Plastid establishment did not require a chlamydial partner

Daryl Domman¹, Matthias Horn¹, T. Martin Embley² & Tom A. Williams²

Primary plastids descend from the cyanobacterial endosymbiont of an ancient eukaryotic host, but the initial selective drivers that stabilized the association between these two cells are still unclear. One hypothesis that has achieved recent prominence suggests that the first role of the cyanobiont was in energy provision for a host cell whose reserves were being depleted by an intracellular chlamydial pathogen. A pivotal claim is that it was chlamydial proteins themselves that converted otherwise unusable cyanobacterial metabolites into host energy stores. We test this hypothesis by investigating the origins of the key enzymes using sophisticated phylogenetics. Here we show a mosaic origin for the relevant pathway combining genes with host, cyanobacterial or bacterial ancestry, but we detect no strong case for Chlamydiae to host transfer under the best-fitting models. Our conclusion is that there is no compelling evidence from gene trees that Chlamydiae played any role in establishing the primary plastid endosymbiosis.
Endosymbiosis is key to the evolutionary success of eukaryotes, from the ancient endosymbiotic origins of mitochondria and chloroplasts to the methanogenic symbionts of anaerobic ciliates and the nutritional symbioses of sap-feeding insects. Cell biology and phylogenetics testify to the prokaryotic origins of these endosymbiotic organelles, but the molecular mechanisms by which their free-living progenitors were originally recruited and integrated with a host cell remain poorly understood. The endosymbiotic capture of a cyanobacterium by a heterotrophic eukaryotic host cell at the origin of the Archaeplastida marked one of the most important events in evolutionary history, for through this symbiosis all plant life would emerge. Other photosynthetic eukaryotes obtained their plastids through secondary endosymbiosis of one of these primary lineages, implying that—with a single exception—all photosynthetic eukaryotes trace the origin of their photosynthetic machinery to the primary cyanobacterial endosymbiosis. However, despite substantial progress on the evolution of plastids and their relationships to free-living cyanobacteria, the initial selective pressure that drove the acquisition and retention of the cyanobacterial endosymbiont remains unclear. Modern plastid and host metabolisms are intimately intertwined, with the chloroplast providing primarily fixed carbon to the host in exchange for a multitude of metabolites, including phosphate derivatives and NAD. However, present-day host–plastid interactions are the product of more than a billion years of co-evolution and the situation may have been very different at the time of the initial endosymbiosis. In addition to the provision of carbohydrates to the host through import by the pathogen through a series of downstream reactions catalysed by the effectors GlgA, GlgP and GlgX, all secreted by the chlamydial effector GlgC; ADP-glucose was subsequently lost. The newly formed relationship between cyanobacterium and host led to the modern plastid and the evolution of new plastidial genomes were transferred, the chlamydial partner was no longer needed and the chlamydial partner was lost. The newly formed relationship between cyanobacterium and host led to the modern plastid and the evolution of modern Archaeplastida, the chlamydial partner might have been lost from the consortium following horizontal transfer (HGT) of the key metabolic genes to the host nucleus. In support of the hypothesis, some modern pathogenic Chlamydia appear to manipulate host metabolism by the secretion of glycolytic enzymes and some of the homologues of these enzymes from environmental Chlamydia were shown to be secreted by the type III secretion system in a Shigella assay. Further, published gene trees for some archaeplastidal enzymes involved in carbohydrate metabolism show the archaeplastidal sequences having been recently transferred to the chlamydial partner. Extant Chlamydiae have probably been associated with eukaryotes for at least 700 million years (17) so it appears reasonable to suggest that they also infected even more ancient eukaryotes. Extant Chlamydiae can infect a tremendously diverse range of eukaryotic hosts such as humans, cattle, pigs, birds, koala, fish, insects and unicellular protists. Notably, Chlamydiae have not been found infecting any member of the Archaeplastida. A proposed evolutionary scenario, coined the ‘ménage à trois’ hypothesis of plastid establishment. Chlamydiae have a broad host range, from humans (where Chlamydia trachomatis is a major cause of sexually transmitted disease) to cattle, fish, isopods and protists. However, extant Chlamydiae are not known to infect any members of the Archaeplastida, although the situation may have been different in the distant past. The ‘smoking gun’ for the mitochondrial and plastid endosymbioses was the detection of an organelle, and although there is currently no evidence for a chlamydia-derived organelle in modern Archaeplastida, the chlamydial partner might have been lost from the consortium following horizontal transfer (HGT) of the key metabolic genes to the host nucleus. In support of the hypothesis, some modern pathogenic Chlamydia appear to manipulate host metabolism by the secretion of glycolytic enzymes and some of the homologues of these enzymes from environmental Chlamydia were shown to be secreted by the type III secretion system in a Shigella assay. Further, published gene trees for some archaeplastidal enzymes involved in carbohydrate metabolism show the archaeplastidal sequences
emerging from within, or clustering with, the *Chlamydiae*. These include the genes encoding the putative effectors GlgX and GlgA mentioned above. These trees are compatible with a chlamydial origin for key components of plant carbohydrate metabolism, providing phylogenetic support for the ménage à trois hypothesis.

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

**Results and Discussion**

Simple methods do not adequately model sequence evolution. Under the ménage à trois hypothesis, archaebplastidal GlgC, GlgA, GlgP and GlgX originated as chlamydial effectors whose coding sequences were later transferred to the host nucleus; as a consequence, their modern-day archaebplastidal homologues are expected to cluster within, or as the sister to, chlamydial genes in sequences were later transferred to the host nucleus; as a consequence, their modern-day archaebplastidal homologues are expected to cluster within, or as the sister to, chlamydial genes in sequences, raising the possibility that these trees are compatible with a chlamydial origin for key components of plant carbohydrate metabolism, providing phylogenetic support for the ménage à trois hypothesis.

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

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

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

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

The enzyme that performs this step is a glycogen phosphorylase, GlgP, which catalyses glycogen breakdown by releasing glycogen chains to within four residues of an α-1,6 branch. Our phylogenetic analyses did not recover a chlamydial, or indeed a cyanobacterial, origin for archaeplastidal glgP under any of the models used. Instead, the archaeplastidial sequences grouped with some other eukaryotes away from both the cyanobacterial and chlamydial clades (Fig. 3b and Supplementary Fig. 2), consistent with vertical descent of GlgP from host glycogen for import into the pathogen.

Under the ménage à trois hypothesis, the original role of the glycogen synthase GlgA is proposed to have been the incorporation of ADP-glucose generated by GlgC into host glycogen for later exploitation by the chlamydial pathogen. This enzyme would therefore have established the initial link between host and cyanobiont metabolism by providing a route for the incorporation of a bacterial metabolite (ADP-glucose) into host energy stores. In agreement with the recent analyses of Ball et al.\(^16\), a phylogeny inferred under the LG model provides moderate support (PP = 0.83) for chlamydial ancestry of the archaeplastidial sequences, although this model was rejected in our posterior predictive simulations both for across-site and across-branch compositional heterogeneity (PP = 0 for across-site compositional heterogeneity, PP = 0.002 for across-branch heterogeneity; see Supplementary Table 1). Indeed, the GlgA alignment proved to be extremely heterogeneous in both across-site and across-branch composition; unusually, even the most general substitution model currently available (CAT + GTR) failed to provide an adequate fit with respect to across-site compositional variation (PP = 0, see Fig. 2e and Supplementary Table 1). Although the tree inferred under CAT + GTR did not fit the data, it did weakly (PP = 0.74) support a Chlamydiaceae plus Archaeplastida clade, consistent with horizontal exchange (Fig. 4a and Supplementary Fig. 5).
The original GlgA alignment of Ball et al.\textsuperscript{16} contains, in addition to the chlamydial and archaeplastidal sequences that are key to the ménage à trois hypothesis, an extensively sampled outgroup that includes distantly related, functionally divergent paralogues of these enzymes from Archaeplastida and bacteria. We reasoned that the large evolutionary distances, functional shifts and associated long branches that characterize this outgroup might be a contributing factor to the failure of the CAT + GTR model to adequately capture the compositional heterogeneity evident in the data set, potentially interfering with the inference of in-group relationships\textsuperscript{35}. To test this idea, we removed most of the outgroup sequences, retaining only the clade that branched closest to the key Chlamydiae/Archaeplastida clade in the initial CAT + GTR analysis, and performed inference on this reduced data set under the same model. The removal of the more distant outgroup sequences significantly reduced the

![Figure 3](https://example.com/figure3.png)

**Figure 3** | Single gene trees for key components of archaeplastidal carbohydrate metabolism implicated in the ménage à trois. (a–d) Phylogenies for GlgC, GlgP, GlgX and UhpC. These trees were inferred under the CAT + GTR model in PhyloBayes, which performed better in our analyses of model fit than the single-matrix models originally used to analyse these genes. With the exception of the Chlamydomonas and Ostreococcus GlgC sequences, the Archaeplastida were recovered as a monophyletic group in all of these trees, suggesting that this pathway was already present in its current form in the last common ancestor of the group. However, the closest outgroup to the Archaeplastida varies among the individual gene trees, as discussed in the main text. We rooted the tree in panel (d) between UhpC and its paralogue GpT. In the other panels, we oriented the trees to most clearly visualize the key relationships between the archaeplastid and chlamydial sequences, and to test the predictions of the ménage à trois hypothesis. Support values are summarized as Bayesian posterior probabilities and branch lengths are proportional to the expected number of substitutions per site, as indicated by the scale bar.
Chlamydiae are summarized as Bayesian posterior probabilities, and branch lengths are proportional to the expected number of substitutions per site. Containing the chlamydial and archaeplastidal sequences; the root positions indicated are based on the topology of the complete analyses. Support values recoded data set under the CAT relationship, recovering an in-group trichotomy between the sequences from Archaeplastida, + CAT across-site and across-branch compositional variation using the evolution of GlgA: Dayhoff recoding and joint modelling of other bacteria. Also evaluated two alternative approaches for modelling the evolution of the GlgA gene is unusually difficult to model, given the high levels of both across-site and across-branch compositional variation observed. Nonetheless, our analyses using a series of better-fitting models suggest that there is no convincing support for a specific Chlamydiae/Archaeplastida relationship.

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

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

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

Finally, we attempted to jointly model the across-branch and across-site compositional variation in the GlgA alignment using the non-stationary CAT + BP model, which combines modelling of across-site composition in the same way as the CAT model with a process in which composition can change at breakpoints (BPs) across the phylogenetic tree, leading to across-branch compositional variation. These analyses also failed to recover a Chlamydiae/Archaeplastida clade (Fig. 4d and Supplementary Fig. 8). Overall, our analyses demonstrate that the evolution of the GlgA gene is unusually difficult to model, given the high levels of both across-site and across-branch compositional variation observed. Nonetheless, our analyses using a series of better-fitting models suggest that there is no convincing support for a specific Chlamydiae/Archaeplastida relationship.
In summary, our phylogenetic analyses suggest a mosaic origin for archaeoplastidal carbohydrate metabolism: the ADP-ribosephorylase GlgC descends from within the cyanobacteria, consistent with an origin from the cyanobacterial endosymbiont; the glycolgen phosphorylase GlgP may descend from the eukaryotic host cell for that endosymbiont, and the glycolgen synthase GlgA, the debranching enzyme GlgX and the hexose phosphate transporter UhpC appear to have bacterial, but not necessarily chlamydial, origins. Thus, in contrast to the predictions of the nénage à trois hypothesis, our analyses suggest that there is no compelling evidence that any of the key genes of archaeoplastidal carbohydrate metabolism were acquired from an ancient chlamydial partner.

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

Methods

Sequences and alignments. The GlgA and GlgX alignments were those used in Ball et al.9 For the other genes, gene families were downloaded from the HOGENOM40 (UhpC and GlgC) or OMA41 (GlgP) databases and augmented with their orthologues from a set of newly sequenced chlamydial genomes (Neochlamydia sp. TUME1 and EPS4, Protoclamydia sp. EI2 and Parachlamydia sp. OEPW1) as well as additional cyanobacterial orthologues. Sequences for GlgC and GlgP were aligned using Muscle 3.8 (ref. 42) and poorly aligning regions were detected and removed using BMGE43 with the BLOSUM30 scoring matrix. UhpC sequences were collected by extracting the top 250 BLAST hits of the Protoclamydia amoebophila homologue (Q6ME88_PARUW) against the UniRef90 database44 and supplemented with the aforesaid chlamydial genomes. The alignment was performed using clustalOmega45 and filtered using GBLOCKS46 by using the parameters ‘-b4 = 3 – b5 = a’. All sequence sets, alignments and Newick tree files have been deposited in FigShare (http://dx.doi.org/10.7684/m9.figshare.1257740).

Phylogenetic analyses. Analyses using the CAT + GTR and CAT + GTR + Dayhoff models were performed using PhyloBayes-MP17 1.5a and analyses using CAT+BP were performed using nhPhyloBayes25. Bayesian analyses using the LG model were performed in PhyloBayes 3.3 (ref. 48). For each analysis, two chains were run in parallel, and the bcpomp and tracccomp programmes were used to assess convergence. We judged that analyses had converged when the maximum discrepancies in bipartition frequencies (bpcomp) and summary statistics (traccomp) between the two chains had all dropped below 0.1, and the effective sample size of each parameter was at least 100, as recommended in the PhyloBayes manual (http://www.phylobyases.org).

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

References
