicted	*P-value*	max. diff.
1	ATP/ADP translocase	orig.	LG	6,569	7.63846 +/- 0.122674	0,000	0,007	0,005	0,015	**0,30**
		sub.	LG	7,309	8.79285 +/- 0.123042	0,000	0,008	0,005	0,013	**0,30**
		orig.	CAT	6,569	6.72637 +/- 0.0982029	0,057	0,007	0,005	0,030	**0,30**
		orig.	C60	6,569	6.73882 +/- 0.0969779	0,036	0,007	0,005	0,096	**0,22**
		orig.	CGTR	6,569	6.71774 +/- 0.104982	0,073	0,007	0,005	0,018	**0,29**
		sub.	GTR	7,309	7.4836 +/- 0.113298	0,059	0,008	0,005	0,004	**0,24**
		orig.	CGREC	3,231	3.2065 +/- 0.0570345	0,650	0,016	0,005	0,000	**0,22**
		sub.	CGREC	3,433	3.41417 +/- 0.0558541	0,639	0,012	0,006	0,003	**0,27**
2	malate dehydrogenase	orig.	LG	8,012	10.4773 +/- 0.133826	0,000	0,023	0,012	0,004	1,00
		sub.	LG	7,354	9,36358 +/- 0,132883	0,000	0,020	0,009	0,000	**0,20**
		orig.	CAT	8,012	7.67321 +/- 0.119262	0,996	0,023	0.0131814 +/- 0.00239219	0,001	0,69
		orig.	C60	8,012	8.42794 +/- 0.13364	0,000	0,023	0.0131789 +/- 0.00243763	0,001	0,68
		orig.	CGTR	8,012	8.149 +/- 0.119604	0,121	0,023	0,012	0,001	0,96
		sub.	GTR	7,354	7,4936 +/- 0,117229	0,108	0,020	0,009	0,000	**0,20**
		orig.	CGREC	3,692	3.65347 +/- 0.059953	0,728	0,042	0,014	0,000	1,00
		sub.	CGREC	3,563	3,5341 +/- 0,0598516	0,681	0,040	0,010	0,000	0,41
3	probable isopentenyl monophosphate kinase (IPK)	orig.	LG	11,242	14.3661 +/- 0.137195	0,000	0,034	0,013	0,000	1,00
		sub.	LG	6,913	8,82695 +/- 0,228281	0,000	0,024	0,015	0,008	**0,10**
		orig.	CAT	11,242	11.3147 +/- 0.123132	0,267	0,034	0,013	0,000	0,92
		orig.	C60	11,242	12.2182 +/- 0.115112	0,000	0,034	0,014	0,000	1,00
		orig.	CGTR	11,242	11.9185 +/- 0.225635	0,002	0,034	0,013	0,000	0,73
		sub.	GTR	6,913	7,22584 +/- 0,199309	0,055	0,024	0,014	0,010	**0,07**
		orig.	CGREC	4,464	4.4876 +/- 0.0652962	0,346	0,035	0,015	0,000	0,68
		sub.	CGREC	3,609	3,56007 +/- 0,105503	0,671	0,037	0,019	0,007	**0,09**
4	2-C-Methyl-d-erythritol 4-phosphate cytidylyltransferase	orig.	LG	11,625	13.4983 +/- 0.146732	0,000	0,031	0.0202317 +/- 0.00192929	0,001	0,99
		sub.	LG	7,486	8,90699 +/- 0,217656	0,000	0,019	0,015	0,136	**0,10**
		orig.	CAT	11,625	11.684 +/- 0.135313	0,325	0,031	0,013	0,000	0,91
		orig.	C60	11,625	12.3983 +/- 0.134872	0,000	0,031	0,013	0,000	0,93
		orig.	CGTR	11,625	11.9973 +/- 0.169707	0,014	0,031	0.0223768 +/- 0.0028907	0,010	0,74
		sub.	GTR	7,486	7,74648 +/- 0,198688	0,089	0,019	0,015	0,110	**0,26**
		orig.	CGREC	4,514	4.83559 +/- 0.127187	0,002	0,024	0.0221474 +/- 0.00411517	0,260	0,52
		sub.	CGREC	3,626	3,54945 +/- 0,0951414	0,782	0,017	0,016	0,350	**0,20**
6	4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase	orig.	LG	5,879	7.22074 +/- 0.123654	0,000	0,010	0,004	0,000	**0,17**
		sub.	LG	6,118	7,85552 +/- 0,128288	0,000	0,013	0,004	0,000	**0,13**
		orig.	CAT	5,879	6.0363 +/- 0.0992333	0,046	0,010	0,004	0,000	**0,16**
		orig.	C60	5,879	6.3068 +/- 0.0960404	0,000	0,010	0,004	0,000	0,32
		orig.	CGTR	5,879	6.11724 +/- 0.106463	0,012	0,010	0,004	0,000	**0,16**
		sub.	GTR	6,118	6,38038 +/- 0,102841	0,006	0,013	0,004	0,000	**0,23**
		orig.	CGREC	3,309	3.32763 +/- 0.0595701	0,371	0,010	0,005	0,015	**0,22**
		sub.	CGREC	3,335	3,34244 +/- 0,0587995	0,429	0,009	0,005	0,027	**0,18**
8	Phosphate Permease	orig.	LG	9,938	12.2718 +/- 0.130136	0,000	0,019	0,009	0,000	0,91
		sub.	LG	6,250	7,12782 +/- 0,130071	0,000	0,009	0,006	0,033	**0,10**
		orig.	CAT	9,938	9.95728 +/- 0.102599	0,418	0,019	0,009	0,001	0,39
		orig.	C60	9,938	10.8473 +/- 0.105344	0,000	0,019	0,009	0,002	0,36
		orig.	CGTR	9,938	10.3178 +/- 0.134432	0,001	0,019	0,009	0,001	1,00
		sub.	GTR	6,250	6,37151 +/- 0,116805	0,139	0,009	0,006	0,026	**0,07**
		orig.	CGREC	4,074	3.97255 +/- 0.0598643	0,951	0,013	0,009	0,096	0,97
		sub.	CGREC	3,079	3,02583 +/- 0,0628281	0,800	0,007	0,007	0,494	**0,09**
9	phosphoglyceromutase	orig.	LG	11,064	13.3313 +/- 0.134739	0,000	0,022	0,010	0,000	1,00
		orig.	CAT	11,064	10.8436 +/- 0.123743	0,960	0,022	0,009	0,000	0,92
		orig.	C60	11,064	11.6718 +/- 0.11663	0,000	0,022	0,009	0,000	0,99
		orig.	CGTR	11,064	11.4079 +/- 0.140917	0,005	0,022	0,009	0,001	1,00
		orig.	CGREC	4,518	4.42229 +/- 0.0608171	0,930	0,013	0,011	0,160	0,89

Table S12: Summary of the results of posterior predictive tests and bpcomp. The best fitting model for each original set and subset according to posterior predictive results is marked in red. Good convergence values are bold.

no	protein	dataset	model	saturation/diversity test:	across-site composition		homogeneity - global test:	across-branch composition		convergence
10	probable isoamylase	orig.	LG	11,775	14.5161 +/- 0.0771883	0,000	0,015	0,004	0,000	1,00
		sub.	LG	8,287	10,639 +/- 0,118596	0,000	0,008	0,004	0,003	**0,26**
		orig.	CAT	11,775	11.7656 +/- 0.0740416	0,547	0,015	0,004	0,000	0,99
		orig.	C60	11,775	12.5822 +/- 0.0705473	0,000	0,015	0,004	0,000	0,90
		orig.	CGTR	11,775	12.2228 +/- 0.0914712	0,000	0,015	0,003	0,000	1,00
		sub.	GTR	8,287	8,41587 +/- 0,114283	0,128	0,008	0,004	0,002	0,37
		orig.	CGREC	4,665	5.01917 +/- 0.0889449	0,000	0,010	0.00893463 +/- 0.001718	0,193	1,00
		sub.	CGREC	3,932	3,86351 +/- 0,05398	0,895	0,013	0,005	0,000	0,48
11	Tyrosyl-tRNA synthetase	orig.	LG	10,632	14.0071 +/- 0.116957	0,000	0,031	0,006	0,000	1,00
		sub.	LG	7,518	9,65831 +/- 0,144341	0,000	0,011	0,006	0,004	**0,21**
		orig.	CAT	10,632	10.725 +/- 0.0914503	0,153	0,031	0,006	0,000	0,51
		orig.	C60	10,632	12.1241 +/- 0.0874198	0,000	0,031	0,006	0,000	1,00
		orig.	CGTR	10,632	11.3479 +/- 0.118808	0,000	0,031	0,006	0,000	1,00
		sub.	GTR	7,518	7,84601 +/- 0,128379	0,004	0,011	0,006	0,002	**0,30**
		orig.	CGREC	4,278	4.28374 +/- 0.0494069	0,451	0,013	0,007	0,010	0,99
		sub.	CGREC	3,566	3,54257 +/- 0,0614998	0,638	0,009	0,007	0,099	**0,13**
12	Cation transport ATPase	orig.	LG	11,960	15.9616 +/- 0.0840774	0,000	0,029	0,006	0,000	1,00
		sub.	LG	6,878	8,31746 +/- 0,121074	0,000	0,014	0,005	0,000	**0,08**
		orig.	CAT	11,960	12.0892 +/- 0.0753213	0,038	0,029	0,006	0	1,00
		orig.	C60	11,960	14.2444 +/- 0.0838762	0,000	0,029	0,006	0	0,58
		orig.	CGTR	11,960	12.7792 +/- 0.19386	0,000	0,029	0.011531 +/- 0.00179335	0,000	1,00
		sub.	GTR	6,878	7,0654 +/- 0,104812	0,032	0,014	0,005	0,001	**0,04**
		orig.	CGREC	4,367	4.38859 +/- 0.0457774	0,307	0,020	0,005	0	0,58
		sub.	CGREC	3,249	3,20711 +/- 0,0572664	0,761	0,007	0,006	0,315	**0,10**
13	cytidylyltransferase	orig.	LG	10,665	13.7086 +/- 0.14089	0,000	0,021	0,010	0,000	1,00
		orig.	CAT	10,665	10.7142 +/- 0.127222	0,342	0,021	0,010	0,000	0,89
		orig.	C60	10,665	11.87 +/- 0.118632	0,000	0,021	0,010	0,001	0,86
		orig.	CGTR	10,665	11.2116 +/- 0.155576	0,000	0,021	0,010	0,000	0,83
		orig.	CGREC	4,392	4.39019 +/- 0.0643508	0,496	0,025	0,012	0,001	0,60
16	hypothetical protein pc1708	orig.	LG	11,075	14.481 +/- 0.208634	0,000	0,033	0,020	0,001	1,00
		sub.	LG	8,195	10,6297 +/- 0,253808	0,000	0,022	0,020	0,195	**0,10**
		orig.	CAT	11,075	11.098 +/- 0.167834	0,429	0,033	0,019	0,003	0,68
		orig.	C60	11,075	12.2177 +/- 0.16583	0,000	0,033	0,019	0,003	0,68
		orig.	CGTR	11,075	11.6247 +/- 0.188637	0,001	0,033	0,019	0,001	1,00
		sub.	GTR	8,195	8,39671 +/- 0,209544	0,157	0,022	0,019	0,171	**0,28**
		orig.	CGREC	4,386	4.45504 +/- 0.0850201	0,187	0,042	0,023	0,005	0,93
		sub.	CGREC	3,688	3,68669 +/- 0,107234	0,489	0,025	0,021	0,174	**0,29**
17	L,L-diaminopimelate aminotransferase	orig.	LG	7,459	9.68349 +/- 0.113028	0,000	0,012	0,004	0,002	**0,27**
		sub.	LG	6,821	9,24947 +/- 0,131864	0,000	0,011	0,005	0,000	0,44
		orig.	CAT	7,459	7.62916 +/- 0.0893501	0,033	0,012	0,004	0,001	**0,29**
		orig.	C60	7,459	7.62496 +/- 0.0888982	0,028	0,012	0,004	0,002	0,36
		orig.	CGTR	7,459	8.09757 +/- 0.179879	0,000	0,012	0.00587081 +/- 0.00070572	0,000	**0,27**
		sub.	GTR	6,821	7,039 +/- 0,118049	0,027	0,011	0,004	0,000	**0,21**
		orig.	CGREC	3,511	3.52607 +/- 0.0534771	0,376	0,014	0,005	0,003	**0,30**
		sub.	CGREC	1,248	3,36631 +/- 0,0611967	0,333	0,011	0,006	0,012	**0,26**
18	Oligoendopeptidase F	orig.	LG	9,140	12.0757 +/- 0.0983642	0,000	0,015	0,004	0,000	1,00
		sub.	LG	7,581	9,10925 +/- 0,131159	0,000	0,010	0,005	0,005	**0,10**
		orig.	CAT	9,140	9.37725 +/- 0.0835416	0,002	0,015	0,004	0,000	0,98
		orig.	C60	9,140	10.3913 +/- 0.0798522	0,000	0,015	0,004	0,000	0,41
		orig.	CGTR	9,140	9.54105 +/- 0.0908996	0,000	0,015	0,004	0,000	0,31
		sub.	GTR	7,581	7,75886 +/- 0,11704	0,063	0,010	0,006	0,005	**0,11**
		orig.	CGREC	3,854	3.86359 +/- 0.0413176	0,397	0,015	0,005	0,000	0,46
		sub.	CGREC	3,524	3,5227 +/- 0,0574825	0,500	0,007	0,006	0,233	**0,13**

Table S13: Summary of the results of posterior predictive tests and bpcomp. The best fitting model for each original set and subset according to posterior predictive results is marked in red. Good convergence values are bold

no	protein	dataset	model	saturation/diversity test:	across-site composition		homogeneity - global test:	across-branch composition		convergence
19	polynucleotide phosphorylase/polyadenylase	orig.	LG	9,335	13.028 +/- 0.0878783	0,000	0,014	0,004	0,000	1,00
		sub.	LG	5,989	8,02548 +/- 0,109324	0,000	0,004	0,004	0,402	**0,07**
		orig.	CAT	9,335	9.44318 +/- 0.0688111	0,055	0,014	0,004	0,000	1,00
		orig.	C60	9,335	11.11 +/- 0.0682809	0,000	0,014	0,004	0,000	0,46
		orig.	CGTR	9,335	13.0322 +/- 0.0906127	0,000	0,014	0,004	0,000	1,00
		sub.	GTR	5,989	6,28993 +/- 0,088961	0,000	0,004	0,004	0,412	**0,23**
		orig.	CGREC	3,951	3.961 +/- 0.0369918	0,399	0,009	0,004	0,007	0,31
		sub.	CGREC	3,190	3,22755 +/- 0,0506648	0,222	0,007	0,005	0,161	**0,16**
20	Predicted sulfur transferase	orig.	LG	7,578	8.67975 +/- 0.196415	0,000	0,029	0.0123108 +/- 0.00169991	0,000	**0,21**
		sub.	LG	7,671	9,63913 +/- 0,187142	0,000	0,032	0,010	0,000	**0,10**
		orig.	CAT	7,578	7.64579 +/- 0.131179	0,300	0,029	0,008	0,000	**0,30**
		orig.	C60	7,578	7.68308 +/- 0.155148	0,246	0,029	0.0116381 +/- 0.00190202	0,000	0,31
		orig.	CGTR	7,578	7.79351 +/- 0.150829	0,072	0,029	0,008	0,000	**0,29**
		sub.	GTR	7,671	7,90378 +/- 0,172341	0,083	0,032	0,009	0,001	0,48
		orig.	CGREC	3,587	3.54402 +/- 0.0703473	0,720	0,029	0,011	0,000	**0,18**
		sub.	CGREC	3,716	3,73035 +/- 0,0727001	0,418	0,026	0,011	0,001	0,67
21	probable S-adenosyl-methyltransferase	orig.	LG	10,462	13.8002 +/- 0.126155	0,000	0,016	0,010	0,025	0,99
		sub.	LG	6,964	9,1378 +/- 0,172376	0,000	0,013	0,008	0,004	**0,10**
		orig.	CAT	10,462	10.4969 +/- 0.109197	0,362	0,016	0,011	0,029	0,84
		orig.	C60	10,462	11.9225 +/- 0.110177	0,000	0,016	0,011	0,030	1,00
		orig.	CGTR	10,462	11.0375 +/- 0.118855	0,000	0,016	0,010	0,026	0,99
		sub.	GTR	6,964	7,17682 +/- 0,148109	0,065	0,013	0,008	0,006	**0,30**
		orig.	CGREC	4,327	4.36912 +/- 0.0591791	0,233	0,015	0,011	0,096	0,76
		sub.	CGREC	3,381	3,41937 +/- 0,0762973	0,292	0,010	0,009	0,376	**0,10**
22	anthranilate phosphoribosyltransferase	orig.	LG	10,362	12.51 +/- 0.122739	0,000	0,023	0.0146145 +/- 0.00228379	0,003	0,98
		orig.	CAT	10,362	10.4453 +/- 0.10388	0,202	0,023	0.0154653 +/- 0.00275574	0,014	0,65
		orig.	C60	10,362	12.0118 +/- 0.11846	0,000	0,023	0.0154814 +/- 0.00273889	0,007	0,90
		orig.	CGTR	10,362	10.9691 +/- 0.218615	0,002	0,023	0.0153722 +/- 0.00273969	0,014	0,81
		orig.	CGREC	4,220	4.5275 +/- 0.112345	0,002	0,018	0.0176487 +/- 0.00320101	0,419	0,70
23	cytosine/adenosine deaminases	orig.	LG	11,872	14.2376 +/- 0.142945	0,000	0,028	0,021	0,030	0,99
		orig.	CAT	11,872	11.7264 +/- 0.140518	0,849	0,028	0,019	0,008	0,99
		orig.	C60	11,872	12.5331 +/- 0.129909	0,000	0,028	0,020	0,011	0,96
		orig.	CGTR	11,872	12.4274 +/- 0.155905	0,000	0,028	0,019	0,005	0,97
		orig.	CGREC	4,482	4.64227 +/- 0.100211	0,046	0,024	0.0246658 +/- 0.00448843	0,466	1,00
24	Na+/H+ antiporter NhaD and related arsenite permeases	orig.	LG	5,090	6.21817 +/- 0.119757	0,000	0,012	0,005	0,000	**0,15**
		sub.	LG	5,796	7,5457 +/- 0,122402	0,000	0,014	0,005	0,000	**0,29**
		orig.	CAT	5,090	5.21768 +/- 0.101213	0,099	0,012	0,005	0,003	**0,17**
		orig.	C60	5,090	5.41314 +/- 0.0950014	0,003	0,012	0,005	0,009	**0,27**
		orig.	CGTR	5,090	5.25238 +/- 0.108057	0,062	0,012	0,005	0,002	**0,18**
		sub.	GTR	5,796	6,04255 +/- 0,0979296	0,002	0,014	0,005	0,001	**0,24**
		orig.	CGREC	2,850	2.84305 +/- 0.0700755	0,525	0,014	0,007	0,013	**0,24**
		sub.	CGREC	2,957	2,9618 +/- 0,0593177	0,461	0,013	0,006	0,002	0,32
25	queuine tRNA-ribosyltransferase	orig.	LG	10,746	14.1897 +/- 0.108826	0,000	0,012	0,007	0,015	1,00
		orig.	CAT	10,746	10.7141 +/- 0.0959455	0,623	0,012	0,007	0,005	0,98
		orig.	C60	10,746	12.0874 +/- 0.0936563	0,000	0,012	0,007	0,008	0,93
		orig.	CGTR	10,746	11.4716 +/- 0.106725	0,000	0,012	0,007	0,008	1,00
		orig.	CGREC	4,336	4.33852 +/- 0.0518978	0,457	0,014	0,008	0,028	0,93
26	rRNA methylases	orig.	LG	11,777	13.9502 +/- 0.140359	0,000	0,029	0.0255107 +/- 0.00234193	0,057	0,41
		orig.	CAT	11,777	11.9357 +/- 0.125824	0,092	0,029	0.0259225 +/- 0.00292805	0,114	0,93
		orig.	C60	11,777	12.7806 +/- 0.118227	0,000	0,029	0.0257562 +/- 0.00264686	0,101	0,98
		orig.	CGTR	11,777	12.5427 +/- 0.205776	0,000	0,029	0.0257833 +/- 0.00301531	0,115	0,74
		orig.	CGREC	4,463	4.83611 +/- 0.129571	0,002	0,036	0.0285356 +/- 0.00568485	0,086	0,60
27	tRNA delta(2)-isopentenylpyrophosphate transferase	orig.	LG	11,885	12.0078 +/- 0.108593	0,127	0,030	0.0227869 +/- 0.00349284	0,035	0,65
		sub.	LG	6,487	7,91882 +/- 0,214337	0,000	0,020	0,014	0,055	**0,10**
		orig.	CAT	11,885	12.061 +/- 0.107364	0,051	0,030	0.0227562 +/- 0.00338936	0,030	0,74
		orig.	C60	11,885	12.9891 +/- 0.109971	0,000	0,030	0.022435 +/- 0.0035495	0,028	1,00
		orig.	CGTR	11,885	12.7709 +/- 0.194575	0,000	0,030	0.0233716 +/- 0.00370962	0,055	0,42
		sub.	GTR	6,487	6,74221 +/- 0,192333	0,088	0,020	0,013	0,056	**0,06**
		orig.	CGREC	4,631	4.8048 +/- 0.0804132	0,013	0,032	0.0206261 +/- 0.00429936	0,016	0,99
		sub.	CGREC	3,406	3,37933 +/- 0,10215	0,589	0,012	0,015	0,760	**0,10**
28	tRNA-dihydrouridine synthase	orig.	LG	7,338	9.51461 +/- 0.146638	0,000	0,011	0,007	0,019	**0,28**
		sub.	LG	6,550	7,71203 +/- 0,181897	0,000	0,008	0,009	0,668	**0,10**
		orig.	CAT	7,338	7.46967 +/- 0.116852	0,121	0,011	0,008	0,029	**0,30**
		orig.	C60	7,338	8.11837 +/- 0.112524	0,000	0,011	0,007	0,032	0,63
		orig.	CGTR	7,338	7.5921 +/- 0.134993	0,026	0,011	0,007	0,016	0,53
		sub.	GTR	6,550	6,72118 +/- 0,161964	0,132	0,008	0,009	0,691	**0,11**
		orig.	CGREC	3,599	3.55163 +/- 0.0655622	0,753	0,027	0,010	0,000	0,90
		sub.	CGREC	3,489	3,46393 +/- 0,0867696	0,593	0,011	0,012	0,543	**0,06**

„I am still confused, but on a higher level.“

Enrico Fermi

Acknowledgements

I want to thank the head of the Department, Michael Wagner, and especially my group leader Matthias Horn for welcoming me. I feel lucky and proud I got the chance to work with my supervisor Daryl Domman whom I truly appreciate very much, especially his enthusiastic and inspiring nature.

Thank you Daryl, for passing the passion for phylogenetics on to me. After achieving a lot of practical experience in the wet lab, I was a bit scared the first moments after switching to computational work. However, thank you for showing me another shade of a biologist's life. In the end, I was happy to end up working in front of the computer for my master's thesis on this exciting topic. At the beginning I was suspicious of bioinformatics but it turned out that an unknown passion has awoken.

Of couse, I also want to thank all DOMiES who made such a familiar working atmosphere possible. Whenever I felt the need for a break I was glad to be able to accompany Florian Wascher in the lab and enjoy some funny conversations. Thank you Florian Höggerl for your support in the wet lab and fruitful discussions about my gels, I was a bit sad you left so early. Thank you Allen, Paul and Daryl for fooling around - you really made me laugh tears sometimes. Also Felicitas', Gabi's and Martina's cheerful tempers always were perfect for putting a smile on my face.

Another award goes to Markus Eich, who bore me whenever I was in a bad or frustrated mood. Thanks for being my personal chauffeur and especially for hittin' the gym with me and doing a lot of relaxing activities.

My dear parents, thank you for your unconditional love and support. Thanks for feeding my brain with delicious food whenever I was too tired, thanks for your motivation and consoling words whenever I was down. Thanks for telling me I should also have breaks, relax. In my whole lifetime I will never be able to compensate what you give to me.

Sabine Felkel, BSc

Curriculum vitae

Personal data

Date of birth	October 12th, 1990 in Vienna
Nationality	Austria
Address	Donaufelderstraße 208/9/10, 1220 Vienna
Contact	sabine.felkel@gmx.at

Education

Since 2014	Diploma thesis at the Department of Micobiology and Ecosystem Science, Divison of Microbial Ecology, University of Vienna
2012 – 2014	Master's degree course in Ecology, University of Vienna
2009 – 2012	Bachelor's degree course in Biology, University of Vienna
2001 – 2009	AHS Franklinstraße 26, 1210 Vienna

Work Experience

2014	Internship at AIT, Tulln

Skills

Wet lab	Especially familiar with PCR, protein and DNA isolation, gel electrophoresis, SDS-PAGE, FISH, cloning, microscopy
Dry lab	MS Office, experience with Linux, ARB, basic knowledge in Bash, Perl, Python and R
Languages	Mother tongue German, fluent English, basic French
Driver's license	Class B
Qualities	Flexible, independent, reliable, on time, ambitious, nosy