Université de Lille
Ecole Doctorale Biologie Santé

Université de Liège
Collège Doctoral BBMC - BIM

THÈSE DE DOCTORAT

en vue de l'obtention du grade de Docteur en Biologie des Organismes et des populations de l'Université de Lille
en vue de l'obtention du grade de Docteur en Sciences de l'Université de Liège

Rédigée et présentée par

Marie LELEU

Implication des Chlamydiales dans l'évolution des Archaeplastida

Testing the Chlamydial footprint in the evolution of Archaeplastida

Directeurs de Thèse :
Pr Steven Ball, Pr Denis Baurain et Dr Ugo Cenci

Thèse soutenue le 22 avril 2022

Membres du jurys

Rapporteurs:	Pr John Archibald, Dalhousie University
	DR Gwenael Piganeau, Sorbonne University
Examinateurs:	DR Angela Falciatore, Sorbonne University
	Pr Eva Nowack, Heinrich-Heine-University
	Dr HDR Ugo Cenci, Lille University
	Pr Steven Ball, Lille University
	Pr Denis Baurain, Liège University
Président:	Pr Annick Wilmotte, Liège University

Abstract

Acquisition of oxygenic photosynthesis by eukaryotes occurred through a unique primary endosymbiosis of a cyanobacterium within a heterotrophic ancestor that resulted in the emergence of the three primary lineages known as the Archaeplastida consisting of the Rhodophyceae (red algae), Chloroplastida (green algae) and Glaucophyta (glaucophytes)

Recently, a paradigm shift in the acquisition of photosynthesis was proposed: the implication of an intracellular obligate pathogen in plastid establishment. This hypothesis, dubbed the Ménage-à-trois Hypothesis (MATH), specifically addresses the central issue of disconnected supply and demand of carbon at the time of plastid endosymbiosis, suggesting an active and direct role of Chlamydiales in the success of primary endosymbiosis, which would have provided many critical genes to the cyanobiont hosted in a common vesicle known as the chlamydial inclusion. The expression and efficient localization of specific genes, such as key transporters and glucan transferases, would have initiated the biochemical fluxes of symbiosis. MATH is supported by molecular, biochemical and phylogenetic evidence but remains highly controversial. The major criticism concerns both the interpretation of single gene phylogenetic trees and the existence of other contributions assigned to different groups of bacteria thereby questioning a specific role for Chlamydiales in the endosymbiotic process. This work aims to test the Ménage à Trois Hypothesis by first evaluating the chlamydial footprint in the evolution of Archaeplastida and then comparing this signal to analogous contributions from other lineages of the bacterial domain. A bioinformatic pipeline was designed to identify all lateral gene transfer (LGT) events between Chlamydia and Archaeplastida, for which a manual analysis of the trees confirmed the occurrence very early during the endosymbiotic process. We then compared this chlamydial signal in the Archaeplastida to control signals, to ensure the specificity of both the bacterial donors and the eukaryotic acceptors involved in these LGT.

Résumé

L'acquisition de la photosynthèse oxygénique par les eucaryotes s'est produite par une endosymbiose primaire entre une cyanobactérie et un ancêtre hétérotrophe. Cet événement majeur de l'évolution a abouti à l'émergence des trois lignées primaires connues sous le nom d'Archaeplastida, à savoir les Rhodophyta (algues rouges), les Chloroplastides (algues vertes) et les Glaucophytes (glaucophytes).

Une récente hypothèse remet en question l'acquisition de la photosynthèse chez les eucaryotes en proposant l'implication d'un pathogène obligatoire intracellulaire lors de l'endosymbiose primaire du plaste. Cette hypothèse, appelée l'hypothèse du Ménage-à-trois (MATH), aborde spécifiquement la question centrale de la déconnexion entre l'offre et la demande de carbone au moment de l'endosymbiose du plaste, suggérant un rôle actif et direct des Chlamydiales dans le succès de l'endosymbiose primaire. Ces pathogènes auraient fourni de nombreux gènes critiques au cyanobiont hébergé dans une vésicule commune connue sous le nom d'inclusion chlamydiale. L'expression et la localisation efficace de gènes spécifiques, tels que les transporteurs clés et les glucanes transférases, auraient initié les flux biochimiques de la symbiose. MATH est soutenue par des preuves moléculaires, biochimiques et phylogénétiques mais reste très controversée. La principale critique concerne à la fois l'interprétation des arbres phylogénétiques monogéniques et l'existence d'autres contributions attribuées à différents groupes de bactéries, remettant ainsi en cause un rôle spécifique des Chlamydiales dans le processus endosymbiotique. Ce travail vise à tester l'hypothèse du Ménage à Trois en évaluant d'abord l'empreinte des Chlamydiales dans l'évolution des Archaeplastida et en comparant ensuite ce signal aux contributions analogues d'autres lignées du domaine bactérien. Un pipeline bioinformatique a été conçu pour identifier les transferts latéraux de gènes (LGT) entre Chlamydia et Archaeplastida, pour lesquels une analyse manuelle des arbres a confirmé l'occurrence très tôt au cours du processus endosymbiotique. Nous avons ensuite comparé ce signal Chlamydien chez les Archaeplastida à des signaux contrôles, afin de nous assurer de la spécificité des donneurs bactériens et des accepteurs eucaryotes impliqués dans ces LGT.

Remerciements

Au moment d'écrire ces mots, je me rends compte du chemin que j'ai parcouru au cours de ces trois années de thèse, autant d'un point de vue scientifique que personnel. Et vous y êtes pour beaucoup.

Je voudrais d'abord commencer par remercier les membres de mon jury. Merci de prendre le temps d'évaluer mon travail. J'espère, dans ces quelques pages, vous transmettre un peu de la passion qui m'a animée durant ce projet.
I would like to begin by thanking the members of my jury. Thank you for taking the time to evaluate my work. I hope, in these few pages, to pass on to you some of the passion that animated me during this project.

J'aimerai ensuite remercier sincèrement mes directeurs de thèse, sans qui rien de tout cela n'aurait été possible: Pr Steven Ball, Pr Denis Baurain et Dr Ugo Cenci. Merci infiniment pour m'avoir fait confiance, pour nos échanges et discussions, mais surtout pour votre curiosité et votre passion de la recherche. Chacun à votre manière, vous avez contribué à ce que je suis devenue et m'avez transmis énormément.

Je tiens à exprimer toute ma reconnaissance au Pr Denis Baurain. Nos échanges et vos conseils ont certes contribué à ma réflexion scientifique, mais m'ont aussi conduit à toujours donner le meilleur de moi-même. Merci de m'avoir guidée tout au long de cette thèse, de m'avoir aidée et conseillée, toujours avec bienveillance.

Un merci particulier au Dr Ugo Cenci. Merci de m'avoir donné ma chance et fait découvrir ce monde passionnant qu'est la recherche. Merci pour ton humour, ta patience, pour m'avoir soutenue et éveillée ma curiosité. Merci pour tout ce que tu m'as transmis.

J'aimerai ensuite remercier l'ensemble des membres des équipes de Liège et de Lille. Merci à Christophe, Malika, Matthieu et Léa à Lille et à Valérian et Raph à Liège, pour leur présence, leur bonne humeur. Merci pour ces discussions et ces rires, ces moments qui font la vie. Grâce à vous ce fût toujours un plaisir de venir au labo.

Je n'oublie pas les personnes qui m'ont épaulée dans ce projet de thèse, mais pas seulement. Mes soutiens à toutes épreuves, présents que je le demande ou non. Ceux qui me connaissent mieux que moi-même et qui font rimer fous rires, confidences et complicité: Alice, Amandine et Louis.

Et enfin, je remercie mes parents. Merci pour votre soutien infaillible, pour avoir toujours cru en moi, pour tout l'amour que vous m'avez donné. J'y suis arrivée, c'est grâce à vous.

Table of contents

Liste des tables et figures

List of Table and Figures

List of Appendix

Version en Français

Introduction

1. Endosymbiose primaire du plaste

a. Endosymbioses et évolution

L'acquisition de la photosynthèse, d'abord par les cyanobactéries puis par les eucaryotes, marque un tournant majeur dans l'évolution de la vie sur Terre. La capacité des organismes à utiliser les électrons de l'eau pour conduire les chaînes de transports d'électrons nécessaires à la réduction du CO2 et son incorporation en matière organique grâce à la lumière, entraîne un changement atmosphérique important résultant du dégagement d'oxygène consécutif à la photolyse de l'eau. Cette modification à abouti à une évolution importante du paysage terrestre.

Depuis les travaux de Lynn Margulis (Sagan, 1967), et de ses prédécesseurs, la communauté scientifique s'accorde sur l'origine bactérienne des organites énergétiques des cellules eucaryotes. La théorie endosymbiotique propose en effet l'internalisation de bactéries qui, suite à leur dégénérescence et intégration métabolique, forment la mitochondrie et le plaste. L'endosymbiose est un phénomène évolutif décrit comme l'internalisation d'organismes vivants par d'autres, le tout aboutissant, en fonction du degré d'intégration, à un nouvel organite, dans ce dernier cas on parle d'organellogénèse. Dans les cas d'endosymbioses dites "primaires" (à distinguer des endosymbioses dites "secondaires" ou "tertiaires") une bactérie est internalisée par un autre organisme. La mitochondrie tire son origine d'un événement endosymbiotique produit il y a environ 2 milliards d'années, entre une Alphaproteobacteria et un proto-eucaryote. Le plaste, quant à lui, issu de l'internalisation d'une cyanobactérie ancestrale par un eucaryote unicellulaire hétérotrophe, aurait eu lieu il y a environ 1,6 milliards d'années (Chan et al., 2011; Rodríguez-Ezpeleta et al., 2005; Strassert et al., 2021; Yoon et al., 2004). Chacun de ces événements a marqué l'histoire évolutive terrestre, notamment par l'avènement des eucaryotes, puis par la propagation de la photosynthèse.

b. L'endosymbiose primaire du plaste et les Archaeplastida

Plusieurs événements endosymbiotiques du plaste ont façonné la diversité et l'évolution des eucaryotes photosynthétiques (figure 1) (Archibald, 2009; Cenci et al., 2015). Le plus ancien et le plus important consiste en l'établissement d'une symbiose entre un eucaryote hétérotrophe unicellulaire et une cyanobactérie photosynthétique. La dégénérescence de cette dernière, au cours du processus d'intégration métabolique, conduira à la formation d'un nouvel organite appelé « plaste » (chloroplastes chez les plantes), permettant ainsi la propagation de la photosynthèse au sein des eucaryotes (McFadden, 2014; Reyes-Prieto et al., 2007; Rodríguez-Ezpeleta et al., 2005). Internalisée en tant que proie par

l'eucaryote, le plus probablement par phagocytose, la cyanobactérie ancestrale se trouve alors dans une vacuole en voie d'acidification (Ball et al., 2015). Contrairement aux endosymbioses secondaires où des vestiges demeurent d'une vacuole d'internalisation fusionnée ou non subséquemment au RE (van Dooren et al., 2001), l'intégrité de la membrane de phagocytose n'a pas été conservée lors de l'endosymbiose primaire du plaste. Notons cependant que le feuillet externe de la membrane externe des plastes possède des caractéristiques distinctes du feuillet interne rapprochant ces derniers respectivement des eucaryotes et procaryotes (Joyard et al., 2010). Ces propriétés sont partagées avec la mitochondrie. On ignore tout des modalités d'échappement de ces endosymbiotes relativement à la phagocytose. La symbiose entre les différents organismes n'est possible qu'en établissant des liens communicant entre la vacuole et le cytoplasme de la cellule hôte. Deux cas de figure sont à distinguer ici constitués d'une part par les endosymbioses transitoires très communes chez tous les eucaryotes (*Rhizobium*, *Frankia*, zooxanthelles, endosymbiontes d'insectes…) qui n'auront pas vocation à donner naissance à des organites (Gil and Latorre, 2019; Poole et al., 2018) et d'autre part les endosymbioses primaires et secondaires de plastes et de la mitochondrie de fréquence exceptionnellement basse voire unique (Archibald, 2015). Dans le cas des endosymbioses transitoires, les données les plus récentes suggèrent la mise en place d'une perméabilité membranaire peu sélective par l'interaction des enveloppes avec des peptides antimicrobiens synthétisés par l'hôte (Masson et al., 2016; Mergaert, 2018; Mergaert et al., 2017). Cette perméabilité non sélective laissant passer les petites molécules de manière non spécifique peut bien sûr s'avérer cytotoxique pour le symbionte et pourrait expliquer dans certains cas sa mort cellulaire (par ex. dans le cas de *Rhizobium*) (Kim et al., 2015). Dans une revue récente, mon laboratoire d'accueil propose que les endosymbioses basées sur un métabolisme énergétique membranaire exigeant un contrôle strict de la perméabilité de ces dernières ne peuvent autoriser la mise en place d'une perméabilité non ou peu sélective (under revision). Ces dernières doivent donc mettre en place en connectivité sélective par l'entremise des transporteurs et ce dès les stades les plus précoces de l'endosymbiose. L'étape limitante des endosymbioses génératrices d'organites correspond à la mise en place de moyens de communication sélective entre les différents partenaires en jeu. Ainsi, la cyanobactérie ancestrale, ou cyanobionte, peut profiter des ressources de son hôte, de même que l'eucaryote peut bénéficier des capacités de photosynthèse de la bactérie, le tout dans un environnement d'échanges contrôlés et régulés. La monopolisation et l'adressage de transporteurs sur l' enveloppe du symbionte constitue donc un élément clé de l'intégration métabolique du plaste.

Cet événement majeur de l'évolution, connu sous le nom d'endosymbiose primaire du plaste, a donné naissance au groupe des Archaeplastida (Adl et al., 2005), ou Plantae si on reprend la classification de Thomas Cavalier-Smith (Cavalier-Smith, 1998), parmi lequel nous pouvons distinguer trois lignées différentes : les Glaucophyta , les Rhodophyta (algues

rouges) et les Viridiplantae ou Chloroplastida (plantes et algues vertes) (figure 1). D'un point de vue phylogénétique, les études divergent, mais semblent s'accorder sur la monophylie des Archaeplastida, et ainsi sur l'unicité de l'endosymbiose primaire du plaste chez ces organismes (Irisarri et al., 2022; Moreira et al., 2000; Rodríguez-Ezpeleta et al., 2005). Cependant, l'ordre de branchement des différentes lignées diverge en fonction des études. La plupart placent la séparation des Glaucophytes en premier lieu, suivie par celle des Rhodophyta et des Viridiplantae. Cette topologie ne fait cependant pas encore consensus, puisque certaines publications identifient les Rhodophyta comme étant les premiers à avoir divergé (Hackett et al., 2007; Moreira et al., 2000; Nozaki et al., 2009; Rodríguez-Ezpeleta et al., 2005; Sato, 2019).

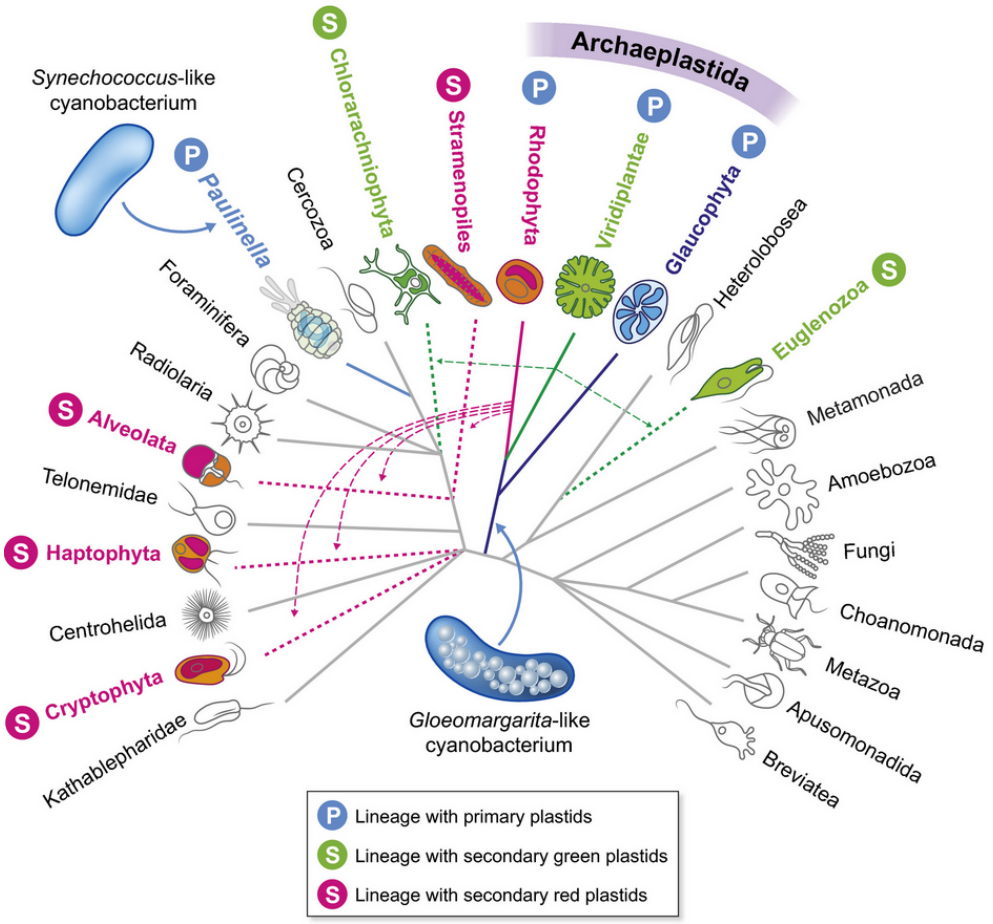


Figure 1 : Distribution de la photosynthèse dans l'évolution des eucaryotes. (repris de Ponce-Toledo et al., 2019). Les lignées photosynthétiques issues d'une endosymbiose primaire sont indiquées par des traits pleins colorés, celles issues d'endosymbioses secondaires par des traits en pointillé. Les deux événements d'endosymbioses primaires connus sont marqués par les flèches bleues, donnant naissance aux Archaeplastida et à *Paulinella*. Les flèches roses et vertes indiquent quant à elle les endosymbioses secondaires de Rhodophyta (rose) ou de Viridiplantae (vert). Les eucaryotes non photosynthétiques sont représentés en gris. (New Phytologist, Volume: 224, Issue: 2, Pages: 618-624, First published: 28 May 2019, DOI: (10.1111/nph.15965))

c. Les autres eucaryotes photosynthétiques

Les Archaeplastida représentent un grand nombre d' eucaryotes photosynthétiques, et sont en tout cas les plus visibles dans le paysage terrestre actuel. Cependant, la dispersion de la photosynthèse ne s'arrête pas à l'endosymbiose primaire du plaste. Une majorité des eucaryotes photosynthétiques, et notamment des algues, sont en effet issus d'endosymbioses secondaires et tertiaires (bien que certaines soient peut-être le résultat de kleptoplastie (Bodył, 2018)) entre des Archaeplastida et d'autres eucaryotes (figure 1) (Archibald, 2009; Keeling, 2010; McFadden, 2001). Ainsi, diverses lignées eucaryotes ont vu le jour suite à l'internalisation d'algues rouges ou vertes par d'autres eucaryotes. Nous pouvons dénombrer au minimum trois évènements endosymbiotiques secondaires principaux, dont l'impact sur la diversité et l'évolution des organismes est considérable. L'internalisation de deux algues vertes est à l'origine de l'apparition des chlorarachniophytes et des euglènes, tandis que l'endosymbiose d'au moins une algue rouge a entraîné l'émergence des haptophytes, cryptomonadales, hétérokontes, dinoflagellés et apicomplexés (figure 1). La situation relative aux endosymbioses rouges semble bien plus complexe, puisque la distribution phylogénétique des organismes possédant des plastes secondaires rouges est peu compatible avec une endosymbiose unique suivie d'une simple évolution verticale. A cela s'ajoute la complexité due à la perte progressive en premier lieu de la photosynthèse, ensuite de l'ADN plastidial et enfin du compartiment plastidial qui s'accompagne de la perte des gènes nucléaires nécessaires au maintien des différents stades. Ces phénomènes ont été particulièrement bien étudiés chez les Apicomplexa et leurs lignées soeurs. Il est très malaisé pour ces raisons de dénombrer le nombre précis d'événements endosymbiotiques en particulier pour les endosymbioses rouges. Contrairement aux Archaeplastida, les plastes des organismes issus d'endosymbioses secondaires ou tertiaires comportent trois à quatre membranes. Ces espèces issues d'endosymbioses secondaires sont considérées comme des algues complexes. Plus récemment, un autre évènement d'endosymbiose primaire menant à l'acquisition de la photosynthèse par les eucaryotes a eu lieu. Totalement indépendant de l'endosymbiose primaire du plaste décrite précédemment, *Paulinella chromatophora* est actuellement le seul représentant de cette association (figure 1). L'intégration métabolique des deux chromatophores n'est à priori pas encore complète. C'est pourquoi l'étude de cet organisme permettrait de mettre en lumière les mécanismes mis en place lors de l'endosymbiose primaire du plaste (Gabr et al., 2020; Marin et al., 2005; Nowack et al., 2008).

d. La cyanobactérie ancestrale

Bien que l'origine endosymbiotique du plaste lors d'un événement unique fasse généralement l'unanimité (Sato, 2021, 2019), l'identification de la cyanobactérie ancestrale à l'origine de l'endosymbiose est encore sujet au débat. Les études divergent en effet, mais

semblent opposer deux origines différentes pour ce cyanobionte : une origine très basale de la cyanobactérie ancestrale, ou une origine parmi des organismes plus récents, similaires aux espèces de la sous section IV (Criscuolo and Gribaldo, 2011; Dagan et al., 2013; Ochoa de Alda et al., 2014; Ponce-Toledo et al., 2017; Turner et al., 1999). Cette dernière hypothèse se base sur les similarités de séquences entre les différentes espèces, procédé débattu en ce qui concerne l'étude de l'endosymbiose primaire du plaste. De plus, le rapprochement de la cyanobactérie ancestrale avec les Nostocales ou les cyanobactéries de la sous section IV peut être dû en partie à la composition en G+C des séquences. En 2011, Criscuolo et Gribaldo investiguent le signal phylogénétique des cyanobactéries au sein des Archaeplastida, et suggèrent une cyanobactérie ancestrale basale. Leurs travaux montrent en effet un branchement phylogénétique des Archaeplastida directement après la divergence de *Gloeobacter violaceus*. Ces résultats sont affinés en 2017, lorsque Ponce-Toledo et al.. placent *Gloeomargarita lithophora* à la base des Archaeplastida pour chacune des conditions et jeux de données testés.

e. Transferts de gènes endosymbiotiques

Tout comme la mitochondrie, le plaste dérive d'une bactérie et possède un génome propre. Comme évoqué précédemment, le succès de l'intégration du plaste repose en grande partie sur la mise en place de liaisons, de points de communication entre les différents partenaires. L'établissement de la connectivité active du plaste a nécessité la mise en place d'un mécanisme d'adressage à l'organite de protéines synthétisées par l'hôte A terme, ces systèmes d'adressage permettent la mise en place des processus plus complexes tels que le transfert de gènes entre les génomes de l'endosymbionte et de l'eucaryote hôte.

Les endosymbioses de manière générale s'accompagnent de pertes massives de gènes (Cavalier-Smith, 2003; McCutcheon and Moran, 2011). En effet l'isolement, à l'intérieur de son hôte, de l'endosymbionte, relativement à la population d'organismes libres d'origine, interdit la réparation des mutations par recombinaison. Ce concept énoncé sous une autre forme par Mueller, et appelé cliquet de Mueller, aboutit à la présence de nombreux pseudogènes inactifs encodant des fonctions désormais inutiles dans l'environnement intracellulaire (Moran, 1996). Le destin de ces pseudogènes sera d'être irrévocablement perdus après la délétion du segment d'ADN correspondant. Ce concept d'isolement de la population naturelle ne s'applique cependant pas aux endosymbiotes transitoires, capables de s'échapper pour recombiner avec leurs partenaires libres tels que *Rhizobium, Frankia*, les zooxanthelles ou *Wolbachia*. Il s'applique toutefois à de nombreux endosymbiotes d'insectes dont le nombre de gènes indispensables à la symbiose se limite au plus à quelques douzaines de fonctions (McCutcheon and Moran, 2011). Ces endosymbiotes, que l'on peut qualifier aussi de transitoires parce qu'ils n'ont pas vocation, tels les mitochondries ou plastes, à demeurer éternellement associés à leurs hôtes font preuve d'un degré élevé d'évolution

convergente dans la structuration de leur génome, aboutissant à de minuscules génomes d'organisation aussi, voire plus, exotique et irrégulière que celles des génomes mitochondriaux. **L'absence de transporteurs dans les membranes de ces endosymbiontes, l'absence de protéome hybride en leur sein, l'absence de système d'adressage, et enfin l'absence, ou la rareté, de transferts horizontaux de gènes de l'endosymbionte au génome nucléaire de l'insecte, démontrent que malgré la remarquable simplification des ces génomes, celle-ci ne corrèle nullement avec et n'est pas à rapprocher de l'émergence d'un véritable organite**.

Les gènes nécessaires à la symbiose chez les plastes et mitochondries se chiffrent, contrairement aux endosymbiontes d'insectes, à plusieurs centaines. Toutefois la nécessaire présence chez les endosymbiontes générateurs d'organites de systèmes d'adressage de protéines à l'organite a permis de partager le codage des fonctions nécessaires à la symbiose avec le génome de l'hôte. Ceci a permis au génome de l'organite de se réduire davantage. L'endosymbiose primaire du plaste s'accompagne ainsi d'une réduction massive du génome du cyanobionte (Cavalier-Smith, 2003). En comparaison aux bactéries actuelles, le génome plastidial représente moins de 5% de la taille du génome cyanobactérien (Green, 2011; Raven and Allen, 2003). Cette réduction peut s'expliquer donc d'une part par la perte irréversible de gènes, due au cliquet de Mueller, et d'autre part par le transfert de gènes au noyau de la cellule hôte. Cette proportion de transferts de gènes endosymbiotiques (EGT) peut s'expliquer notamment par la nécessité pour l'hôte de contrôler l'organite et la disponibilité de ses ressources et de protéger le plus de fonctions indispensables possibles de l'effet néfaste du cliquet de Mueller en permettant la réparation des mutations par les processus de recombinaison de l'hôte impliquant, notamment, son cycle sexué. Cette forme particulière de transfert latéral de gène (LGT), puisque le transfert s'effectue entre organismes de différentes espèces, est caractérisée par un retour des protéines produites vers l'organite, ici donc le plaste. La plupart des EGT participent aux fonctions de photosynthèse et de maintenance spécifique de l'organite (machinerie et métabolisme). Privilégier le transfert de gènes plutôt que la traduction des protéines au sein du plaste résiderait dans la balance de coût énergétique demandé pour les deux phénomènes. En effet, selon Steven Kelly, 2021, d'un point de vue énergétique, il est plus avantageux pour la cellule de transférer le gène dans son propre noyau et ensuite de traduire la protéine dans le cytoplasme, plutôt que de laisser cette fonction au plaste. Chez les plantes vertes, plus de 1000 protéines sont présentes dans le plaste, pour moins d'une centaine de gènes dans le génome plastidial (Vries and Archibald, 2018). Selon les études, entre 600 et 5000 gènes auraient été transférés depuis la cyanobactérie ancestrale jusqu'au génome eucaryote de l'hôte lors de l'endosymbiose primaire du plaste (Martin et al., 2002; Price et al., 2012). Cependant, les cyanobactéries ne sont pas les seules à avoir contribué au génome des eucaryotes photosynthétiques. De manière surprenante, la deuxième source de transferts de gènes chez les Archaeplastida provient des Chlamydia (Price et al., 2012; Qiu et al., 2013; Stephens et al., 1998). Plusieurs études établissent en effet un listing

des gènes chlamydiens retrouvés chez les Archaeplastida, allant jusqu'à remettre en question l'endosymbiose primaire du plaste comme impliquant uniquement l'hôte et la cyanobactérie ancestrale pour préférer une vue centrée sur au minimum trois organismes distincts impliqués dans une symbiose commune (Ball et al., 2015, 2013; Becker et al., 2008; Cenci et al., 2017, 2016; Huang and Gogarten, 2007; Moustafa et al., 2008).

2. Hypothèse du Ménage à Trois

a. Description et évolution des Chlamydia

Une hypothèse récente remet en question notre vision du mécanisme d'acquisition de la photosynthèse chez les eucaryotes, en invoquant la participation d'un troisième partenaire lors de l'endosymbiose primaire du plaste. Cette hypothèse, appelée Hypothèse du Ménage à Trois (MATH), propose l'implication directe d'un pathogène de type Chlamydia (Ball et al., 2015; Facchinelli et al., 2013).

Les Chlamydiae sont des pathogènes intracellulaires obligatoires pour lesquels on distingue 16 familles différentes que l'on peut de manière générale regrouper en deux principales : les Chlamydiaceae et les Chlamydia environnementales. Ces dernières sont en effet principalement isolées depuis des prélèvements environnementaux. Contrairement aux Chlamydiaceae, qui infectent spécifiquement les métazoaires, les Chlamydia environnementales ont un spectre d'infection plus large (Omsland et al., 2014). Les hôtes de ces pathogènes sont assez variables parmi les invertébrés, poissons, mammifères et protistes divers. Leur mode de vie intracellulaire obligatoire entraîne une importante réduction de leur génome (Henrissat et al., 2002) et la mise en place de mécanismes de survie au sein de la cellule hôte. La sécrétion de facteurs de virulence par le système de sécrétion de type 3, associé à la traduction et l'adressage de transporteurs spécifiques, permet à ces pathogènes non seulement d'assurer leur protection face aux défenses de la cellule hôte, mais aussi d'accaparer son métabolisme énergétique, notamment via l'import d'ATP et de glucose-6-phosphate. La sortie des Chlamydia de la cellule hôte entraîne la mort de celle-ci par lyse cellulaire (AbdelRahman and Belland, 2005; Omsland et al., 2014).

La pathogénicité humaine de certains Chlamydia, notamment de *Chlamydia trachomatis*, a accéléré les études sur ces organismes. A la fin des années 1990, le séquençage du génome de *C. trachomatis* est publié (Stephens et al., 1998). Parmi les 894 gènes qui composent son génome, 35 sont recensés d'origine eucaryote. Les premières analyses en déduisent un phénomène de transfert de gènes horizontal entre ces pathogènes et leurs hôtes eucaryotes. Or, les études phylogénétiques montrent une proximité de ces gènes non pas avec les animaux cibles des Chlamydiaceae en général et de *C trachomatis* en particulier mais avec les eucaryotes photosynthétiques (Ball et al., 2013; Becker et al., 2008; Collingro et al., 2011;

Huang and Gogarten, 2007; Moustafa et al., 2008). L'évolution de l'état phagotrophe à phototrophe des eucaryotes, suite à l'endosymbiose primaire du plaste, entraîne un épaississement de la parois cellulaire, les membranes n'étant donc plus exposées à l'infection par des Chlamydia. De fait, aucune infection d'Archaeplastida par des Chlamydia n'a été rapportée à ce jour. De manière surprenante, des gènes chlamydiens ont été retrouvés chez les Viridiplantae d'abord, et ensuite chez les Rhodophyta lorsque le génome de *Cyanidioschyzon merolae* fut disponible. Ainsi, ces LGT entre les pathogènes et les eucaryotes photosynthétiques impactent deux lignées sur les trois qui composent les Archaeplastida. Ceci, combiné à l'incapacité des Chlamydia à infecter des cellules végétales, amène donc à déterminer le timing de ces LGT chez l'ancêtre commun des Archaeplastida.

A la suite de ces premières découvertes, plusieurs études s'attardent sur cette relation entre Chlamydia et Archaeplastida, répertoriant d'abord les transferts de gènes entre les différents groupes. En 2007, Huang and Gogarten sont les premiers à évoquer l'implication potentielle des Chlamydia lors de l'endosymbiose primaire du plaste. L'année suivante, Moustafa et al., et Becker et al., font des propositions similaires sur base d'observations analogues. Bien que les premières études sur le sujet dessinent les contours de l'hypothèse du ménage à trois, c'est l'étude du métabolisme des polysaccharides de réserve chez les Archaeplastida qui laisse entrevoir une contribution chlamydienne au cœur même du processus symbiotique. En effet, une étude préalable de l'évolution du métabolisme des polysaccharides de réserve chez les Archaeplastida a permis de reconstruire l'état probable de ce métabolisme lors de l'initiation du processus endosymbiotique (Deschamps et al., 2008). Cette reconstruction repose sur trois hypothèses : 1°) la monophylie des Archaeplastida, 2°) la compartimentation exclusivement cytosolique de ce métabolisme, 3°) la perte très rapide du métabolisme des polysaccharides de réserve chez le symbiote. L'argumentaire très solide soutenant ces hypothèses peut être retrouvé dans de nombreuses revues (Ball et al., 2013; Baum, 2013; Cenci et al., 2017; Facchinelli et al., 2013; McFadden, 2014). Nous reproduisons sa forme initialement proposée dans le schéma ci-contre (Figure 2).

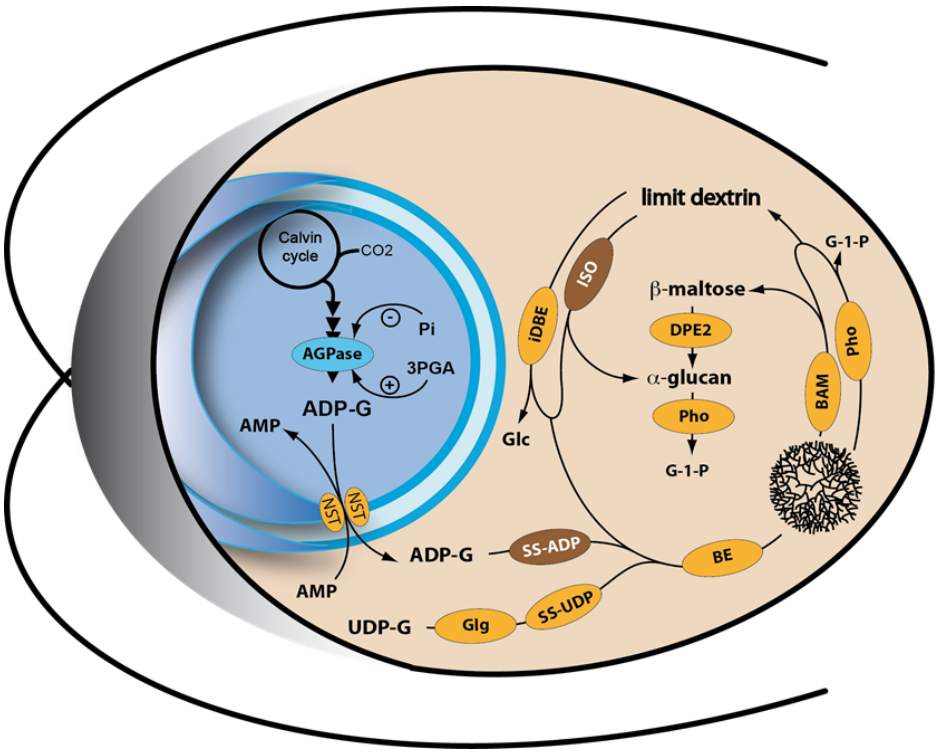


Figure 2 : Flux endosymbiotiques chez l'ancêtre des Archaeplastida. (repris de Deschamps et al., 2008). Dans le cyanobionte, l'ADP-glucose pyrophosphorylase (AGPase) répond à la disponibilité du carbone photosynthétique en synthétisant l'ADP-glucose (ADPG) qui chez les cyanobactéries est normalement destiné à la synthèse de glycogène. Le nucléotide-sucre est transporté par un translocateur nucléotide-sucre/triose-phosphate (NST) qui provient du système endomembranaire de l'hôte, comme cela est proposé dans Weber et al. 2006. L'ADP-glucose est polymérisé en glycogène sans aucune interférence avec les voies de l'hôte. La synthèse fait intervenir une amidon/glycogène synthase soluble (SSADP) nécessitant de l'ADP-glucose. Les chaînes produites sont ramifiées par une enzyme de branchement (BE), puis incorporée dans le glycogène. Indépendamment de la photosynthèse et du cyanobionte, l'hôte est encore capable comme la mjorité des eucaryotes d'alimenter le stockage en glucose grâce à l'utilisation d'une amidon/glycogène synthase soluble nécessitant de l'UDP-glucose, amorcée par la glg (glycogénine). La mobilisation du glucose à partir de l'amidon dépendra entièrement des besoins de l'hôte par le biais d'enzymes hôtes qui comprennent des phosphorylases (Pho), des b-amylases (BAM), et une α-1,4 glucanotransférase spécifique du maltose (DPE2). L'origine phylogénétique de chaque enzyme est représentée soit par une couleur bleue (origine cyanobactérienne avérée), soit par une couleur brune (origine bactérienne) soit encore par une couleur jaune(origine hôte). Le cyanobionte est représenté en bleu. (d'après Mol Biol Evol, Volume 25, Issue 3, March 2008, Pages 536–548, https://doi.org/10.1093/molbev/msm280)

Le flux métabolique débute dans le symbionte par la synthèse d'ADP-glucose, un nucléotide sucre uniquement et exclusivement consacré à la synthèse du glycogène chez les bactéries. Cette enzyme chez les cyanobactéries est très finement couplée au cycle de Calvin et à la photosynthèse par l'entremise des substrats et des effecteurs (le 3-PGA) et inhibiteurs (Pi) de régulation. Elle est aujourd'hui retrouvée dans tous les plastes des algues vertes et plantes terrestres et présente une phylogénie l'enracinant juste à côté de *Gloeomargarita*

lithophora (Deschamps et al., 2008). Son maintien chez le symbionte n'aurait aucun sens si l'ADP-glucose produit n'était pas utilisé. Or, rappelons qu'il était proposé que le symbionte soit devenu incapable d'accumuler du glycogène, comme c'est le cas chez tous les endosymbiontes documentés à ce jour. Pour résoudre ce paradoxe, les auteurs de cette hypothèse ont proposé l'existence d'un transporteur d'ADP-glucose encodé par l'hôte et responsable de l'efflux dans le cytosol de carbone photosynthétique. En effet, à l'époque de cette étude, il avait été montré que tous les transporteurs exportant des sucres variés des plastes au cytosol des algues vertes et rouges, des plantes vertes et des diatomées font partie d'une même famille appelée PPT. Leur origine est en outre indiscutablement monophylétique (Colleoni et al., 2010; Weber et al., 2006). Ils sont dérivés de transporteurs du système endomembranaire des eucaryotes, notamment de transporteurs de GDP-mannose, qui est un analogue structural de l'ADP-glucose. De plus, il avait été aussi montré que certains des ces transporteurs de plantes vertes avaient pour propriété de s'intégrer aux membranes mitochondriales lorsqu'ils sont exprimés chez la levure, sans nécessiter de séquence d'adressage (Loddenkötter et al., 1993). La proposition de la présence d'un transporteur ancien de ce type, bien que non démontrée, était donc plausible, même si très spéculative. Cette spéculation a d'ailleurs été fortifiée par la démonstration *in vitro* de l'efficacité des transporteurs actuels de GDP-mannose pour le transport de l'ADP-glucose. Une fois dans le cytosol de l'hôte, l'utilisation de ce nucléotide sucre était a priori impossible puisque ce métabolite n'existe chez aucun eucaryote. Pourtant, malgré le fait que l'intégralité des gènes impliqués dans le métabolisme eucaryote du glycogène cytosolique soient présents dans la reconstruction du métabolisme des polysaccharides de réserve de l'ancêtre commun des Archaeplastida, celle-ci exigeait aussi la présence de deux gènes bactériens additionnels. Ceux-ci devaient exprimer deux activités dans le compartiment cytosolique de l'hôte qui constitue le seul endroit où se trouvaient les substrats et produits des enzymes correspondantes. Initialement, les gènes encodant ces fonctions étaient pensés d'origine cyanobactérienne, sur base d'analyses phylogénétiques rudimentaires anciennes et d'échantillons de séquences insuffisamment représentatifs. Ces gènes encodaient d'une part une glycogène synthase utilisatrice d'ADP-glucose, et non pas d'UDP-glucose comme chez les eucaryotes, et d'autre part une isoamylase que l'on sait aujourd'hui impliquée dans la transition évolutive du glycogène à l'amidon. Dans leur reconstruction, Deschamps et al., avaient correctement vu que la glycogène synthase, puisqu'elle incorporait l'intégralité du surplus d'ADP-glucose dû à la photosynthèse de l'endosymbionte dans les réserves carbonées de l'hôte, matérialisait la réalité moléculaire du lien symbiotique unissant le symbionte à son hôte. Le flux symbiotique proposé avait pour conséquence inattendue qu'il résolvait l'asynchronie entre l'offre et la demande de carbone photosynthétique au moment de l'endosymbiose, alors qu'il n'existait aucune intégration métabolique et aucune régulation croisée entre les partenaires permettant de résoudre cette asynchronie. En effet, le symbiote pouvait continuer à divertir une fraction du flux photosynthétique vers ses réserves sans

aucun impact négatif sur sa physiologie (comme le font toutes les cyanobactéries libres) et ce, à un moment où l'hôte eucaryote n'en aurait pas un besoin immédiat, tandis que l'hôte eucaryote pourrait mobiliser le surcroît de réserves disponible dans son cytosol en utilisant son propre réseau de mobilisation du glycogène, par exemple à l'obscurité, ou à n'importe quel autre moment où le symbionte ne pourrait lui fournir le carbone nécessaire. Le seul impact négatif sur le partenariat endosymbiotique aurait été la disparition du pool de glycogène dans le compartiment bactérien. Néanmoins, nous savons que des mutants de cyanobactéries libres dépourvus de glycogène sont parfaitement viables en lumière continue mais souffrent de carence en ATP à l'obscurité. Notons qu'il est aujourd'hui établi que les trois lignées issues de l'endosymbiose hébergent, dans l'enveloppe interne de leurs plastes, des protéines importatrices d'ATP de phylogénie indiscutablement chlamydienne.

La proposition de reconstruction décrite ci-dessus avait été initialement bien reçue. Leurs auteurs ont réalisé par la suite que les phylogénies pointant que l'origine cyanobactérienne des enzymes devaient être rediscutées à la lumière de données additionnelles. En effet, une origine cyanobactérienne peut encore aujourd'hui être rejetée pour ces gènes tandis qu'un signal phylogénétique fort les apparente aux Chlamydia, et ce même si un petit groupe de Proteobacteria ne peut, en particulier pour la glycogène synthase être totalement rejetés comme source possible. Les incertitudes de ce type sont néanmoins communes dans les phylogénies simples gènes. Afin d'expliquer la présence d'enzymes chlamydiennes, Ball et al., 2013, ont proposé que ces enzymes étaient des effecteurs sécrétés par le système de sécrétion de type III des Chlamydia dans le cytosol de l'hôte afin de manipuler le flux carboné. Plusieurs groupes de microbiologistes étudiant le cycle d'infection des Chlamydia chez les animaux ont trouvé cette proposition séduisante et prouvé *in vitro* et *in vivo* que toutes les enzymes du métabolisme du glycogène des Chlamydia sont, en effet, des effecteurs métaboliques de leur cycle d'infection (Ball et al., 2013; Gehre et al., n.d.; Lu et al., 2013). Une telle hypothèse d'implication lors de l'endosymbiose, si vérifiée, aurait pour conséquence de rendre les trois génomes interdépendants dans l'établissement des flux symbiotiques. Cette hypothèse très rapidement appelée « ménage à trois » (MATH, Ménage À Trois Hypothesis) explique de manière détaillée les raisons de la conservation longue d'un pathogène/symbionte chlamydien dans le cytosol de l'ancêtre commun des Archaeplastida et l'importance du signal phylogénétique qui en découle dans les génomes de ses descendants actuels.

b. Mécanisme d'action impliqué dans MATH

Historiquement, il existe deux versions différentes de l'hypothèse du ménage à trois. La première propose l'infection préalable de l'eucaryote hétérotrophe par le Chlamydia, puis la phagocytose de la cyanobactérie (Facchinelli et al., 2013). L'échappement immédiat du symbionte de la vacuole est alors nécessaire pour initier l'endosymbiose. La deuxième version quant à elle implique une internalisation simultanée des deux bactéries qui se retrouveraient alors protégées dans la vésicule d'inclusion chlamydienne (Facchinelli et al., 2013). Dans cette deuxième version, la proximité entre les partenaires bactériens entraînerait une facilitation des phénomènes de conjugaison et de transfert de gènes, ce qui permet de repousser l'échappement du futur plaste à un avenir plus lointain. Au fur et à mesure des analyses et des disponibilités des données de séquençage, la deuxième version de l'hypothèse a été privilégiée et sera détaillée dans la suite de ce manuscrit. La raison de ce choix réside dans l'interprétation faite par Gehre et al. d'un ensemble de résultats expérimentaux portant sur l'infection de cellules humaines ou animales par les Chlamydiaceae. Leurs approches étaient basées sur la première utilisation rapportée de mutants de bactéries intracellulaires obligatoires pour les Chlamydia d'une part et sur l'examen des conséquences de constructions de "gene silencing" pour la cellule hôte d'autre part. Ces expériences visaient à comprendre les mécanismes d'accumulation du glycogène dans la vésicule d'inclusion du pathogène. Cette étude a abouti à proposer un modèle détaillé sur le fonctionnement des flux métaboliques actuels impliquant les polysaccharides de réserve et les effecteurs chlamydiens représentés dans la Figure 3. L'étude de *Chlamydia trachomatis* et *Chlamydia muridarum* avait été privilégiée parce que ces pathogènes sont les seuls où l'accumulation de glycogène dans la vésicule atteint des niveaux suffisants pour leur quantification par des colorations cytologiques. Les Chlamydiales « environnementales » se distinguent des Chlamydiaceae sur le plan du métabolisme du glycogène par deux aspects. Le premier réside en la nature effectrice de l'intégralité des enzymes impliquées, y compris l'ADP-glucose pyrophosphorylase, alors que cette dernière n'est pas effectrice chez les Chlamydiaceae. Le deuxième est constitué par une préférence marquée de la glycogène synthase des Chlamydiaceae pour l'UDP-glucose alors que les enzymes correspondantes des autres Chlamydiales sont incapables d'utiliser ce substrat et font preuve d'une sélectivité absolue pour l'ADP-glucose.

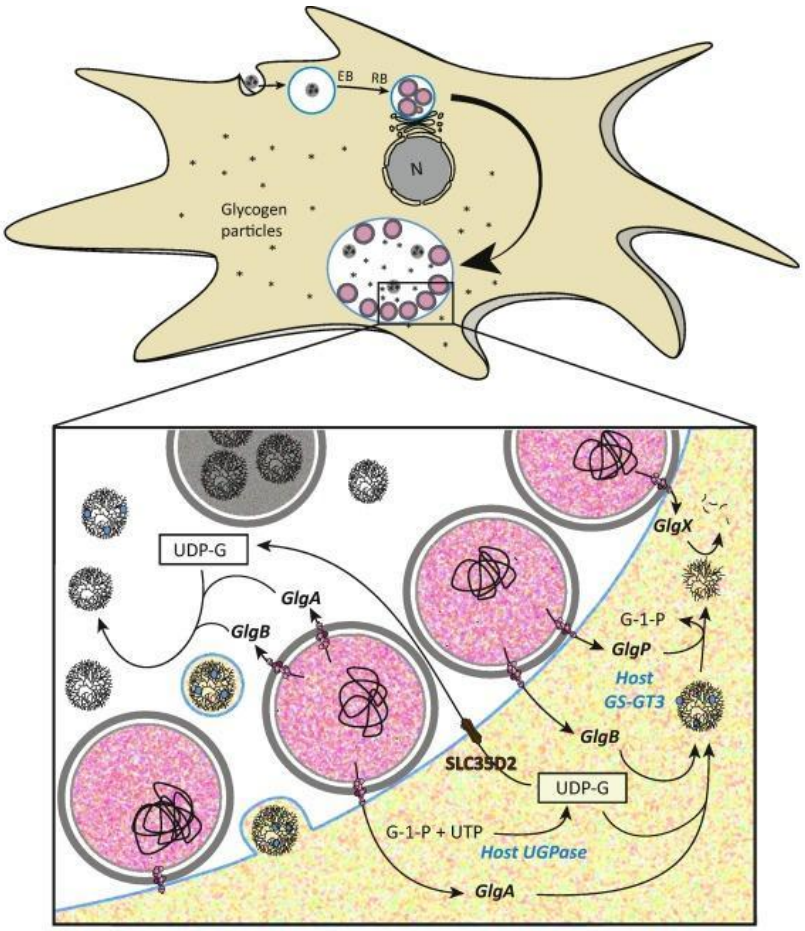


Figure 3: Métabolisme du glycogène chez *Chlamydia trachomatis*. (repris de Cenci et al., 2017) Les corps réticulés (RB) de *C. trachomatis* sont présentés en rose. Ces bactéries à réplication active sécrètent, en tant que protéines effectrices, les enzymes du métabolisme du glycogène bactérien dans la vésicule d'inclusion et le cytosol par l'intermédiaire du T3SS (en rose sur les enveloppes bactériennes faisant face à l'inclusion (en blanc) ou au cytosol (en beige)). GlgA (glycogène synthase) allonge les chaînes de glucose par des liaisons α-1,4 à partir de sucre-nucléotides activés (UDP-Glc ou ADP-Glc), et GlgB (enzyme de ramification) introduit les branches α-1,6 dans le glycogène. GlgC (ADP-glucose pyrophosphorylase), l'enzyme responsable de la synthèse de l'ADP-Glc, substrat spécifique des bactéries, n'est pas sécrétée par les pathogènes et reste à l'intérieur des bactéries. La synthèse du glycogène dans l'inclusion se fait par deux voies. Une voie mineure consiste en l'importation par bourgeonnement de vésicules de glycogène hôte à partir du cytosol (représenté par des particules noires avec la glycogène synthase hôte liée (GS-GT3) représentée par des cercles bleus). La principale voie de synthèse du glycogène dans l'inclusion dépend des effecteurs chlamydiens. La glycogène synthase chlamydiale est capable d'utiliser l'UDP-Glc qui est importé dans l'inclusion par SLC35D2, un transporteur humain d'UDP-Glc recruté à la membrane de l'inclusion. Dans le cytosol, le métabolisme du glycogène implique à la fois l'hôte et les enzymes effectrices chlamydiennes. L'UDP-glucose pyrophosphorylase cytosolique de l'hôte définit la seule enzyme responsable de la synthèse de l'UDP-Glc utilisée à la fois dans l'inclusion et dans le cytosol. On pense que la GlgC (ADP-glucose pyrophosphorylase) bactérienne dirige la synthèse d'une partie du glycogène dans les corps élémentaires (EBs, représentés en gris) lorsque les RBs se différencient à nouveau en EBs. Les effecteurs chlamydiens de la dégradation du glycogène (GlgP (glycogène phosphorylase) et GlgX (enzyme de débranchement du glycogène)) sont également sécrétés dans le cytosol et la vésicule d'inclusion. GlgP dégrade les chaînes externes des grains de glycogène tandis que GlgX débranche la particule restante pour permettre une digestion supplémentaire par GlgP. Ce processus produit de courts malto-oligosaccharides (représentés par de petites lignes noires) qui ne peuvent être utilisés que par les pathogènes. Contrairement à la plupart des glycogènes synthases procaryotes, l'enzyme GlgA de *C. trachomatis* utilise efficacement l'UDP-Glc et l'ADP-Glc. (Cenci et al., 2017)

En extrapolant les résultats expérimentaux obtenus chez les Chlamydiaceae aux autres Chlamydiales, (Cenci et al., 2017) ont proposé pour l'ensemble des Chlamydiales les flux métaboliques détaillés dans la Figure 4 (panneau A). Notons que chez les Chlamydiales le translocateur d'UDP-glucose aurait nécessairement été substitué par un transporteur d'ADP-glucose dont nous avons déjà discuté l'origine. Il devient dès lors aisé d'imaginer l'internalisation simultanée d'une cyanobactérie et d'un pathogène (Figure 4 panneau B). A la manière des interactions biotiques caractérisant l'infection des cellules de plante par *Agrobacterium tumefasciens*, il est possible de proposer que les Chlamydiales aient évolué des mécanismes de manipulation des flux métaboliques des cyanobactéries à leur avantage. Sur le plan du métabolisme carboné, la transmission par conjugaison des gènes encodant des transporteurs clés aurait permis l'efflux de carbone photosynthétique tout en palliant les désordres occasionnés par cet efflux. De tels transporteurs sont constitués par UhpC, une protéine chlamydienne responsable de l'efflux de carbone sous forme de Glucose-6-P encore présente aujourd'hui chez les plastes de glaucophytes où elle est seule responsable possible de l'efflux de carbone photosynthétique, et NTT, un importateur d'ATP présent chez tous les plastes primaires qui aurait permis à la cyanobactérie de survivre à la carence en ATP provoquée à l'obscurité par cet efflux. Cette exportation de glucose-6-P aurait permis d'alimenter la synthèse de glycogène dans la lumière de l'inclusion chlamydienne au bénéfice du pathogène. Ce ne serait qu'en cas d'excès de synthèse d'ADP-glucose dans l'inclusion que ce dernier serait exporté de la vésicule au bénéfice de l'hôte, grâce à l'inversion du flux d'importation d'ADP-glucose sur la membrane de l'inclusion (Figure 4 panneau B). La dissection phylogénétique du métabolisme du tryptophane par Cenci et al., 2016, a mis en lumière une contribution importante des Chlamydiales à ce métabolisme impactant 3/4 enzymes distinctes sur les 7 qui sont responsables de la synthèse de cet acide aminé. Le rôle du tryptophane semble central pour la réplication des Chlamydia. En effet, deux mécanismes de défense antichlamydienne ont été documentés chez les eucaryotes, l'un dans les cellules de souris et l'autre dans les cellules humaines. Ces deux mécanismes totalement différents aboutissent néanmoins au même effet final : la mise en carence sélective pour le tryptophane de la cellule infectée, signant par là une compétition féroce entre l'hôte et le pathogène pour cet acide aminé. Sur base de ces observations, Cenci et al. (2016) ont proposé que les gènes chlamydiens nécessaires à la biosynthèse du tryptophane retrouvés aujourd'hui soient les vestiges d'une interaction ancienne consistant en la transfection d'un opéron chlamydien entier chez la cyanobactérie et engendrant une surproduction de tryptophane. Cette proposition est singulièrement renforcée par la présence dans la membrane interne des plastes primaires d'un transporteur de tryptophane d'origine chlamydienne. Un tel transporteur aurait eu pour effet de permettre l'efflux du tryptophane dans l'inclusion chlamydienne, expliquant la finalité du transfert de gènes impliqués dans le métabolisme du tryptophane. Sachant que seule une demi-douzaine au plus de transporteurs chlamydiens ont été dénombrés sur la membrane interne du plaste, il est difficile d'accepter que la présence de ce transporteur soit

dûe à une simple coïncidence plutôt qu'à un vestige d'une interaction biotique ancienne. L'importance des transferts conjugatifs de gènes dans les interactions biotiques au cœur des hypothèses MATH ont conduit les scientifiques proposant cette hypothèse à préférer l'arrivée simultanée des partenaires endosymbiotiques dans une inclusion commune, facilitant par là les transferts conjugatifs. Selon les auteurs, cette interaction aurait été du même type que celle induite aujourd'hui par *Agrobacterium*, mais se différencie par le fait qu'elle implique une bactérie pathogène et une cyanobactérie dans un environnement intracellulaire, et non une bactérie pathogène et des cellules de plante dans un environnement extracellulaire. Cette interaction endosymbiotique transitoire ne signe pas, toujours selon ces auteurs, le début de l'endosymbiose primaire du plaste, mais elle en aurait assuré le succès en portant la cyanobactérie à un niveau de préadaptation sans précédent. En effet, cette bactérie autotrophe, à l'origine libre et autonome et donc peu pourvue en transporteurs, est désormais enrichie par un arsenal chlamydien de protéines sur son enveloppe assurant sa connectivité dans l'environnement intracellulaire. Ce sera l'échappement accidentel du cyanobionte de la vésicule chlamydienne par hemifusion entre la membrane de l'inclusion et la membrane externe de la cyanobactérie qui signe, selon les auteurs de MATH, le début de l'endosymbiose plastidiale proprement dite. La connectivité entre les différents partenaires et leurs différents métabolismes définissent donc un aspect primordial de la mise en place de l'endosymbiose.

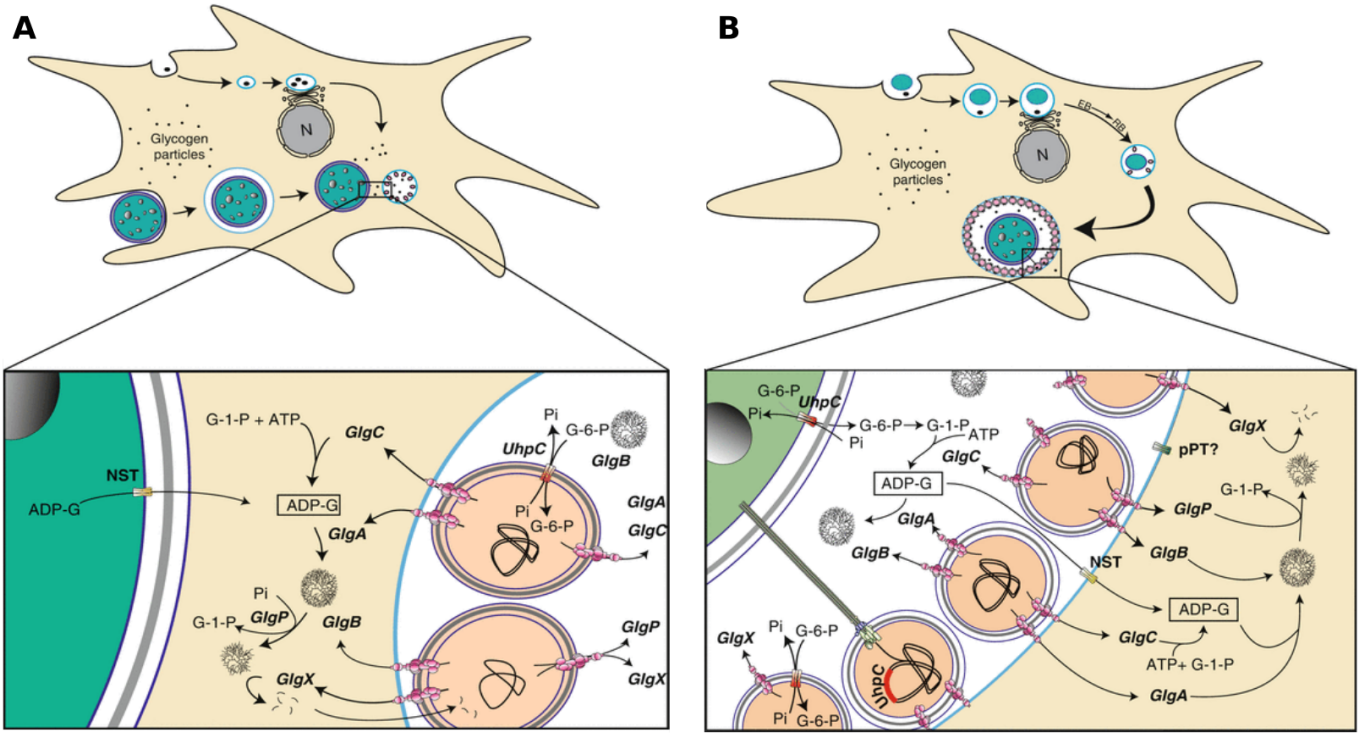


Figure 4: Deux scénarios alternatifs expliquant l'hypothèse du ménage à trois et leur implication métabolique. (repris de (Ball et al., 2015). (A) L'hypothèse classique du ménage à trois (premier scénario NST). Le panneau (A) récapitule le scénario précédemment détaillé dans la figure 3. Le cyanobionte de grande taille (en bleu-vert avec les granules d'amidon affichés) est représenté entrant dans l'hôte indépendamment du

pathogène chlamydien (à l'échelle), un simple point noir épais. Une section montrant l'interaction tripartite est agrandie et encadrée. Les corps réticulés des chlamydia sont représentés attachés par leur TTS (système de sécrétion de type 3, représenté en rose) à la membrane de la vésicule d'inclusion (en bleu clair). Les particules de glycogène sont représentées en noir à l'intérieur de la vésicule d'inclusion et dans le cytosol de l'hôte. Seules les enzymes de provenance chlamydiale sont affichées et abrégées par leurs symboles génétiques : GlgA glycogène synthase, GlgB enzyme de ramification, GlgC ADP-glucose pyrophosphorylase, GlgP glycogène (maltodextrine) phosphorylase , et GlgX GlgX type de DBE directe. Le transporteur NST (ADP-glucose transporter) est mis en évidence sur la membrane du cyanobionte qui a été atteint indépendamment d'un mécanisme d'importation de protéines TIC-TOC inexistant. (B) L'hypothèse du ménage à trois modifié (premier scénario de l'UhpC). Le cyanobionte est représenté entrant avec un corps élémentaire chlamydial. Le corps élémentaire se différencie en corps réticulés attachés par leur TTS à la vacuole phagocytaire modifiée, empêchant ainsi la phagocytose du cyanobionte. Le T4SS (système de sécrétion de type quatre affiché en vert, bleu et noir) responsable du transfert conjugatif de l'ADN transfère ad minima les gènes UhpC et NTT (protéine d'importation de l'ATP) de la chlamydia pour qu'ils soient exprimés dans le cyanobionte. UhpC est représenté en rouge sur la membrane interne du cyanobiont, tandis que NST est représenté en jaune sur la vésicule d'inclusion. Le transporteur d'ADP-glucose de faible affinité (NST) transporte le trop-plein d'assimilation du carbone vers le cytosol où il sera métabolisé.

c. La synthèse de la ménaquinone chez les Archaeplastida

Selon les études, et la stringence des méthodes appliquées, l'estimation du nombre de gènes chlamydiens chez les Archaeplastida varie entre 30 et 100 (Ball et al., 2013; Becker et al., 2008; Collingro et al., 2011; Huang and Gogarten, 2007; Moustafa et al., 2008). Ces transferts sont répartis chez les eucaryotes photosynthétiques, impactant jusqu'aux trois lignées issues de l'endosymbiose du plaste. Ces gènes ne semblent pas être distribués aléatoirement, mais plutôt restreints à certaines voies métaboliques. La première voie métabolique ayant été identifiée comme impactée serait celle du glycogène décrite précédemment. Ces premières études, de même que l'analyse précitée du métabolisme du tryptophane, semblent privilégier la piste de la présence d'interactions biotiques sophistiquées par l'entremise de transferts conjugatifs de gènes encodant des voies métabolique entières et des transporteurs membranaires associés, la connectivité métabolique avec l'hôte étant assurée par la sécrétion d'effecteurs métaboliques (Cenci et al., 2017). Si de tels transferts ont existé, il est logique d'espérer en trouver des traces dans les génomes des plastes actuels. De récentes études montrent en effet l'impact des transferts conjugatifs de gènes chlamydiens vers les plastes sur les voies de synthèse de la ménaquinone (vitamine K) (Cenci et al., 2018).

Les vitamines K constituent un groupe de vitamines liposolubles composées d'un noyau naphthoquinone attaché à une chaîne poly-isoprenyl de longueur et saturation variables. Chez les vertébrés, elles ont des propriétés chélatrices et interviennent dans plusieurs processus moléculaires, dont notamment la coagulation sanguine, le métabolisme osseux et la signalisation cellulaire (Vos et al., 2012). Chez les organismes photosynthétiques, ces vitamines sont associées au photosystème I, participant au transport d'électrons (Reumann, 2013). Deux formes sont principalement retrouvées : la vitamine K1, aussi appelée phylloquinone, présente en majorité chez les plantes (Oostende et al., 2008), algues vertes (Lefebvre-Legendre et al., 2007) et cyanobactéries (Collins and Jones, 1981), et la vitamine

K2 ou ménaquinone, chez les archées, bactéries (Collins and Jones, 1981), Rhodophyta (Yoshida et al., 2003) et diatomées (Ikeda et al., 2008). Contrairement aux bactéries et archées, les animaux et la plupart des protistes ne possèdent pas la voie de synthèse de ces vitamines, pourtant nécessaires à la survie cellulaire, et doivent donc avoir un apport extérieur (Li, 2016). La principale différence entre les différentes molécules réside dans la saturation de la chaîne poly-isoprényl (figure 5A) ; la présence de doubles liaisons impacte en effet leurs propriétés de diffusion et donc leurs capacités biochimiques. Ces quinones isoprénoides polyinsaturées membranaires suivent une réduction réversible en deux étapes, qui leur confère une capacité de conduction d'électrons entre différents complexes protéiques, comme ceux impliqués dans la photosynthèse ou la respiration (Nowicka and Kruk, 2010). En conditions aérobies, l'ubiquinone et la plastoquinone sont les molécules prédominantes. Cependant, la ménaquinone (mais pas la phylloquinone) prend le relais lors de la présence de faibles concentrations d'oxygène, autant pour la respiration anaérobie chez les bactéries que pour la phosphorylation oxydative sous conditions microaérophiliques chez les bactéries et métazoaires (Li, 2016; Sharma et al., 2012). La réduction de la chaîne isoprénoïde insaturée de la ménaquinone permet de former la phylloquinone, dont la chaîne latérale est partiellement saturée.

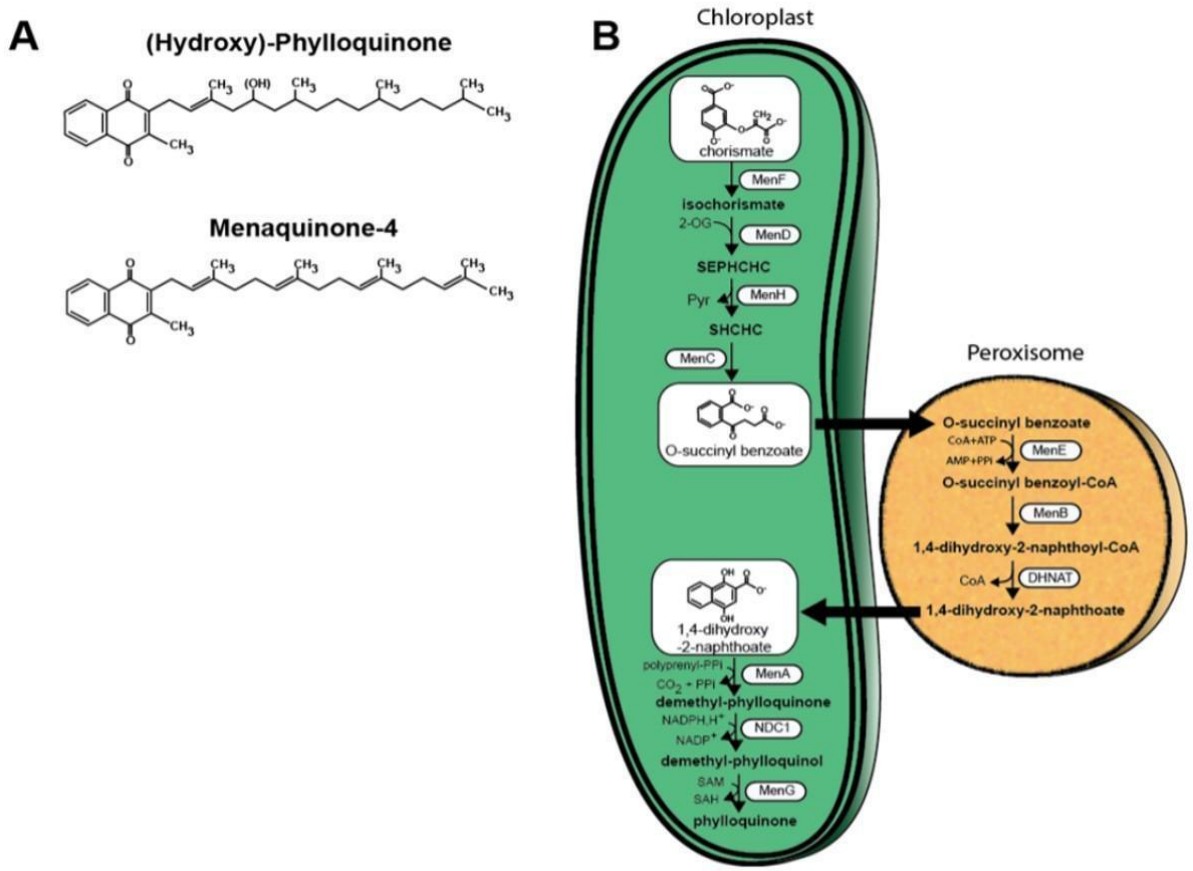


Figure 5 : Structure et synthèse de la Vitamine K chez les Viridiplantae (Adapté de Cenci et al., 2018). A, Structure de la Vitamine K1 (phylloquinone) et Vitamine K2 (ménaquinone). B, Synthèse de la phylloquinone chez les plantes. Le chorismate est converti en quatre étape dans le chloroplaste en O-succinyl benzoate (OSB), par une protéine appelée « Phyllo » représentant la fusion des produits protéiques des gènes bactériens MenF,

MenD, MenC et MenH de la voie de synthèse de la ménaquinone. OSB est alors diffusé au peroxysomes où il est converti en DHNA (1,4-dihydroxy-2-naphtoate) en trois étapes catalysées par MenE, MenB et DHNAT, puis retourne vers le plaste. La chaîne isoprénoïde polyprénylée est ajoutée au DHNA par MenA, puis le noyau DHNA est réduit par Ndc1 (NADPH déshydrogénase) et méthylée par MenG.

Deux voies métaboliques permettent la conversion du chorismate en ménaquinone ou phylloquinone chez les Bactéries et les Archées (voies Men et Futalosine) (Zhi et al., 2014). Les cyanobactéries et les plantes ne comportent quant à elles que la voie Men classique (figure 5B). Chez les plantes, les quatre premières étapes de cette voie de synthèse sont réalisées dans le plaste par la protéine « Phyllo » correspondant à la fusion des quatre enzymes nécessaires à la conversion du chorismate en o-succinyl benzoate, qui est alors acheminé dans le peroxysome pour former le noyau naphthoquinone, avant de revenir dans le plaste où il est prénylé par MenA, réduit par Ndc1 et méthylé par MenG avant d'être associé au PSI (Eugeni Piller et al., 2011; Reumann, 2013) (Figure 5B).

D'un point de vue phylogénétique, l'origine de la voie Men reste floue. Cette voie métabolique étant présente chez les bactéries, mais absente chez la plupart des eucaryotes, il se peut qu'elle ait été transmise à l'ancêtre commun des Archaeplastida lors de l'endosymbiose primaire du plaste. Une première analyse datant de 2008 évoque la possibilité d'une origine non-cyanobactérienne de la majorité des enzymes impliquées (Gross et al., 2008). Une récente étude publiée fin 2018 reprend le sujet en utilisant les nouvelles données génomiques disponibles, et montre que l'ensemble des 7 gènes constituant le cluster responsable de la synthèse de la ménaquinone, encodé au sein du génome plastidial, notamment chez Cyanidiophytina, est d'origine chlamydienne (Cenci et al., 2018). Cette observation est soutenue par des phylogénies simples gènes et la concaténation de MenF et MenD, et vient corroborer l'hypothèse de l'implication du pathogène lors de l'endosymbiose primaire du plaste. Il est aussi important de noter que MenF et MenD pourraient être affiliées à la protéine phyllo retrouvée chez les algues rouges et vertes dans le génome nucléaire de l'hôte. Toutefois l'argumentation phylogénétique soutenant l'affiliation est complexe et mériterait d'être revue à la lumière des modalités différentes d'évolution des génomes nucléaires et plastidiaux. Cette analyse est d'autant plus importante que si l'affiliation venait à être confirmée, celle-ci placerait le transfert du cluster Men avant la diversification des algues rouges et vertes. Au vu de l'absence de la voie Men chez les eucaryotes autres que les Archaeplastida et leurs dérivés, et de la rareté des LGT vers le génome plastidial, il est plausible que le cluster entier responsable de la synthèse de la ménaquinone ait été transmis en une seule fois depuis les Chlamydia jusqu'à l'ancêtre du plaste. Cette observation revêt une importance considérable dans la mesure où elle suggère la réalité des transferts conjugatifs requis dans les interactions biotiques au cœur de MATH. Il était d'autant plus inespéré d'observer de telles reliques que la plupart des plastes ont diminué la taille de leur

génome jusqu'à ne laisser que les gènes responsables du codage de composants des chaînes de transport d'électrons des photosystèmes et des protéines nécessaires à leur traduction, qui ne sont justement pas concernés par les transferts conjugatifs et les interactions biotiques de MATH. Selon les auteurs de MATH, ce ne serait pas un hasard s'ils ont pu être observés chez les Cyanidiophytina puisque les génomes plastidiaux de ces organismes sont les plus gros chez les Archaeplastida, ayant conservé le plus grand nombre de fonctions cyanobactériennes d'origine non comprise dans les photosystèmes ou l'appareil de traduction.
De manière intéressante, cette étude montre aussi l'absence d'une partie de la voie Men, en amont de DHNAT, chez les Glaucophytes, un des trois groupes issus de l'endosymbiose primaire du plaste, et *Galdieria sulphuraria* (Cyanidiophytina), alors que des analyses biochimiques réalisées chez *Cyanophora paradoxa* attestent de la présence d'hydroxyphylloquinone et donc de la capacité de ces organismes à le synthétiser. Certaines cyanobactéries présentent les mêmes caractéristiques : les Gloeobactérales et *Gloeomargarita lithophora* ne possèdent ni la voie Men ni la voie Futalosine, mais synthétisent tout de même de la vitamine K.

Cette similarité biochimique amène à se questionner sur une possible origine commune de la voie de synthèse de la ménaquinone chez ces organismes. Identifier cette voie alternative d'abord chez les Gloeobactérales et *G. lithophora*, puis chez les Glaucophytes, permettrait, si celles-ci sont identiques, d'émettre l'hypothèse de la sélection de cette voie métabolique lors de la séparation des Archaeplastida, et donc de préciser la nature de la cyanobactérie à l'origine du plaste.

3. Une Hypothèse controversée

a. MATH, une hypothèse soutenue et controversée

L'hypothèse MAT est soutenue par des analyses moléculaires, biochimiques et phylogénétiques. L'implication des Chlamydia lors de l'endosymbiose primaire du plaste permettrait donc de dessiner des mécanismes sous-jacents à l'endosymbiose jusqu'alors inconnus. D'un point de vue fonctionnel, la présence d'une Chlamydia dans la vésicule d'inclusion se justifie de par la protection de la cyanobactérie mais aussi de par le transfert de gènes clés permettant la mise en place des flux biochimiques de l'endosymbiose. La Chlamydia agirait donc comme un connecteur entre les deux autres partenaires, permettant l'export des ressources photosynthétiques de la cyanobactérie à la cellule hôte sans pour autant qu'elle n'en pâtisse. L'implication d'un troisième partenaire permettrait aussi d'expliquer la rareté de l'événement au cours de l'évolution.
Cependant, cette hypothèse est controversée au sein de la communauté scientifique. Certaines études remettent en question le timing et la pertinence des LGT, alors que d'autres

questionnent l'interprétation des phylogénies simples gènes, ou l'impact réel des Chlamydia en comparaison de celui d'autres groupes bactériens (Dagan et al., 2013; Deschamps, 2014; Domman et al., 2015).

b. Critiques et aspects méthodologiques

De par l'ancienneté du signal plastidial, que ce soit d'un point de vue purement cyanobactérien ou pris dans un contexte MATH, le signal phylogénétique visualisable dans les arbres simples gènes est considérablement affaibli et sujet à d'important artéfacts phylogénétiques. La difficulté d'identification des EGT, due à la perte de signal mais aussi à l'accélération de l'évolution de ces gènes à la suite de leur adaptation à un nouvel environnement, entraîne la prise de précaution maximale tant qu'à l'interprétation des phylogénies générées. De plus, l'identification de ces EGT est méthode-dépendante. La reconstruction phylogénétique réalisée par maximum de vraisemblance semble surestimer la proportion de gènes transférés par comparaison aux méthodes bayésiennes, à priori moins sensibles aux artéfacts phylogénétiques. En ré-analysant manuellement les transferts de gènes identifiés par Becker et al., (2008), Moreira and Deschamps, (2014), confirment en effet seulement 17 des 55 phylogénies initiales suggérant un transfert de gènes chlamydiens vers les Archaeplastida.

Le recours à des outils de sélection visant à automatiser les protocoles et à diminuer les biais d'interprétation, et ainsi à quantifier le signal endosymbiotique dans sa globalité, comporte cependant quelques biais en lui-même. En effet, ces outils analysent les bipartitions des arbres et sélectionnent ceux présentant un branchement d'intérêt, dans notre cas par exemple un branchement entre les Chlamydia et les Archaeplastida. Plusieurs aspects de l'analyse sont donc négligés par ces approches, tels que la prise en compte de la topologie complète de l'arbre ou la diversité taxonomique. Une approche de sélection manuelle des arbres présente elle aussi certains inconvénients et certains biais, notamment au vu de la variabilité de l'interprétation individuelle, mais permet de confirmer un transfert de gène d'origine endosymbiotique en prenant en compte l'arbre dans sa globalité. L'écart entre les sélections faites par les deux approches, et même par différents observateurs, peut être surprenant. En effet, suite à l'identification automatique des transferts de gènes entre diatomées et Archaeplastida, deux équipes différentes investiguent ces résultats par une approche manuelle. Combinées, l'analyse manuelle des arbres des deux études ne récupère que 10% des transferts identifiés en automatique (Moreira and Deschamps, 2014).

Il est important de garder en tête que la plupart des études réalisées sur l'hypothèse du ménage à trois et l'identification des transferts de gènes entre Chlamydia et Archaeplastida datent du début des années 2000 (Becker et al., 2008; Huang and Gogarten, 2007; Moustafa et al., 2008). Les résultats obtenus sont donc en accord avec les données génomiques, protéomiques et transcriptomiques disponibles à l'époque. Or, la diversité taxonomique prise

en compte dans les analyses phylogénétiques peut avoir un impact non négligeable sur les résultats obtenus. Les avancées techniques liées au séquençage ont considérablement augmenté la quantité (et parfois, mais pas toujours, la qualité) des données disponibles. Cependant, il existe encore aujourd'hui un déséquilibre représentatif de la diversité présente dans les bases de données, notamment du NCBI. En effet, en ne prenant en compte que RefSeq procaryote de mars 2021, 93% des données disponibles sont représentées par trois principaux phyla : les Proteobacteria, Firmicutes et Actinobacteria (Léonard et al., 2021). Les 50 autres phyla bactériens ne représentent alors que 7% de la base de données. Cette disproportion marquée peut être aussi visualisée dans les analyses phylogénétiques mises en place, et nécessite donc un effort d'échantillonnage représentatif de la diversité réelle du vivant.

c. Le rôle prédominant des Chlamydia et les transferts de gènes endosymbiotiques

Une des critiques principales apportées contre l'hypothèse du ménage à trois réside dans l'implication réelle du Chlamydia lors de l'endosymbiose primaire du plaste en comparaison des autres organismes (Dagan et al., 2013). Certains reconnaissent en effet les transferts de gènes chlamydien vers les Archaeplastida, mais les mettent en perspective avec les transferts de gènes identifiés pour les autres groupes bactériens, mettant de côté l'aspect fonctionnel de l'hypothèse. Dagan et al., 2013, évaluent les proportions de transferts de gènes en fonction des différents phyla procaryotes. Selon leurs résultats, les Chlamydia arrivent en 6ème position, derrière les cyanobactéries, Proteobacteria, Firmicutes, Actinobacteria et Bacteroidetes. Ainsi, avant de prendre en compte l'implication des pathogènes dans l'évolution des Archaeplastida, il faudrait aussi investiguer l'impact des autres groupes, le signal chlamydien devenant alors une partie seulement de la contribution bactérienne globale.

Il est utile de s'interroger ici sur la signification des transferts latéraux bactériens en général qui semblent associés au processus endosymbiotique. La mise en place de la symbiose entre la cyanobactérie et la cellule hôte eucaryote s'accompagne d'une dégénérescence de la bactérie, se traduisant par un nombre important de transferts de gènes depuis le génome cyanobactérien. Ces EGT (Endosymbiotic Gene Transfer), par définition, proviennent du symbionte et agrémentent le génome nucléaire eucaryote de nouvelles fonctions (Figure 6). Il se peut cependant, pour différentes raisons, qu'une version cyanobactérienne d'un gène eucaryote déjà en place lui soit préférée, actant ainsi le "grand remplacement" des gènes eucaryotes souches. Toujours dans un contexte endosymbiotique, d'autres sources bactériennes contribuent à la mise en place de la symbiose et au nouveau génome eucaryote. Ces ERGT (Endosymbiotic Related Gene Transfer), pour l'essentiel, portent sur des fonctions présentes chez les cyanobactéries mais dont les gènes ne sont pas retrouvés tels quels dans le noyau eucaryote, qui leur a préféré une version bactérienne

d'origine différente. Une explication plausible à la préférence pour des gènes "immigrés" plutôt que pour les gènes du symbionte cyanobactérien pourrait trouver sa source dans un examen des conditions qui sont exigées par le(s) système(s) de translocation au plaste des protéines synthétisées dans le cytosol de l'hôte. Il est utile de rappeler ici que les protéines destinées au plaste ne sont pas mises en conformation sous forme du repliement séquentiel de microdomaines à la sortie du tunnel du ribosome comme le sont la plupart des protéines cytosoliques bactériennes et eucaryotes; pas plus qu'elles ne suivent les voies de translocation et de mise en conformation des protéines vers d'autres compartiments (voie sec par ex) (Jarvis and Soll, 2002). Par contre, selon leur destination finale, elles se doivent d'interagir avec des chaperones spécifiques les maintenant dénaturées à leur sortie dans le cytosol et leur permettant d'être transportées et d'interagir ensuite avec des protéines diverses des translocons externes et internes du plaste ainsi qu'avec d'autres chaperones présidant à leur correcte mise en conformation finale *in situ* (Jarvis and Soll, 2002). Les séquences primaires des gènes cyanobactériens n'ont pas évolué pendant des centaines de millions d'années pour se « plier » à cet exercice. Il se peut donc que, par hasard, une séquence permette, en effet, de suivre le bon chemin, sans avoir à subir de mutations, auquel cas le gène plastidial sera très rapidement remplacé par une version nucléaire cyanobactérienne authentique de ce gène. Par contre, il reste possible que ce remplacement ne soit possible que si la séquence subit un certain nombre de mutations différant la sélection définitive d'une séquence dupliquée au noyau de l'hôte voire interdisant son remplacement par la séquence cyanobactérienne d'origine. Dans ce cas, le remplacement du gène d'origine se fera plus rapidement par une séquence étrangère nécessitant moins de mutations et qui entrera en compétition avec le gène cyanobactérien même si le transfert de cette séquence étrangère était, à l'origine, de fréquence bien inférieure à celle des gènes du symbionte. Comme le partenariat endosymbiotique ne vise pas nécessairement à reproduire le protéome d'origine de la cyanobactérie mais seulement à assurer la pérennité de la relation symbiotique au bénéfice de l'hôte, ce transfert latéral sera sélectionné. **Il est utile d'insister ici sur la dépendance exclusive de cette sélection relativement aux propriétés physicochimiques de la protéine concernée. Deux enzymes d'une même voie de synthèse ou de dégradation d'un même organisme n'afficheront pas la même adaptabilité sur ce plan. En d'autres termes, l'occurrence de deux remplacements de ce type émanant d'une même source bactérienne dans une même voie métabolique serait un événement extraordinairement rare voire inexistant**. Par contre les gènes cyanobactériens se comporteront différemment puisque c'est ce protéome qui guide et ordonne le processus de migration des gènes au noyau. Si la proposition MATH est incorrecte, on pourrait donc s'attendre 1°) à ce que le signal phylogénétique chlamydien dans le génome des Archaeplastida soit de même nature et d'importance identique voire inférieure comme l'affirme Dagan et al. aux autres signaux bactériens, 2°) que les transferts multiples dans une même voie métabolique soient par conséquent inexistants.

Par contre, si la proposition MATH est correcte, le chlamydia devenant lui-même un symbionte, ces ERGT devraient par définition être considérés comme des EGT et leurs propriétés se rapprocher de celles du signal EGT cyanobactérien. En particulier on peut s'attendre d'une part à ce que des transferts conjugatifs anciens aboutissent à des signaux multiples dans une voie et d'autre part que des effecteurs métaboliques au cœur du processus symbiotique aboutissent au même résultat.

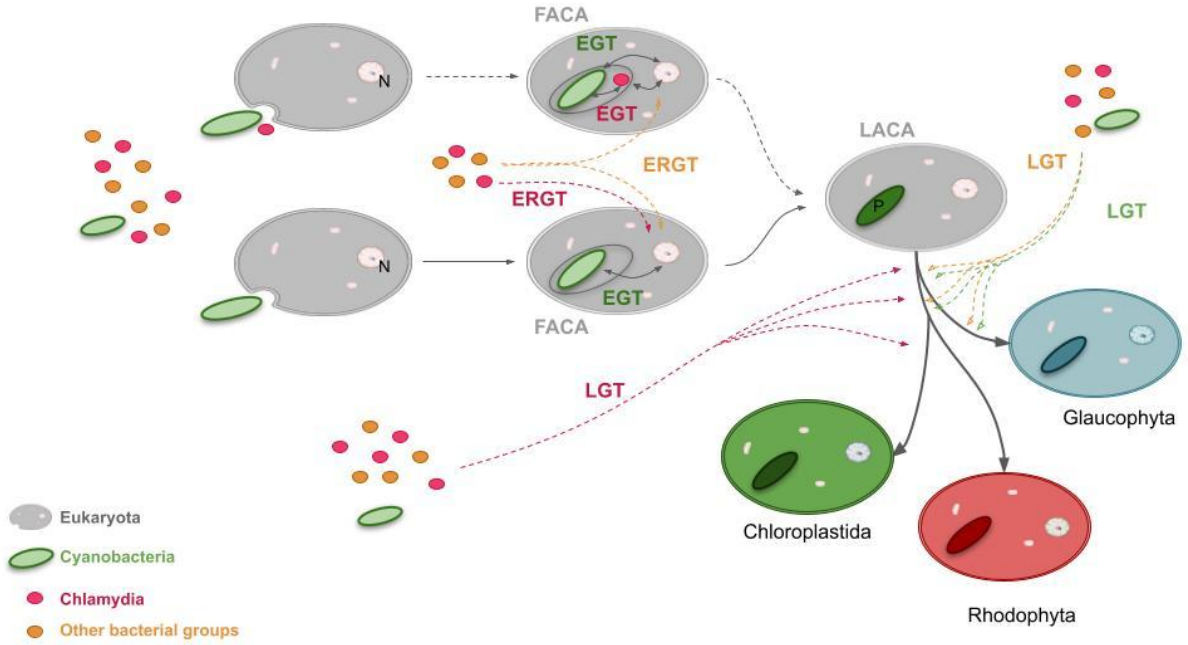


Figure 6 : Transferts de gènes horizontaux impliqués dans l'évolution des Archaeplastida. L'émergence des trois lignées d'Archaeplastida (Glaucophyta, Rhodophyta et Viridiplantae) fait suite à l'endosymbiose primaire du plaste, c-à-d l'internalisation d'une cyanobactérie photosynthétique par un eucaryote hétérotrophe (représenté par les flèches pleines). Le premier ancêtre commun des Archaeplastida (FACA) présente donc une cyanobactérie internalisée, en cours de dégénérescence. A ce stade, le génome du symbionte se réduit et un grand nombre de gènes sont transférés dans le génome nucléaire de la cellule hôte (EGT pour Endosymbiotic Gene Transfer). Les contributions bactériennes sont multiples au cours de l'évolution des Archaeplastida (flèches colorées hachurées), et peuvent intervenir dans un contexte endosymbiotique, que l'on appelle alors ERGT (Endosymbiotic Related Gene Transfer) mais aussi dans un contexte plus tardif de l'évolution au travers de LGT (Lateral Gene Transfer) prenant place alors lorsque le plaste est déjà intégré dans le dernier ancêtre commun des Archaeplastida (LACA) et ses descendants. Récemment, l'Hypothèse du Ménage à Trois propose l'implication d'un pathogène de type chlamydia lors de l'endosymbiose du plaste (flèches grises hachurées). Le chlamydia est alors un symbionte transitoire. Les transferts de gènes depuis les pathogènes peuvent donc être des trois types décrits : EGT et LGT, en fonction du contexte et du timing de transfert. A la différence des EGT, les LGT et ERGT peuvent être caractérisées comme des contributions "externes" à la symbiose et donc faire appel à plusieurs sources bactériennes (que ce soit au niveau taxonomique qu'au sein même d'un groupe d'espèces). Les transferts de gènes liés au contexte endosymbiotiques (EGT et ERGT) sont quand à eux plus susceptibles d'être retrouvés dans une grande diversité des Archaeplastida puisque réalisés lors de la mise en place de la symbiose avant le dernier ancêtre commun des Archaeplastida.

Objectifs du projet

Le but de ce projet est de tester l'Hypothèse du Ménage à Trois (MATH), et donc de préciser le rôle potentiel d'un partenaire Chlamydia dans l'endosymbiose primaire du plaste. A cette fin, l'étude sera organisée en deux parties complémentaires. Dans un premier temps, il s'agira d'examiner la nature du signal phylogénétique chlamydien chez les Archaeplastida et de vérifier qu'il puisse être interprété comme un vestige de MATH. Ensuite, dans un deuxième temps, il faudra comparer ce signal par rapport à d'autres signaux, choisis comme contrôles, afin de contrôler la spécificité de l'interaction, que ce soit du point de vue du donneur Chlamydia (contrôles bactériens) que de l'hôte Archaeplastida (contrôles eucaryotes). Ainsi nous pouvons identifier deux principales problématiques à ce projet :

- Existe-t-il un signal chlamydien chez les Archaeplastida spécifiquement lié à l'endosymbiose primaire du plaste ? Le cas échéant, ce signal est-il déjà connu de la littérature et confirme-t-il les fondations métaboliques de MATH ?
- Ce signal chlamydien chez les Archaeplastida est-il alors différent par comparaison avec d'autres signaux bactériens ?

Un protocole bioinformatique a donc été mis en place pour cribler les transferts latéraux de gènes (LGT) entre Chlamydia et Archaeplastida, puis une analyse manuelle des arbres générés pour les gènes retenus permet de repositionner ces transferts au sein de l'évolution des eucaryotes, répondant alors à la première partie de la problématique posée. L'automatisation complète du protocole permet ensuite de quantifier les signaux de différents groupes contrôles, tant bactériens qu'eucaryotes. et donc de vérifier le caractère unique de la contribution des Chlamydia aux Archaeplastida.

Résultats

1. Identification du signal chlamydien chez les Archaeplastida

a. Démarche générale

La démarche d'analyse du projet se décompose de manière générale en trois phases : i) le crible des LGT entre Chlamydia et Archaeplastida pour établir le signal phylogénétique, ii) l'analyse manuelle des arbres générés pour vérifier que le signal identifié puisse être un vestige d'une endosymbiose primaire du plaste tripartite et iii) l'automatisation complète du protocole, mimant l'analyse manuelle, afin de permettre une mise-en-perspective de ce signal à la lumière de contrôles positifs et négatifs, que ce soit d'un point de vue phylogénétique, fonctionnel ou métabolique.

Le développement d'un pipeline bioinformatique permet l'identification des LGT entre les différents groupes taxonomiques cibles et se compose de trois étapes majeures. Basé sur une sélection de génomes et de protéomes de qualité, représentative de la diversité du vivant, les groupes orthologues (OG ou clusters) intégrant le protocole sont i) d'abord filtrés sur des critères de présence des organismes cibles, avant d'être ii) enrichis avec la totalité des données génomiques et protéomiques sélectionnées et filtrées, puis iii) la reconstruction phylogénétique de chaque groupe entraînera la sélection des arbres sur des critères de topologie. Cette étape identifie chaque arbre présentant un branchement phylogénétique (clan) entre les espèces cibles (ici, entre Chlamydia et Archaeplastida). L'analyse manuelle des arbres sélectionnés par le pipeline confirme ou invalide le caractère endosymbiotique du LGT. En parallèle, le recensement des gènes répertoriés comme chlamydiens chez les Archaeplastida dans la littérature, ainsi que la comparaison de cet inventaire avec les résultats obtenus et l'analyse manuelle des arbres correspondants, permet la validation du protocole. L'ensemble des LGT identifiés par le pipeline puis validés par l'analyse manuelle constitue donc le signal phylogénétique global. A partir de l'analyse manuelle des arbres, autant de la sélection du pipeline que de l'inventaire de la littérature, le protocole bioinformatique est ensuite ajusté pour prendre en compte les observations faites et ainsi automatiser le pipeline. Cette conceptualisation automatique du protocole mime les résultats des analyses manuelles et permet la comparaison du signal chlamydien chez les Archaeplastida à d'autres. Deux contrôles sont alors dessinés: le contrôle des organismes "donneurs" des LGT (ici les Chlamydia), grâce à l'étude du signal de différents groupes bactériens chez ces eucaryotes photosynthétiques, et le contrôle des organismes "accepteurs" des LGT (ici les Archaeplastida) par l'étude du signal chlamydien chez d'autres groupes eucaryotes. La spécificité du rôle chlamydien lors de l'endosymbiose primaire du plaste peut ensuite être évaluée par la caractérisation et la comparaison de chaque sélection de LGT. Il s'agit d'abord

de quantifier le nombre de transferts de gènes, puis d'en évaluer la congruence du signal, pour ensuite analyser la diversité des organismes présentant ces transferts et enfin les remettre dans un contexte métabolique et fonctionnel.

b. Crible des LGT Chlamydia - Archaeplastida

Notre étude se base sur l'identification des transferts de gènes entre Chlamydia et Archaeplastida qui, pris ensemble et dans leur globalité, reconstruisent en partie l'histoire évolutive commune de ces organismes, que l'on définit ici comme signal phylogénétique. Ainsi, dans l'idée d'avoir une identification de ces LGT la plus précise possible, en limitant au maximum la proportion de faux-positifs due à des contaminations ou à une sur-représentation taxonomique, nous avons opté pour l'utilisation d'une base de données de qualité et représentative de la diversité. La qualité de chaque protéome ou génome entrant dans le pipeline a donc été évaluée. Parmi l'ensemble des espèces sélectionnées pour cette partie du projet se trouvent 72 eucaryotes, dont 54 eucaryotes photosynthétiques, 33 Chlamydia, 48 cyanobactéries, et deux ensembles de 49 et 92 bactéries, sans distinction de taxonomie, desquels nous avons retiré les Chlamydia et cyanobactéries (annexe 9).
Le protocole bioinformatique mis au point pour cette étude comporte trois étapes principales (Figure 7). Une pré-sélection des arbres générés, comprenant uniquement les 57 eucaryotes, les Chlamydia et les cyanobactéries, permet de mettre en compétition directe les trois partenaires supposés de l'Hypothèse du Ménage à Trois, et ainsi repérer les transferts de gènes pour lesquels le signal chlamydien l'emporte clairement sur celui des cyanobactéries. L'enrichissement de cette pré-sélection avec le reste des génomes et protéomes permet ensuite de tester la robustesse du signal chlamydien, cependant, seule l'analyse manuelle des arbres valide ces transferts et permet d'en déterminer le potentiel contexte endosymbiotique.
210451 groupes orthologues ont été créés par OrthoFinder à partir des protéomes des 57 eucaryotes (dont 412 Rhodophyta et 8 Viridiplantae). Parmi eux, seuls 8583 contenaient au moins 2 Viridiplantae et / ou Rhodophyta. Après enrichissement de chaque groupe orthologue avec les protéomes de Chlamydia et cyanobactéries, nous avons retenu pour la suite du protocole uniquement ceux présentant au moins une séquence chlamydienne. Ainsi, 987 groupes orthologues ont subi le filtrage des séquences et la reconstruction phylogénétique. L'analyse automatique des bipartitions a révélé 201 arbres présentant un clan phylogénétique avec au moins un Chlamydia et au moins un Archaeplastida (Figure 7).

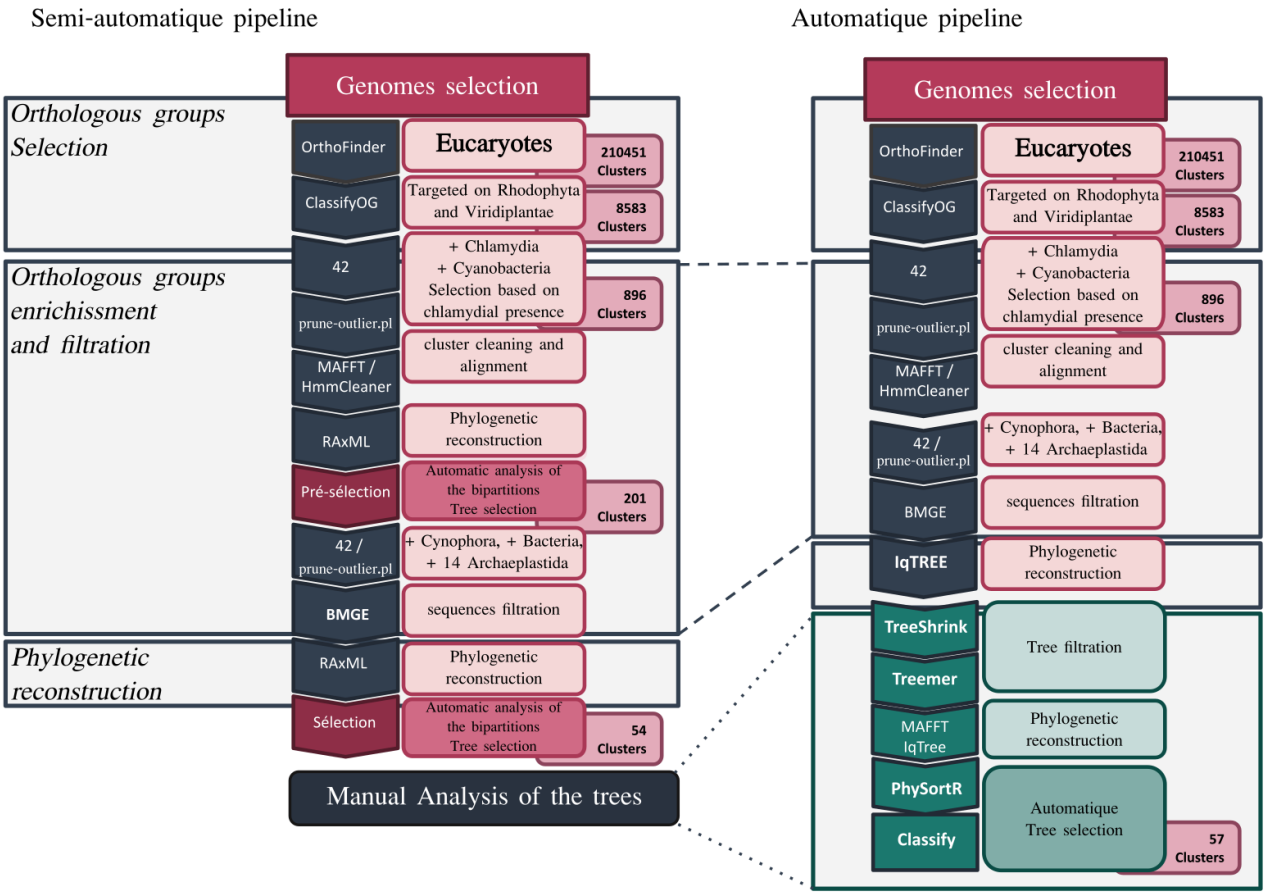


Figure 7: Organigramme méthodologique du crible des LGT Chlamydia-Archaeplastida. A gauche, organigramme du pipeline semi-automatique, basé sur la sélection des génomes et protéomes utilisés, puis composé des trois principales étapes : i) création et sélection des groupes orthologues, ii) enrichissement et filtrage de ces groupes et iii) reconstruction phylogénétique des alignements pré-sélectionnés. L'enchaînement des outils utilisés apparaît au centre de l'organigramme. Ce pipeline est semi-automatique puisqu'une analyse manuelle des arbres générés vient valider la sélection. Le nombre de groupes orthologues restants à chaque étape du protocole est indiqué sur la partie droite de l'organigramme. Le pipeline automatique, à droite, est quant à lui une adaptation du pipeline semi-automatique calibré par l'analyse manuelle des arbres, de sorte à ce que les deux versions du pipeline soient équivalentes. Les principales étapes restent les mêmes, à l'exception du retrait de la pré-sélection intermédiaire des groupes orthologues. Le remplacement de l'analyse manuelle des arbres permet de réduire les biais du protocole et de l'appliquer aux autres groupes bactériens.

Les 201 alignements correspondants ont ensuite été enrichis avec le reste des protéomes et génomes, notamment provenant d'autres bactéries, mais aussi de glaucophytes. En effet, consciente de la qualité moindre des données disponibles pour les Glaucocystophyceae, celles-ci n'ont pas été intégrées dans les étapes initiales du pipeline. Comme précédemment décrit, une étape de filtrage des clusters est nécessaire avant de réaliser la reconstruction phylogénétique. Cependant, en ce qui concerne cette deuxième

étape du protocole, plusieurs outils ont d'abord été testés afin d'optimiser les résultats, autant en qualité qu'en temps de calcul (Figure 8).

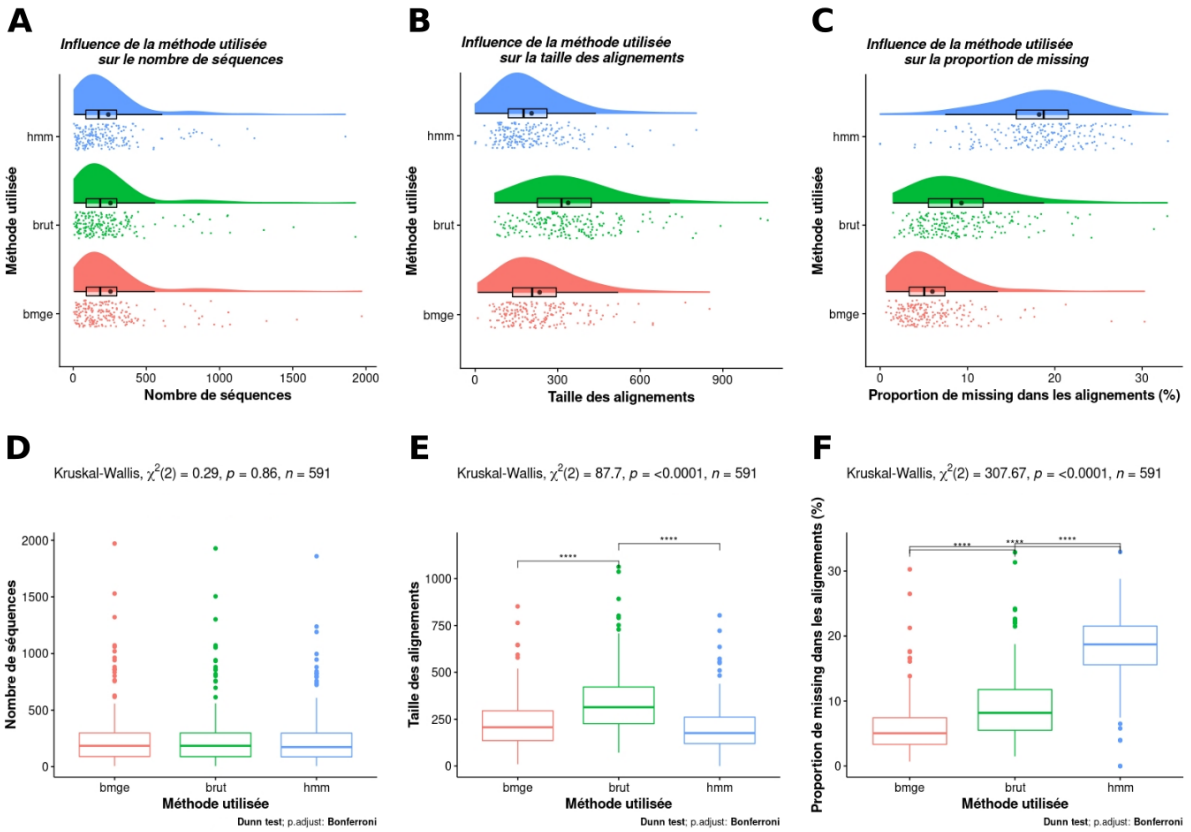


Figure 8: Tests de l'influence de trois méthodes de filtration de séquences sur les alignements. Le même échantillon de groupes orthologues sélectionnés suite à l'étape de pré-sélection du pipeline semi-automatique a été nettoyé par HmmCleaner (en bleu), BMGE (et plus précisément par ali2phylip.pl mask BMGE, en rouge) et uniquement par ali2phylip.pl("brut", en vert). Pour chacune des méthodes utilisées, le nombre de séquences total dans les alignements (A, D), la longueur des alignements (B, E), ainsi que la proportion de "missing" (C, F) ont été comparés. L'évaluation statistique, réalisée par un test de Kruskal-Wallis, suivi d'un test de Dunn (p-adjust Bonferroni), montre une différence significative entre les méthodes en ce qui concerne la longueur des alignements (E) et la proportion de "missing" (F).

Ainsi, nous avons comparé trois méthodes de filtration des groupes orthologues : 1) HmmCleanner combiné à ali2phylip.pl; 2) ali2phylip.pl combiné à BMGE et 3) ali2phylip.pl seul (nommés respectivement HmmCleaner, BMGE et brut). Pour chaque condition, la reconstruction phylogénétique s'est effectuée avec RAxML, sous un modèle LG4X et ultrafast-bootstrap, sur un même échantillon issu de la pré-sélection du pipeline. Nous avons ainsi pu comparer les sélections obtenues en sortie de pipeline, mais aussi l'influence de ces méthodes sur les alignements eux-mêmes, de sorte à en garder la version optimale. Les méthodes de filtration citées impactent les alignements dans leur globalité, et influencent la reconstruction phylogénétique par la suite. Les effets principaux de ces méthodes portent sur la longueur des alignements et leur proportion de sites manquants ("missing"). Ainsi,

l'analyse des distributions du nombre de séquences total dans les arbres montre une stabilité, une similarité des profils pour les trois conditions testées (Figure 8A), avec en moyenne environ 250 séquences par arbre. Aucune de ces trois méthodes ne réduit donc trop drastiquement l'échantillon et permet de conserver la diversité taxonomique de chaque arbre liée aux étapes d'enrichissement. Cependant, la longueur des alignements et la proportion de sites manquants montrent des différences significatives entre les trois conditions. Concernant la taille des alignements d'abord, HmmCleaner et BMGE semblent présenter le même profil de distribution (Figure 8B), avec une longueur moyenne située autour de 220 acides aminés, par rapport à la condition "brut". Cette différence est confirmée par un test statistique de Kruskal-Wallis (Figure 8E). Ceci est attendu puisque HmmCleaner et BMGE sont des méthodes de filtration des alignements et donc nécessairement en réduisent la taille. Les distinctions s'accentuent en analysant l'impact de chaque condition sur la proportion de sites manquants dans les alignements. En effet, les trois méthodes testées présentent ici des profils de distributions différents (Figure 8C), avec une proportion de "missing" plus élevée pour HmmCleaner, située autour de 20%, par rapport à la méthode "brut" (~10%) et BMGE (~6%). Ces différences sont significatives selon un test de Kruskal-Wallis (Figure 8F). Aussi, en mettant en relation les deux derniers paramètres analysés, nous constatons pour HmmCleaner une taille d'alignement identique à BMGE pour une proportion de sites manquants supérieure. Comme dit précédemment, ces caractéristiques influencent la reconstruction phylogénétique, et donc, dans notre cas, influencent aussi le nombre d'arbres sélectionnés par notre protocole. En sortie du pipeline, 67 arbres sont sélectionnés pour HmmCleaner, contre 54 pour BMGE et 54 pour "brut", dont 37 sont communs aux trois approches. A première vue, la condition HmmCleaner semble être celle à privilégier, puisqu'elle permet l'identification d'un plus grand nombre de LGT potentiels. Cependant, une analyse manuelle rapide invalide une grande majorité des arbres sélectionnés uniquement par HmmCleaner (seuls deux arbres sur les 27 sélectionnés uniquement par HmmCleaner sont en effet confirmés en analyse manuelle). Ainsi, nous avons choisi pour la suite du protocole d'utiliser la combinaison ali2phylip.pl et BMGE pour le filtrage des séquences. L'analyse automatique des bipartitions par clans-label.pl des arbres générés par RAxML identifie alors 54 arbres montrant une association phylogénétique entre au moins une Chlamydia et un Archaeplastida (Figure 7). Cette sélection reste cependant imprécise. Une analyse manuelle des arbres est nécessaire non seulement pour valider le protocole, la méthodologie de manière générale, mais aussi pour valider les transfert de gènes identifiés et en vérifier le caractère endosymbiotique.

c. Analyse manuelle des arbres

L'analyse manuelle des arbres générés par le pipeline a été réalisée en partenariat avec Dr Ingrid Lafontaine et Dr Clotilde Garrido. Les buts de cette analyse manuelle sont multiples. Dans un premier temps, elle permet de valider la méthodologie mise en place pour l'identification des transferts de gènes entre les organismes cibles. A cette fin, elle intervient d'ailleurs tout au long de la mise au point du protocole. Ensuite, elle est nécessaire pour valider les transferts, c'est-à-dire distinguer les arbres correctement sélectionnés par la méthodologie des faux positifs. Enfin, seule l'analyse manuelle des arbres permet de replacer le transfert de gène identifié dans un contexte endosymbiotique. Ainsi, cette partie de l'étude constitue une étape importante dans l'établissement de la nature du signal chlamydien chez les Archaeplastida. Cependant, l'analyse des arbres simples gènes reste difficile et leur interprétation varie selon les observateurs et les critères d'évaluation retenus. De plus, en fonction des méthodes de reconstruction appliquées, la topologie des arbres elle-même peut être modifiée. Cette particularité fait d'ailleurs partie des critiques retenues contre l'Hypothèse du Ménage à Trois. Aussi, pour essayer de limiter le biais d'interprétation des arbres, et objectiver au maximum les résultats obtenus, trois observateurs indépendants se sont chargés d'analyser manuellement chacun des 54 arbres sortis du pipeline précédemment décrit selon des critères définis en amont.

Le premier des critères pris en compte dans l'analyse des arbres est la qualité et la diversité des espèces présentes dans le clan Chlamydia - Archaeplastida. En effet, une diversité plus importante, autant du côté donneur (Chlamydia) que du côté accepteur (Archaeplastida et plus généralement eucaryotes photosynthétiques) des LGT, témoigne d'un transfert plus ancien, ayant concerné l'ancêtre commun d'un plus grand nombre d'organismes. Ceci est surtout vérifié lorsque le transfert considéré implique plus d'une lignée d'Archaeplastida. La même logique peut s'appliquer lorsque les espèces les plus basales des donneurs sont présentes dans le clan d'intérêt. En effet, la représentation chlamydienne dans le clan, et plus particulièrement la présence des espèces basales, confirmerait l'origine du transfert de gène depuis ce groupe. Ces deux caractéristiques, autant la présence des espèces basales que le transfert de gènes chez de multiples lignées d'Archaeplastida, permettent d'estimer la direction et la temporalité de ces LGT, pointant alors comme receveur l'ancêtre commun de ces eucaryotes photosynthétiques. Le deuxième critère pris en compte lors de cette analyse des arbres phylogénétiques tient de la présence d'intrus dans le sous-arbre d'intérêt, qui peut en effet traduire des transferts plus disséminés dans l'évolution et le monde vivant en fonction de la diversité des espèces présentes. En d'autres mots, la présence potentielle de ces espèces non-cibles peut aussi bien témoigner de plusieurs événements de transfert indépendants les uns des autres, dans un contexte endosymbiotique ou non, que d'un LGT non lié à l'endosymbiose primaire du plaste. L'analyse de la diversité des organismes dans le

clan permet généralement de différencier les différents cas de figure. Enfin, la topologie générale de l'arbre est étudiée, de sorte à identifier les possibles paralogues en cas de famille multigénique, mais aussi, dans le cas où les arbres montrent un clan entre les Chlamydia et une ou deux lignées d'Archaeplastida, à visualiser la position phylogénétique des autres lignées d'Archaeplastida. Ainsi, la présence dans le même arbre d'un clan Chlamydia-Archaeplastida et d'un autre clan Cyanobacteria-Archaeplastida, atteste de la temporalité MATH du LGT. A l'aide de ces critères, les arbres phylogénétiques analysés sont classés en trois catégories différentes : 1) en faveur de l'Hypothèse du Ménage à Trois, c'est-à-dire montrant un signal pouvant être replacé dans un contexte endosymbiotique, 2) pas en faveur de MATH et 3) incertain, impossible à conclure.

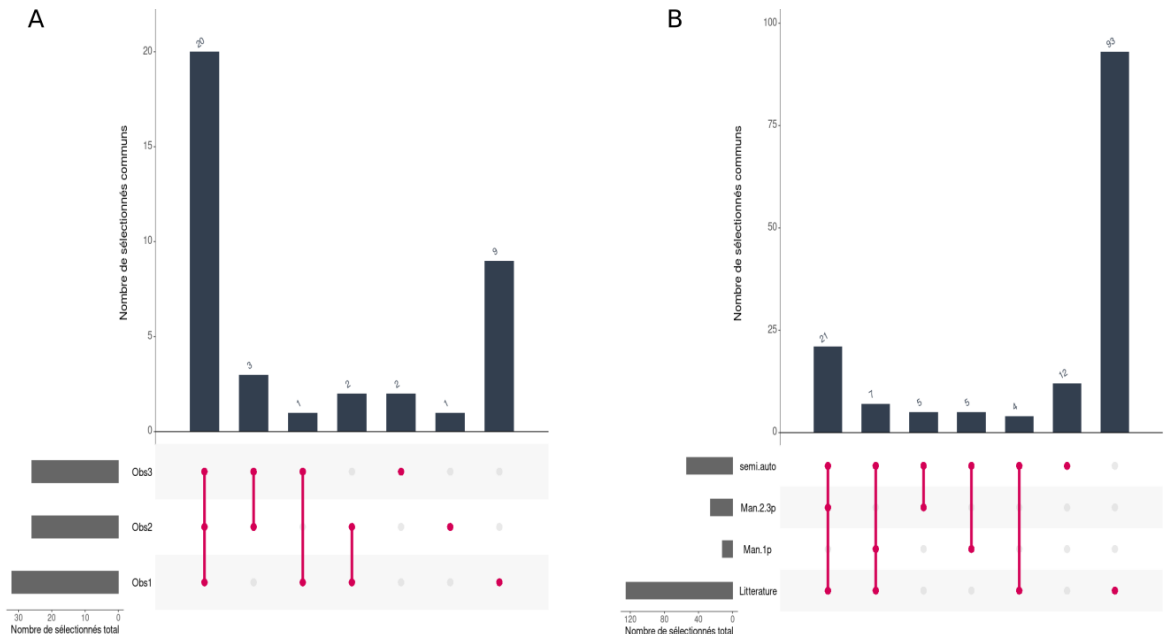


Figure 9: Diagrammes récapitulatifs de la sélection du pipeline semi-automatique et de l'analyse manuelle des arbres générés. Le pipeline semi-automatique a sélectionné 54 arbres avec un clan regroupant Chlamydia et Archaeplastida, qui sont ensuite manuellement analysés par trois observateurs. 37 de ces arbres sont confirmés comme présentant un transfert de gène lié à l'endosymbiose primaire du plaste par au moins un des trois observateurs. Les diagrammes représentent les intersections entre les différents ensembles considérés. Chaque point rose permet de visualiser les groupes pris en compte. Le panel A reprend les résultats de l'analyse manuelle des arbres en fonction des observateurs (uniquement les arbres confirmant un transfert de gène). Le détail de l'ensemble de la sélection du pipeline semi-automatique est présenté dans le panel B, comparé à l'analyse manuelle et aux données de la littérature. Obs1 : observateur 1 ; Obs2 : observateur 2 ; Obs3 : observateur 3, semi.auto : pipeline semi-automatique ; Man.2.3p : arbres confirmés par 2 ou 3 observateurs lors de l'analyse manuelle ; Man.1p : arbres confirmés par 1 seul observateur lors de l'analyse manuelle ; Littérature : gènes rapportés comme chlamydiens chez les Archaeplastida dans (Ball et al., 2013; Becker et al., 2008; Cenci et al., 2018, 2017; Huang and Gogarten, 2007; Moustafa et al., 2008).

Sur 54 arbres analysés, 37 sont sélectionnés par au moins un observateur comme étant en faveur de l'Hypothèse du Ménage à Trois, c'est-à-dire montrant un transfert de gène pouvant être relié à un contexte endosymbiotique, parmi lesquels 20 font l'unanimité et 6 autres sont sélectionnés par deux observateurs sur trois (Figure 9A). En analysant plus en détail les différences de sélection entre les trois observateurs, nous remarquons que les 12 arbres sélectionnés par un seul sont en majorité catégorisés comme incertains ou impossibles à conclure pour les deux autres. De manière générale, nous décidons donc de valider un arbre comme étant compatible avec l'Hypothèse du Ménage à Trois si au moins deux observateurs sur trois l'ont classifié comme tel. De ce fait, 26 arbres sortant du pipeline sont validés par l'analyse manuelle et appuient une implication chlamydienne dans l'évolution des Archaeplastida.

d. Congruence du signal

L'analyse manuelle permet donc de valider les transferts de gènes visibles sur les arbres phylogénétiques simples gènes, et de potentiellement confirmer leur occurrence dans un contexte endosymbiotique. Une supermatrice, créée par SCaFoS (Roure et al., 2007) à partir des seules séquences de Chlamydia et d'Archaeplastida présentes dans les clans des 26 arbres sélectionnés et pour laquelle la reconstruction phylogénétique est réalisée par IQ-TREE (Nguyen et al., 2015), nous informe ensuite sur la congruence du signal global chlamydien chez les Archaeplastida. En d'autres termes, l'analyse manuelle des arbres permet d'identifier et de valider les gènes MATH, alors que la concaténation permet de déterminer si leur signal est congruent. Quel que soit le modèle utilisé pour la reconstruction phylogénétique (LG4X, C20 ou C60), lorsque les arbres sont enracinés sur les similichlamydia (Chlamydia basales), nous observons le regroupement des trois lignées d'Archaeplastida, branchées directement avec les Chlamydia environnementales (Figure 10 + annexe 1). De plus, au sein du clan Archaeplastida, la topologie obtenue respecte les relations évolutives connues chez les algues et végétaux. Ce regroupement des Archaeplastida appuie la congruence du signal, et suggère donc une histoire évolutive similaire des gènes sélectionnés, du moins pour la partie de leur évolution post-endosymbiotique. En effet, dans le cas contraire, les Archaeplastida auraient été dispersés au sein des Chlamydia.

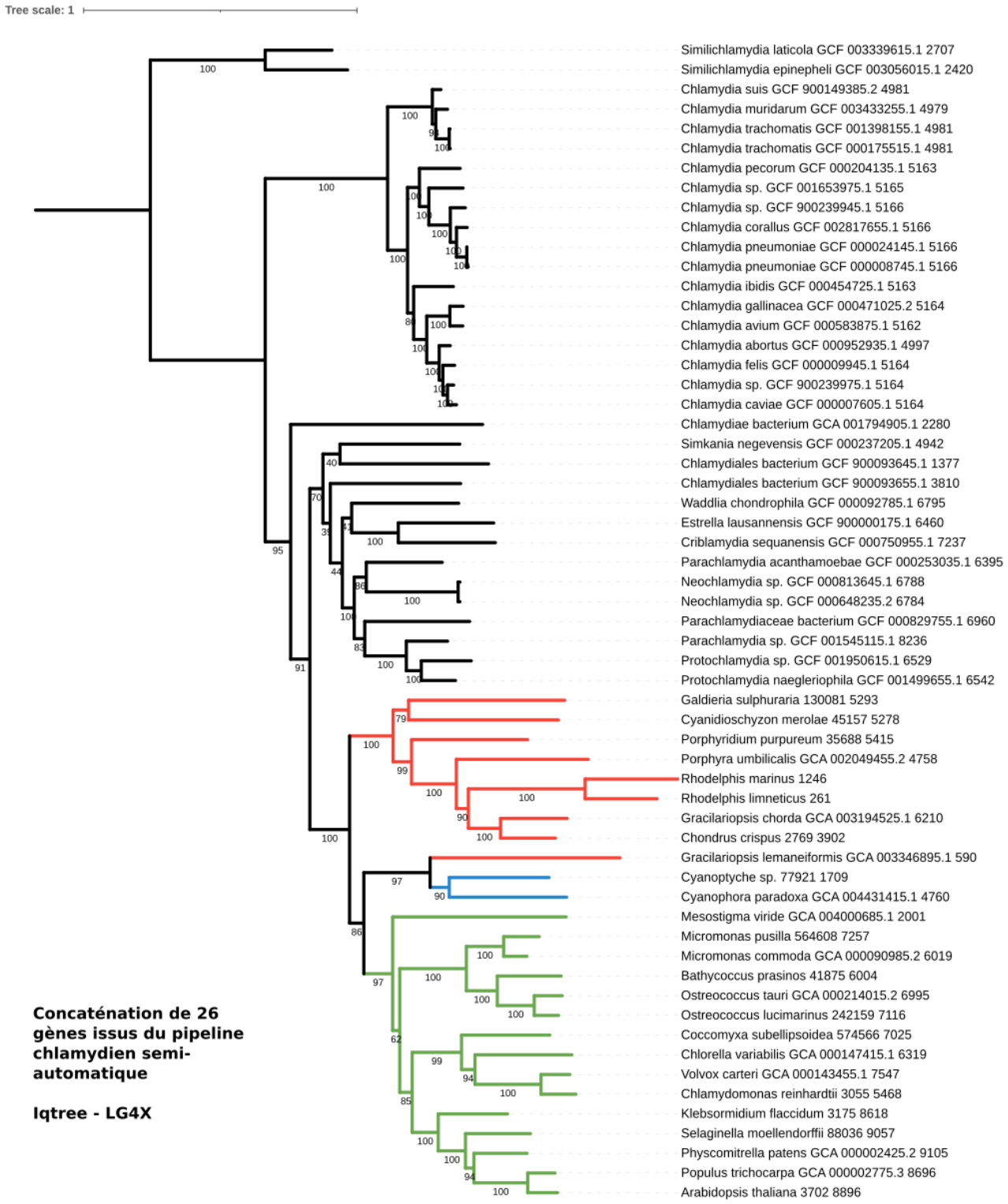


Figure 10: Arbre phylogénétique issu de la concaténation des 26 gènes retenus par l'analyse manuelle. L'arbre a été obtenu par IQ-TREE, sous un modèle LG4X, après concaténation des séquences chlamydiennes et archaeplastiennes des 26 gènes confirmés en analyse manuelle. Les valeurs de bootstrap sont indiquées sur les branches. L'enracinement est manuel sur les simili-Chlamydia. L'échelle compte en nombre de substitutions des acides aminés par site. En rouge : Rhodophyta et Rhodelphea, en vert : Viridiplantae et en bleu : Glaucophyta.

e. Inventaire des gènes chlamydiens dans la littérature et corrélation avec notre analyse

Plusieurs études ont par le passé essayé d'éclaircir l'impact des Chlamydia dans l'évolution des Archaeplastida, notamment en réalisant un catalogue des gènes transférés des premiers aux seconds. A partir de ces études, nous avons repris toutes les séquences d'Archaeplastida rapportées comme étant d'origine chlamydienne dans Ball et al., 2013; Becker et al., 2008; Cenci et al., 2018, 2017; Huang and Gogarten, 2007; Moustafa et al., 2008. Cet inventaire de la littérature comprend 150 séquences qui, une fois dérépliquées pour éviter les doublons dû aux synonymes, se distribuent dans 128 de nos groupes orthologues de départ. En sortie de pipeline, 32 OG sélectionnés parmi les 54 sont aussi retrouvés dans la littérature et, parmi les 26 validés en analyse manuelle, 21 sont aussi identifiés dans cet inventaire (Figure 9B). L'ensemble des arbres correspondant à l'inventaire (en particulier ceux non retenus) ont alors été analysés manuellement pour valider et comprendre la sélection du pipeline. C'est au cours de cette analyse manuelle que nous avons pu remarquer une limite de notre pipeline : la majorité des arbres de l'inventaire sont considérés comme des faux positifs selon nos méthodes. Cependant, à la réflexion, certains auraient bien dû être sélectionnés. En approfondissant l'observation de ces arbres, il apparaît que les arbres en question présentent des clans pour lesquels Chlamydia et Archaeplastida branchent ensemble, mais aussi avec d'autres espèces. Ainsi, l'outil en charge de l'analyse automatique des bipartitions ne pouvait pas identifier le sous arbre d'intérêt puisqu'il ne prend pas en compte les clans interrompus par des "intrus", c'est-à-dire des espèces non cibles. Dans ce cas-ci il s'agit donc de toutes les bactéries et eucaryotes non photosynthétiques.

L'analyse manuelle des arbres, autant de la sélection du pipeline que ceux correspondant à l'inventaire de la littérature, nous permet non seulement de valider et d'ajuster nos méthodes, mais aussi, et surtout, de servir de base pour automatiser entièrement le pipeline de sorte qu'il mime les résultats obtenus à la main. En effet, en fonction des observations réalisées manuellement, nous avons paramétré nos méthodes afin de corriger les biais de sélection et retranscrire de façon automatique nos critères d'analyse.

2. Significativité du signal chlamydien chez les Archaeplastida

L'analyse de l'hypothèse du ménage à trois nécessite d'abord d'identifier le signal chlamydien chez les Archaeplastida, puis ensuite seulement de déterminer si ce signal est différent par rapport à d'autres, de sorte à évaluer l'impact spécifique des Chlamydia sur l'origine des eucaryotes photosynthétiques, non seulement d'un point de vue quantitatif mais aussi d'un point de vue fonctionnel et métabolique. Jusqu'à présent, nous nous sommes

concentrés sur l'identification de ce signal et de sa nature phylogénétique, montrant ainsi la présence de transferts de gènes entre Chlamydia et Archaeplastida dont l'origine remonte à l'endosymbiose primaire du plaste.
Il est maintenant nécessaire de replacer ce signal identifié au sein de l'histoire évolutive globale des bactéries et eucaryotes photosynthétiques, afin de déterminer si les Chlamydia ont joué un rôle particulier lors de l'endosymbiose primaire du plaste. Ceci implique donc de comparer le signal chlamydien chez les Archaeplastida à d'autres signaux bactériens, mais aussi d'évaluer la spécificité du lien avec les Archaeplastida en comparaison à d'autres groupes eucaryotes. Ce contrôle du signal chlamydien chez les Archaeplastida est aussi mis en relief par rapport au signal cyanobactérien, utilisé ici comme référentiel.

a. Automatisation du pipeline

L'étape limitante du protocole décrit dans le premier chapitre reste l'analyse manuelle des arbres. En effet, cette étape est certes nécessaire pour valider les transferts de gènes et surtout pour identifier les gènes dits "MATH", mais chronophage et critiquable. Puisqu'il s'agit maintenant de comparer différents signaux phylogénétiques, autant du côté accepteur que du côté donneur des LGT, et non pas d'établir un listing exhaustif des gènes MATH, nous calibrons le pipeline sur les résultats de l'analyse manuelle pour mettre au point un protocole entièrement automatisé. De ce fait, en se basant sur les observations faites en analyse manuelle, sur les résultats obtenus et sur les critères d'analyses mis en place précédemment, nous avons ajusté le pipeline bioinformatique (Figure 7).

L'étape de pré-sélection des arbres, ne portant encore que sur les espèces cibles retrouvées dans un contexte MATH (eucaryotes, Chlamydia et cyanobactéries), a tout d'abord été supprimée, de sorte à gagner du temps de calcul. Ensuite, toujours dans l'idée de gagner du temps de calcul, nous avons remplacé RAxML (Stamatakis, 2014) par IQ-TREE (Nguyen et al., 2015). En effet, pour la reconstruction phylogénétique des mêmes alignements sous les mêmes modèles, RAxML et IQ-TREE donnent des résultats très similaires. La différence des scores de maximum de vraisemblance pour chaque paire d'arbres est négligeable, même si systématiquement en faveur de RAxML (score moyen de 27500, avec une différence moyenne entre RAxML et IQ-TREE de 32). Par conséquent, la sélection finale par le pipeline précédemment décrit identifie en majorité les mêmes arbres comme ayant une interaction phylogénétique entre au moins un Chlamydia et un Archaeplastida. Cependant, IQ-TREE est (beaucoup) plus rapide. Pour valider entièrement le changement de méthode, les 8 arbres sélectionnés par RAxML et non par IQ-TREE ont été manuellement analysés.

L'automatisation du pipeline repose entièrement sur l'analyse manuelle faite précédemment, autant sur les résultats obtenus que sur les critères mis en place pour l'étude des arbres. En

effet, chaque étape de l'analyse manuelle a été traduite en outil bioinformatique automatisé. La qualité de l'inférence des donneurs et des accepteurs est d'abord améliorée par Treeshrink (Mai and Mirarab, 2018) et Treemer (Menardo et al., 2018) qui, respectivement, retirent les longues branches et dérépliquent les arbres, sans pour autant réduire le nombre d'espèces présentes. PhySortR (Stephens et al., 2016) et classify-ali.pl (https://metacpan.org/dist/Bio-MUST-Core) combinés s'occupent alors de l'identification des transferts de gènes en eux-mêmes, tout en prenant en charge les autres critères établis lors de l'analyse manuelle. PhySortR est un package R qui permet d'identifier les arbres phylogénétiques présentant un clan composé d'espèces prédéterminées, tout en prenant en compte la possibilité d'espèces non cibles coupant les clans d'intérêt (les intrus), ainsi que la proportion d'espèces cibles dans le clan sélectionné par rapport à la totalité de l'arbre. Plusieurs tests de ces paramètres ont été réalisés pour ajuster au mieux la performance de sélection du package. Ainsi, la proportion minimale d'espèces cibles dans le sous-arbre correspondant au branchement d'intérêt par rapport à l'arbre complet est fixée à 30% et nous autorisons la présence de 10% d'intrus. En d'autres termes, est sélectionné chaque arbre pour lequel au moins 30% de la totalité des espèces cibles (Chlamydia et eucaryotes issus de l'endosymbiose primaire du plaste en ce qui concerne le pipeline chlamydien) de l'arbre total sont présentes dans le sous- arbre d'intérêt, et pour lequel moins de 10% des espèces de ce sous- arbre sont autres que ces espèces cibles, est sélectionné. Alors que PhySortR gère les critères de diversité relative générale des arbres et de présence d'intrus pouvant couper les clans, classify-ali.pl affine la sélection faite en filtrant les sous-arbres identifiés sur un critère de diversité absolue. Ici, nous plaçons le minimum d'espèces présentes dans le clan d'intérêt à 5, dont au moins 2 donneurs et 3 accepteurs, pour qu'un arbre soit sélectionné. L'ensemble des modifications apportées au pipeline sont matérialisées sur la Figure 7 .

Concrètement, si on reprend le crible des LGT entre Chlamydia et Archaeplastida, la nouvelle version du pipeline sélectionne 57 arbres ayant une interaction phylogénétique entre au moins 2 Chlamydia et 3 Archaeplastida. En comparaison, le pipeline semi-automatique précédent sortait 54 arbres, dont 26 étaient validés en analyse manuelle (2 ou 3 avis). Parmi ces 57 arbres, 26 sont communs à la précédente sélection, et tous sont confirmés par l'analyse manuelle par au moins un observateur (21 arbres sont sélectionnés par 2 ou 3 observateurs, 5 le sont par un seul) (Figure 11A). L'analyse automatique manque donc 28 arbres (54-26), quoique essentiellement douteux voire invalides (23/28), tout en amenant un supplément de 31 arbres (57-26). En observant plus en détail ces arbres additionnels, notamment en les comparant à leur correspondant du pipeline semi-automatique, nous constatons, comme lors de l'analyse de l'inventaire de la littérature, que ce sont surtout les arbres présentant des clans interrompus par des intrus qui n'ont pas été identifiés par le premier pipeline. L'ajustement du protocole, notamment par le choix de PhySortR, permet donc de les récupérer ces faux-négatifs et d'assembler un catalogue plus complet des gènes MATH. Par contre, 5 arbres

confirmés en analyse manuelle par au moins deux des trois observateurs ne sont pas sélectionnés par cette nouvelle version du pipeline. L'objectif étant de comparer les signaux phylogénétiques des différents groupes bactériens, nous acceptons toutefois ces nouveaux faux-négatifs. En effet, il est important de garder en tête que le but du protocole automatique présenté ici n'est pas d'établir un catalogue exhaustif des gènes MATH. De ce fait, les paramètres ont été choisis pour optimiser cette comparaison, en toute conscience d'une sensibilité pas forcément maximale, conduisant au rejet possible de quelques gènes MATH.

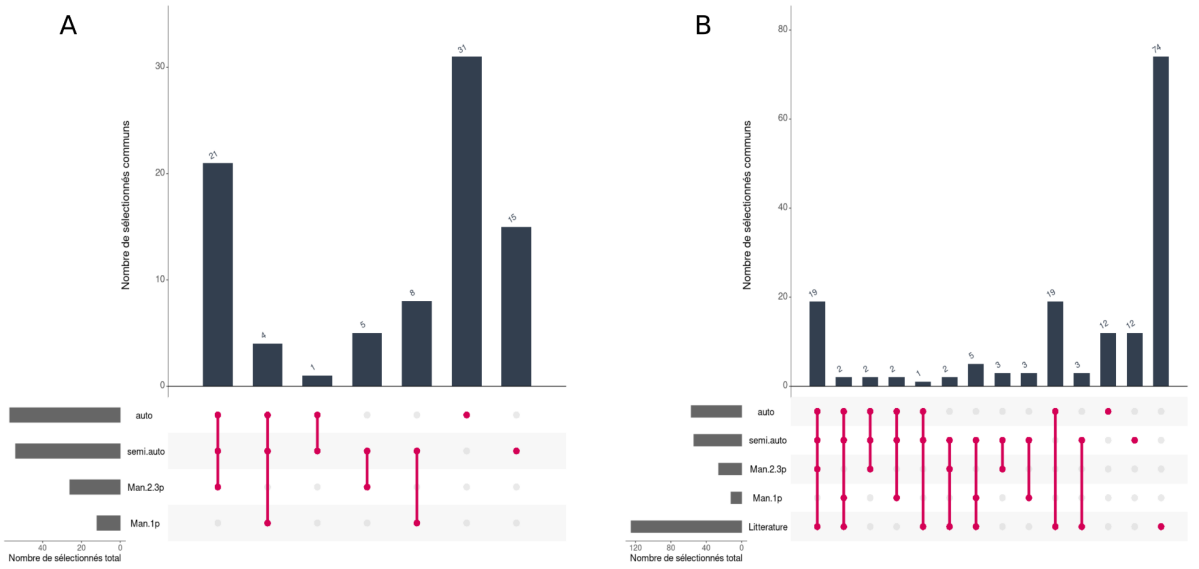


Figure 11: Diagrammes récapitulatifs des sélections en fonction des différentes méthodes. Les diagrammes représentent les intersections entre les différents ensembles considérés. Chaque point rose permet de visualiser les groupes pris en compte. Le pipeline semi-automatique (semi.auto) a sélectionné 54 arbres avec un branchement entre Chlamydia et Archaeplastida, qui sont ensuite manuellement analysés par trois observateurs. 26 de ces arbres sont confirmés comme présentant un transfert de gène lié à l'endosymbiose primaire du plaste par au moins deux des trois observateurs (Man.2.3p), 15 autres ne sont validés que par un seul des observateurs (Man.1p). Le pipeline automatique (auto) a quant à lui identifié 57 arbres avec une relation phylogénétique entre Chlamydia et Archaeplastida. Le panel A permet de visualiser les sélections communes entre les différentes méthodes. Le panel B reprend aussi les gènes chlamydiens identifiés chez les Archaeplastida dans la littérature. semi.auto : pipeline semi-automatique ; Man.2.3p : arbres confirmés par 2 ou 3 observateurs lors de l'analyse manuelle ; Man.1p : arbres confirmés par 1 seul observateur lors de l'analyse manuelle ; Littérature : gènes rapportés comme chlamydiens chez les Archaeplastida dans (Ball et al., 2013; Becker et al., 2008; Cenci et al., 2018, 2017; Huang and Gogarten, 2007; Moustafa et al., 2008).

b. Pipelines contrôles

Deux types de contrôles sont envisagés pour évaluer dans son ensemble le signal chlamydien chez les Archaeplastida : un contrôle des "donneurs" à l'origine des transferts de gènes et un contrôle des "accepteurs". Le premier permet d'évaluer le signal chlamydien chez

les Archaeplastida en comparaison d'autres groupes bactériens, le deuxième de déterminer l'impact des Chlamydia spécifiquement chez les Archaeplastida par rapport aux autres eucaryotes. Si MATH est vraie, la fréquence de transfert de gènes (EGT) depuis les chlamydia doit être plus conséquente chez les Archaeplastida par rapport aux autres groupes bactériens (ERGT), ainsi le nombre de LGT sélectionnés par le pipeline chlamydien serait supérieur à ceux sélectionnés par les pipelines contrôles. Dans la même idée, le signal chlamydien global, issu de la concaténation des gènes identifiés devrait montrer un profil congruent (EGT), à l'inverse des signaux bactériens qui eux traduiraient la multiplicité des sources de transferts (ERGT). Dans tous les cas, une réorientation du pipeline sur l'identification des LGT d'intérêt est nécessaire.

En ce qui concerne le côté "accepteur" de ces contrôles, nous avons choisi de réorienter le pipeline sur l'identification des transferts de gènes entre les Chlamydia et les Amoebozoa puis entre les Chlamydia et les Fungi. Au vu de leur ancienneté, et de leur prédilection pour les transferts de gènes, les amibes jouent ici le rôle de contrôle positif et sont susceptibles d'avoir intégré un nombre important de gènes chlamydiens dans leur génome. A l'inverse, les Fungi, connus pour être plus rétifs aux transferts de gènes, correspondent au contrôle négatif. Étant donné qu'il s'agit ici de l'identification de transfert de gènes chlamydiens vers différents groupes eucaryotes, la réorientation du pipeline porte sur les toutes premières étapes du protocole, et notamment sur le filtre de sélection appliqué sur les groupes orthologues.
Concernant les champignons, le dataset de 57 eucaryotes utilisé pour créer les groupes orthologues par OrthoFinder contient 11 espèces. Nous considérons que ce nombre est suffisant pour comparer leur signal à celui des Archaeplastida, puisque nous comptons 12 espèces de ces derniers. A partir des 210451 groupes orthologues initiaux, classify-mcl.pl en retient 5808 ayant au moins 2 espèces de Fungi. Après enrichissement avec les cyanobactéries et les Chlamydia, seuls 625 groupes orthologues sont sélectionnés sur le critère de la présence d'au moins un Chlamydia. Ces clusters continuent ensuite dans le protocole comme précédemment décrit (Figure 7).
Pour les Amoebozoa, par contre, seules 3 espèces sont présentes dans le dataset initial ayant servi pour la création des groupes orthologues. Avant donc de filtrer les clusters sur la présence de ces espèces cibles, nous les enrichissons avec 10 génomes - protéomes supplémentaires d'amibes. 5734 clusters sont alors sélectionnés sur présence d'au moins deux Amoebozoa puis 1558 après enrichissement en Chlamydia et cyanobactéries pour poursuivre le protocole. La sélection des génomes et protéomes d'Amoebozoa, ainsi que le pipeline correspondant, a été réalisée par Clotilde Garrido. Ainsi, 1 seul arbre a été identifié par le pipeline comme ayant une interaction phylogénétique entre au moins 2 Chlamydia et 3 Fungi, et 16 ont été sélectionnés pour le pipeline ré-orienté vers les amibes (Figure 12).

En ce qui concerne le côté “donneur” de ces contrôles, nous nous sommes basés sur les résultats de Dagan et al., 2013, afin de choisir les groupes bactériens vers lesquels réorienter le pipeline. En effet, ces auteurs ont estimé les contributions en transferts de gènes de chaque groupe bactérien chez les Archaeplastida. Selon leurs résultats, la contribution chlamydienne n'arriverait qu’en 6ème position, après les cyanobactéries, les Proteobacteria, les Firmicutes, les Actinobacteria et les Bacteroidetes. De ce fait, outre les cyanobactéries, nous avons choisi d’inclure tous ces groupes dans nos analyses, afin de les comparer au signal chlamydien chez les Archaeplastida. Le signal cyanobactérien, majoritaire dans l’évolution des Archaeplastida, servira quant à lui de référentiel de comparaison et d’évaluation de la sensibilité de notre pipeline.

La réorientation du pipeline s’effectue donc sur chaque groupe choisi et nécessite avant toute chose d’ajuster la sélection des espèces entrant dans le protocole bioinformatique (Figure 7). Pour chaque groupe bactérien identifié, nous avons eu recours à TQMD (ToRQuEMaDA: tool for retrieving queried Eubacteria, metadata and dereplicating assemblies Léonard et al., 2021) pour produire une liste d’espèces cibles, représentative de la diversité du groupe. Le critère principal de ces sélections par TQMD était de produire des listes d’espèces similaires à celle de Chlamydia en nombre (la structure phylogénétique de chaque sélection étant beaucoup plus difficile à contrôler). En effet, en vue de comparer les différents résultats obtenus, il est important d’homogénéiser les données et méthodes. Quatre listes d’organismes ont donc été produites, contenant respectivement 36 Proteobacteria, 37 Bacteroidetes, 20 Actinobacteria et 22 Firmicutes. L’entrée des génomes et protéomes associés à ces listes dans le protocole va donc dépendre du signal bactérien que l’on veut quantifier (Figure 12, annexe 9. De plus, la liste des bactéries générales, celle sans distinction de taxa, est elle aussi affectée par l’orientation du pipeline. En effet, cette liste de 49+92 organismes, entrant dans la deuxième phase d’enrichissement, représente la diversité complète des bactéries, et contient donc différents organismes cibles selon l’orientation. C’est pourquoi en fonction du signal étudié, les espèces cibles présentes en sont retirées.

Le jeu de groupes orthologues de départ reste ici le même que précédemment, ainsi que le premier filtre sur la présence d’au moins 2 Viridiplantae et/ou Rhodophyta. La réorientation du pipeline intervient par la suite. Au lieu d’enrichir avec les Chlamydia, puis de sélectionner les groupes orthologues sur un critère de présence de ces pathogènes, chaque pipeline est enrichi avec les protéomes des organismes cibles pour lesquels on cherche à évaluer le signal, puis sont sélectionnés uniquement les groupes orthologues présentant au minimum 1 espèce cible. Ainsi, 1143 clusters continuent le protocole des Bacteroidetes, 1147 pour les Proteobacteria, 1002 pour les Firmicutes et 1024 pour les Actinobacteria et suivent les mêmes étapes du pipeline automatique développé pour les Chlamydia.

En ce qui concerne les cyanobactéries, puisque le dataset de 48 protéomes est déjà présent dans tous les pipelines contrôles, il a suffi de réorienter le pipeline initial au moment de la sélection des clusters. En effet, en reprenant le protocole décrit pour le pipeline chlamydien,

après enrichissement en cyanobactéries et en Chlamydia, pour garder la mise en compétition du signal bactérien cible, le filtre de sélection des groupes orthologues se fait sur la présence d'au moins une cyanobactérie et non pas sur la présence des Chlamydia. 1572 OG continuent alors le pipeline.

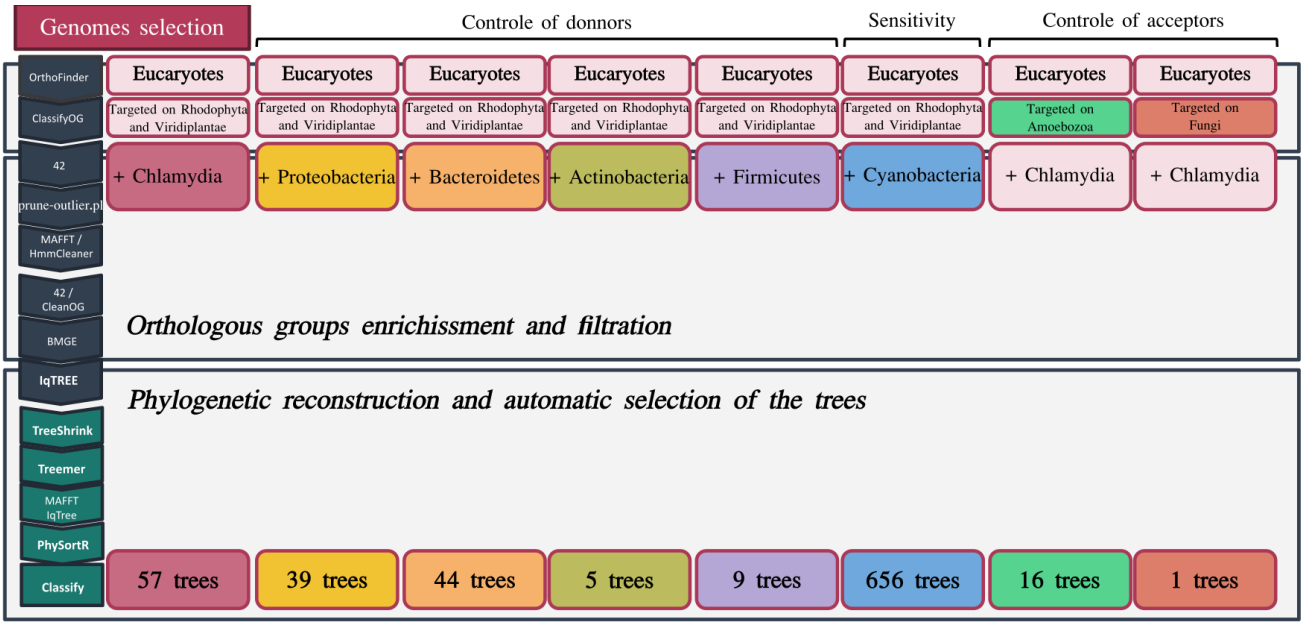


Figure 12: Organigramme de la réorientation du pipeline sur l'identification des transferts latéraux de gènes impliquant des groupes contrôles. Le protocole méthodologique du pipeline automatique, détaillé en Figure 7, visible sur la gauche du schéma, est inchangé. Les modifications apportées apparaissent en couleurs vives. Deux types de contrôles sont réalisés : le contrôle des donneurs à l'origine des LGT et celui des accepteurs. En ce qui concerne les contrôles donneurs, la réorientation du pipeline s'effectue sur la première étape d'enrichissement des groupes orthologues, afin d'identifier les LGT entre Archaeplastida et différents groupes bactériens. En rose : Chlamydia, en jaune : Proteobacteria, en orange : Bacteroidetes, en vert : Actinobacteria et en violet : Firmicutes. L'identification des LGT entre Cyanobacteria et Archaeplastida (en bleu) sert de référentiel de comparaison et d'évaluation de la sensibilité du protocole. Du côté des contrôles accepteurs, le réorientation du pipeline permet d'identifier les LGT entres les Chlamydia et deux groupes cibles, et s'est effectuée sur la sélection des groupes orthologues intégrant le protocole. Pour les contrôles bactériens, en effet, celui-ci est ciblé sur la présence d'au moins 2 Vridiplantae et / ou Rhodophyta, alors que, pour les accepteurs, le tri est ciblé sur la présence d'au moins 2 Amoebozoa (en vert vif) ou 2 Fungi (en rouge).

La suite du protocole reste ici aussi inchangée. Pour tous les pipelines testés, les arbres sont sélectionnés si un minimum de 2 donneurs et 3 accepteurs sont présents ensemble dans un même clan. Par conséquent, la réorientation du pipeline sur chaque groupe bactérien sélectionne 44 arbres ayant une relation phylogénétique entre 2 Bacteroidetes et 3 Archaeplastida, 39 pour les Proteobacteria, 9 pour les Firmicutes, 5 pour les Actinobacteria et 656 pour les cyanobactéries (Figure 12). Cette dernière proportion est attendue, puisque l'ancêtre du plaste était une cyanobactérie. L'estimation des gènes cyanobactériens chez les Archaeplastida varie entre 600 et 5000 en fonction des études. Nos méthodes sont donc, comme attendu, dans le bas de la fourchette de sensibilité décrite dans les précédentes

publications. Toutefois, puisque le but est de comparer différents signaux à stringence identique, nous acceptons cette perte de signal identifié par nos méthodes. De manière générale, nous pouvons constater un nombre de transferts de gènes supérieur pour les Chlamydia, avec 57 arbres sélectionnés, en comparaison des autres groupes bactériens. Par rapport aux Bacteroidetes, deuxième groupe comptant le plus grand nombre de transferts de gènes vers les Archaeplastida, les Chlamydia ont contribué pour environ 23% de plus à ce panachage génétique. Selon nos méthodes, les Chlamydia représentent donc le premier contributeur de transferts de gènes, après les cyanobactéries, chez les Archaeplastida. Ce premier résultat permet de confirmer l'impact relatif indéniable des Chlamydia sur l'évolution des lignées photosynthétiques, mais il n'est pas suffisant pour valider ou réfuter l'hypothèse du Ménage à Trois en tant que telle.

La concaténation des gènes sélectionnés par les pipelines contrôles, autant par création d'une supermatrice que d'un super arbre, révèle un profil de congruence similaire pour chacun (figures 12 à 16). Le pipeline contrôle Fungi n'ayant mené qu'à sélectionner un seul arbre, la concaténation n'a pas été réalisée pour ce dernier. Si l'on se concentre sur l'étude spécifique du signal chlamydien chez les eucaryotes, la concaténation de la sélection amibes montre une séparation franche entre les eucaryotes et les bactéries, mais semble être moins bien supportée que pour les Archaeplastida (figures 12 et 13). En effet, bien que le regroupement des Amoebozoa ensemble montre un signal monophylétique, les valeurs de bootstrap soutiennent moins bien la relation phylogénétique, autant entre les deux groupes taxonomiques qu'au sein même de la diversité chlamydienne, que lorsqu'on analyse le signal chlamydien chez les Archaeplastida. Ceci, combiné à la différence en terme de nombre de gènes transférés, témoigne donc plutôt en faveur d'une relation privilégiée des Chlamydia avec les Archaeplastida.
Du côté des contrôles bactériens, en ne prenant que les organismes cibles présents dans les clans identifiés pour chaque pipeline, la reconstruction phylogénétique de chaque sélection bactérienne, que ce soit en LG4X, en C20 ou en C60, montre un regroupement des Archaeplastida, respectant de plus la topologie connue de l'évolution de ces organismes. Cet ensemble monophylétique est souvent branché avec les représentants les plus basaux des groupes bactériens évalués. Pour chaque pipeline contrôle, les différentes concaténations testées (figures 15 à 16 + annexes 2 à 4), présentent des profils similaires. Les quelques différences notables résident dans la topologie interne des Archaeplastida, pour lesquels ce sont les Rhodophyta ou les Glaucophyta qui divergent en premier, selon les arbres considérés. Cependant, cette instabilité est déjà connue, l'ordre de diversification de trois lignées primaires faisant encore débat. Au sein des différentes sélections, le signal est donc congruent et ne permet pas de différencier une contribution privilégiée des Chlamydia chez les eucaryotes par rapport aux autres groupes bactériens testés.

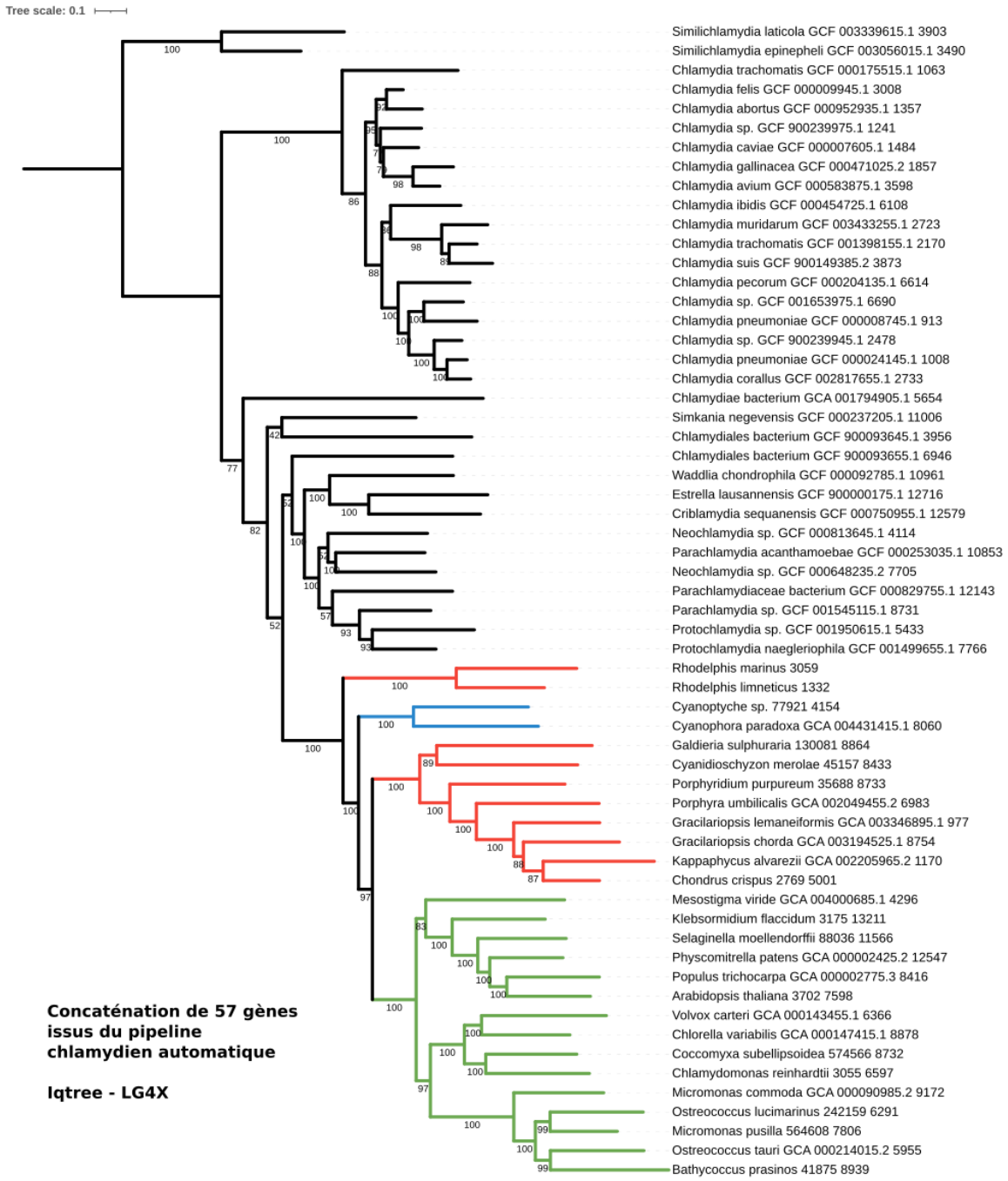


Figure 13: Arbre phylogénétique issu de la concaténation des 57 gènes sélectionnés par le pipeline automatique chlamydien. L'arbre a été obtenu par IQ-TREE, sous un modèle LG4X, après concaténation des séquences chlamydiennes et archaeplastidiennes des gènes sélectionnés par le pipeline. Les valeurs de bootstrap sont indiquées sous les branches. L'enracinement est manuel sur les simili-Chlamydia. L'échelle compte en nombre de substitutions des acides aminés par site. En rouge : Rhodophyta et Rhodelphea, en vert : Viridiplantae et en bleu : Glaucophyta.

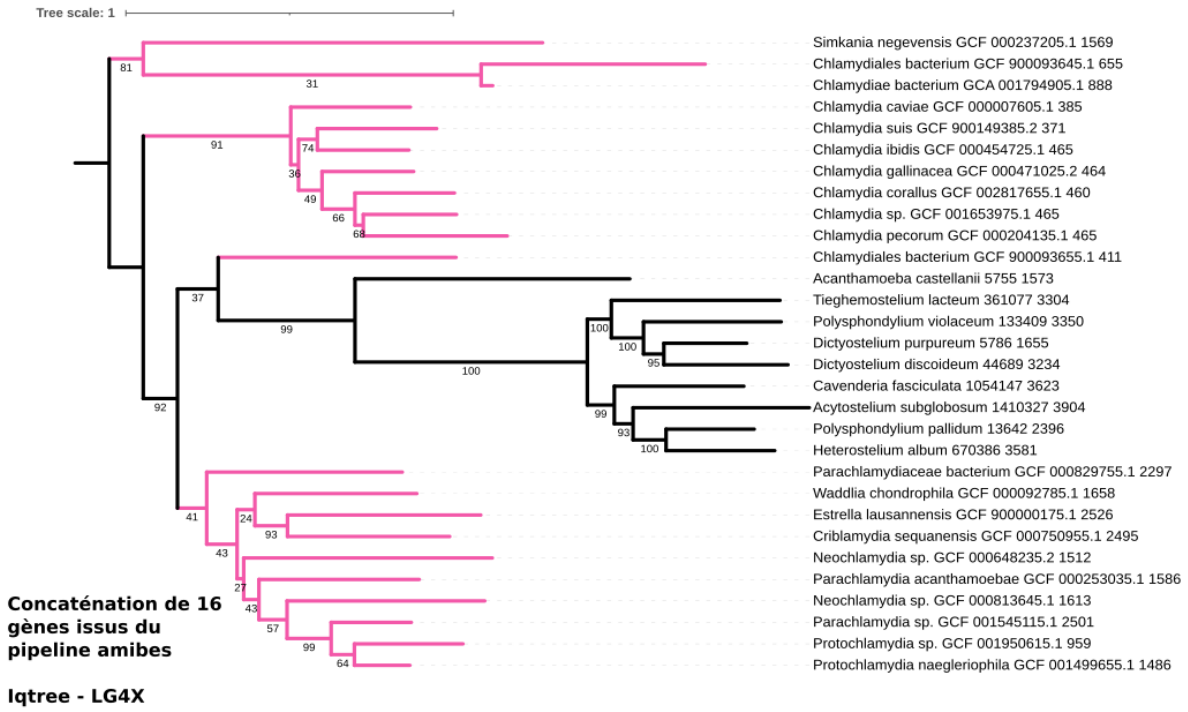


Figure 14: Arbre phylogénétique issu de la concaténation des 16 gènes sélectionnés par le pipeline automatique amibes. L'arbre a été obtenu par IQ-TREE, sous un modèle LG4X, après concaténation des séquences de Chlamydia et d'Amoebozoa des gènes sélectionnés par le pipeline. Les valeurs de bootstrap sont indiquées sur les branches. L'enracinement est manuel sur les Chlamydia basales. L'échelle compte en nombre de substitutions des acides aminés par site. En rose : les Chlamydia, en noir : les amoboezoa.

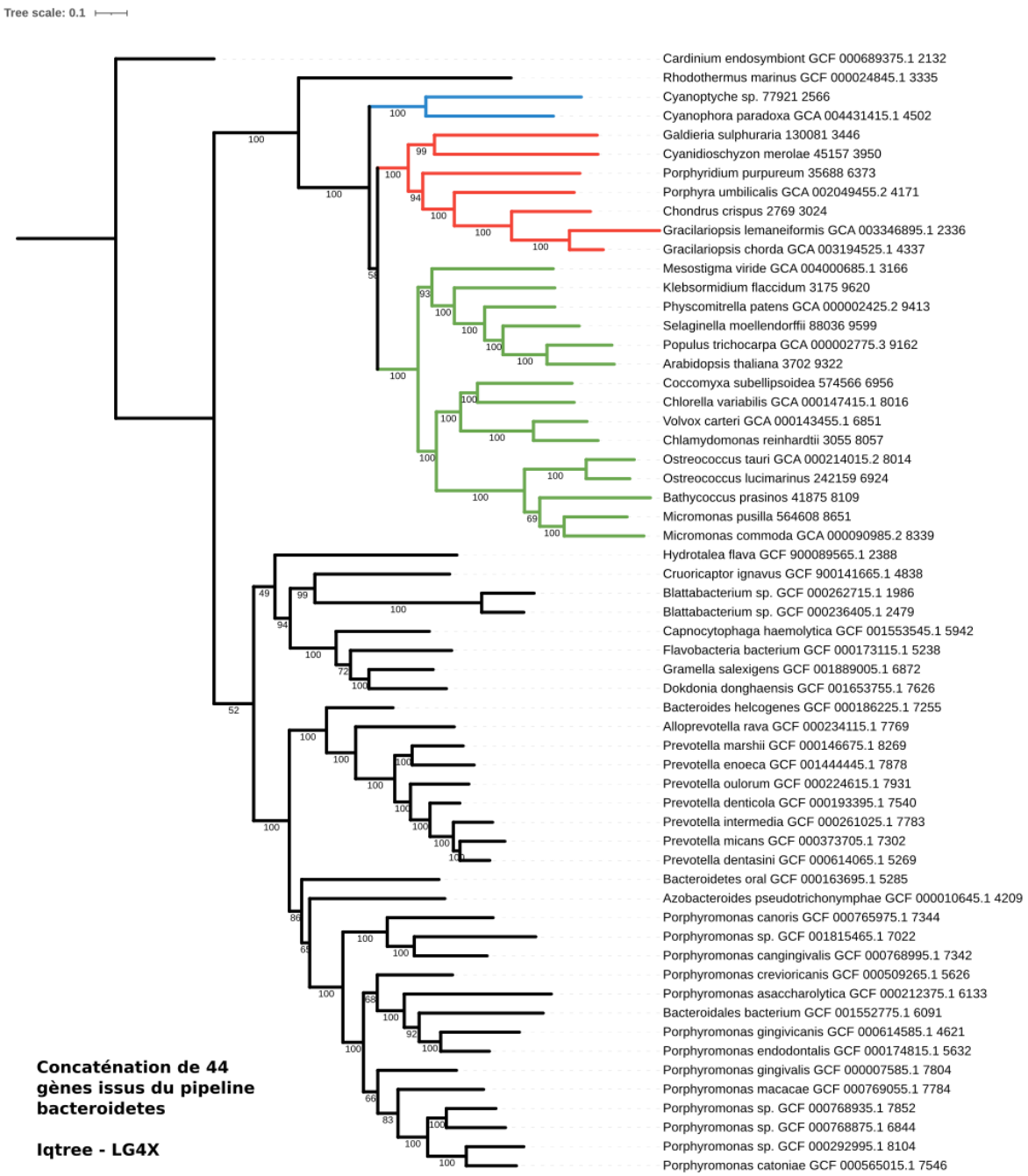


Figure 15: Arbre phylogénétique issu de la concaténation des 44 gènes sélectionnés par le pipeline automatique Bacteroidetes. L'arbre a été obtenu par IQ-TREE, sous un modèle LG4X, après concaténation des séquences de Bacteroidetes et d'Archaeplastida des gènes sélectionnés par le pipeline. Les valeurs de bootstrap sont indiquées sur les branches. L'enracinement est manuel sur les Bacteroidetes basales. L'échelle compte en nombre de substitutions des acides aminés par site. En rouge : Rhodophyta et Rhodelphea, en vert : Viridiplantae et en bleu : Glaucophyta.

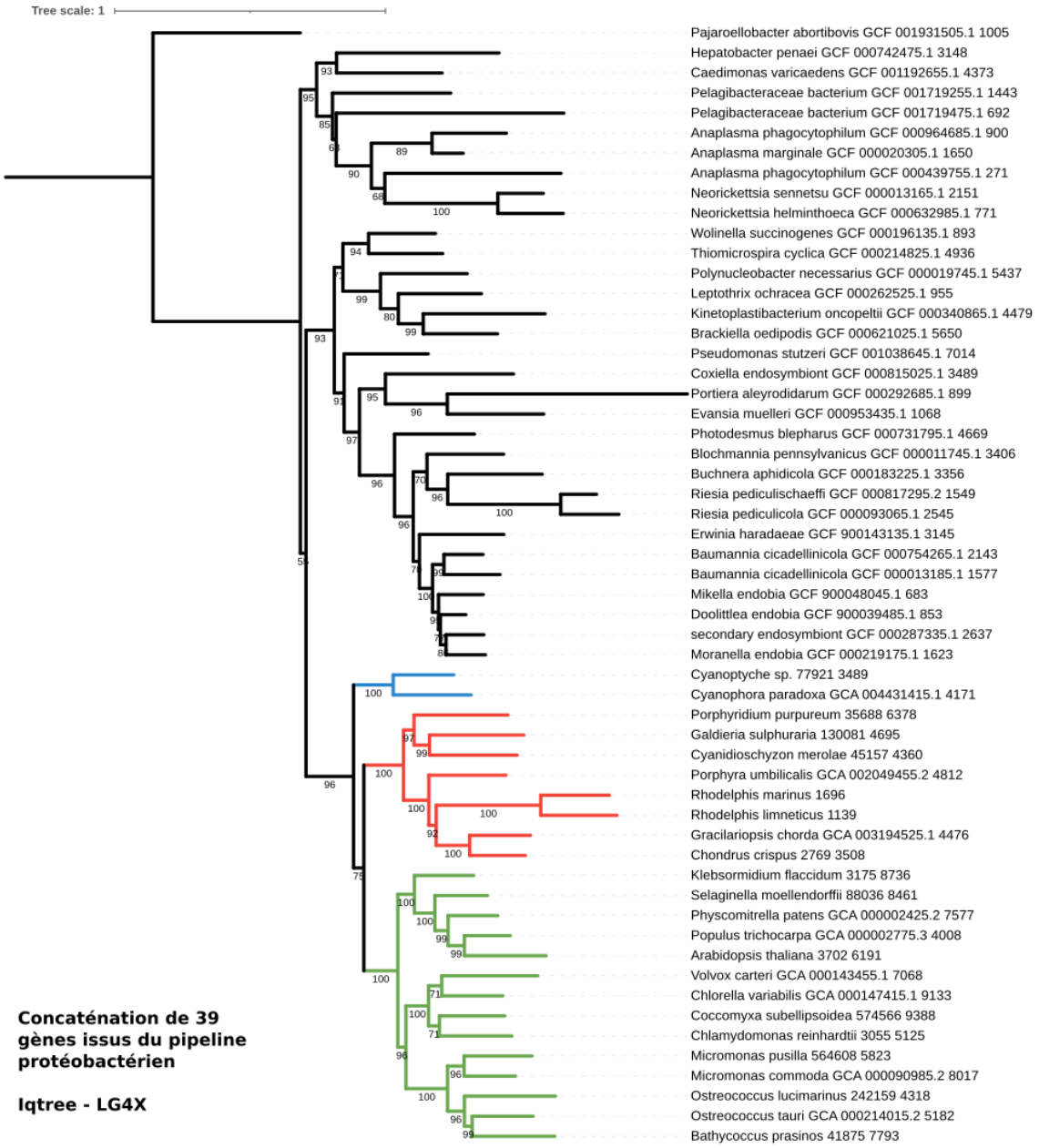


Figure 16: Arbre phylogénétique issu de la concaténation des 39 gènes sélectionnés par le pipeline automatique Proteobacteria. L'arbre a été obtenu par IQ-TREE, sous un modèle LG4X, après concaténation des séquences de Proteobacteria et d'Archaeplastida des gènes sélectionnés par le pipeline. Les valeurs de bootstrap sont indiquées sur les branches. L'enracinement est manuel sur les Proteobacteria basales. L'échelle compte en nombre de substitutions des acides aminés par site. En rouge : Rhodophyta et Rhodelphea, en vert : Viridiplantae et en bleu : Glaucophyta.

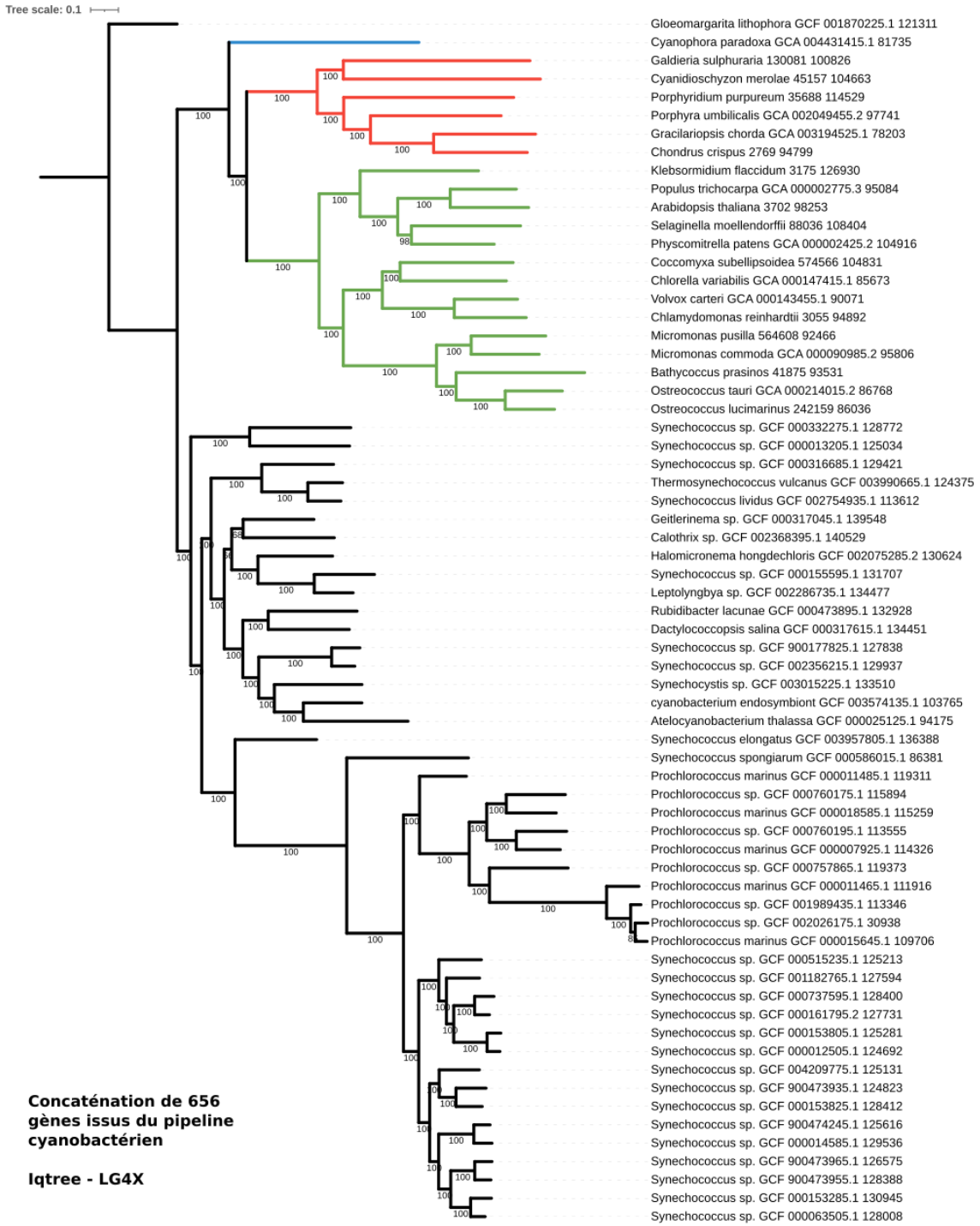


Figure 17: Arbre phylogénétique issu de la concaténation des 656 gènes sélectionnés par le pipeline automatique Cyanobacteria. L'arbre a été obtenu par IQ-TREE, sous un modèle LG4X, après concaténation des séquences de Cyanobacteria et d'Archaeplastida des gènes sélectionnés par le pipeline. Les valeurs de bootstrap sont indiquées sur les branches. L'enracinement est manuel sur les cyanobactéries basales. L'échelle compte en nombre de substitutions des acides aminés par site. En rouge : Rhodophyta et Rhodelphea, en vert : Viridiplantae et en bleu : Glaucophyta.

c. Tropismes et diversité

Jusqu'à présent nous avons identifié la nature phylogénétique de plusieurs signaux bactériens chez les Archaeplastida. Pour chacun, il existe bien une contribution bactérienne chez ces eucaryotes photosynthétiques, pour lesquelles le signal apparaît congruent. Les phylogénies identifiées par les pipelines, bien que sélectionnés de façon à mimer une analyse manuelle des arbres, peuvent témoigner de transferts disparates au cours de l'évolution des Archaeplastida et non pas d'une origine évolutive commune, comme prédite par MATH. Pour aller plus loin dans l'analyse, et définir le signal dans ses subtilités, nous avons étudié l'ensemble des transferts de gènes identifiés en prenant plus spécifiquement en compte la diversité des Archaeplastida impactée. En effet, si MATH est vraie, si le chlamydia a spécifiquement contribué à la mise en place des flux biochimiques nécessaires à l'endosymbiose, les transferts de gène ayant lieu chez l'ancêtre commun des Archaeplastida (EGT ou ERGT) sont plus susceptibles d'être retrouvés dans une diversité plus importante des descendants de l'endosymbiose, conservés alors en majorité chez 2 ou 3 lignées d'archaeplastida. On s'attend dans ce cas, à ce que le signal chlamydien soit similaire au signal cyanobactérien.

En partant du principe que les transferts de gènes issus de l'endosymbiose primaire du plaste seraient plus généralement distribués au sein des eucaryotes photosynthétiques, impactant alors plusieurs lignées d'Archaeplastida, la contribution spécifique d'un groupe bactérien devrait alors se faire sentir dans la diversité des clans sélectionnés. En effet, le transfert de gènes chez l'ancêtre commun des Archaeplastida suppose une co-évolution par la suite des organismes donneurs et accepteurs et une distribution du gène transféré plus importante dans la diversité. Au contraire, un transfert de gène plus tardif dans l'évolution sera caractérisé par une distribution moins importante parmi les donneurs et accepteurs, en nombre d'espèces mais aussi en nombre de lignées différentes impactées. L'étude de la diversité au sein des clans identifiés permet donc en partie de distinguer les LGT établis chez l'ancêtre commun des Archaeplastida, et donc potentiellement liés à l'endosymbiose primaire du plaste, des LGT plus tardifs.

Nous avons répertorié les différents profils de distribution des gènes transférés pour chaque pipeline au sein des Archaeplastida selon deux types d'analyses: i) l'analyse des tropismes topologiques des arbres sélectionnés, c'est à dire le nombre de lignées d'Archaeplastida réceptrices du LGT (Figure 18A et B), et ii) une analyse plus détaillée des lignées impliquées dans les transferts (Figure 18 C et D). Parmi les 57 arbres sélectionnés par le pipeline chlamydien, 16 montrent un branchement avec une seule lignée d'Archaeplastida (28%), 17 avec 2 lignées (30%) et 24 avec 3 (42%). La majorité des transferts de gènes chlamydiens (72%) impactent donc au moins 2 lignées d'Archaeplastida. Sans prendre en compte le signal actinobactérien ni le signal Firmicutes, puisque seulement 5 et 9 LGT y ont été identifiés au

total, cette proportion de transferts multi-lignées est supérieure à celle observée pour les autres groupes bactériens contrôles, et plutôt similaire à celle des cyanobactéries. En effet, le profil Bacteroidetes est inverse au profil Chlamydia, avec une présence de 18 LGT (41%) chez 1 seule lignée d'Archaeplastida, 16 (36%) chez 2 lignées et 10 (23%) chez les 3 lignées (Figure 18A-B). Cette prévalence pour les transferts n'impactant qu'une seule lignée d'Archaeplastida est aussi visible pour les Proteobacteria puisque 17 LGT uniques (44%) sont identifiés, contre 6 LGT (15%) chez 2 lignées, et 16 (41%) chez 3 lignées. La prédominance des transferts vers une seule lignée d'Archaeplastida, pour les Bacteroidetes et Proteobacteria notamment, peut traduire un phénomène de transfert latéral de gène non concerté et/ou plus tardif par rapport à l'émergence de ces organismes. De ce fait, nous ne pouvons affirmer leur potentielle origine endosymbiotique dans ces cas-là. Cependant, les proportions des LGT identifiés chez 2 ou 3 lignées d'Archaeplastida, respectivement de 72%, 59% et 56% pour les Chlamydia, Bacteroidetes et Proteobacteria, ne sont pas à négliger puisqu'ils mettent en évidence une ancienneté du transfert, ayant eu lieu chez l'ancêtre commun de ses organismes.

Pour aller plus en détail dans l'analyse de la diversité impactée par les LGT identifiés, nous avons répertorié les arbres sélectionnés pour chaque pipeline en fonction des lignées d'Archaeplastida réceptrices du transfert (Figure 18C-D). Ainsi, nous pouvons noter des profils différents de distribution dans la diversité des Archaeplastida en fonction du groupe bactérien à l'origine du transfert. Dans un premier temps, nous pouvons constater que, là encore, les profils Chlamydia et Cyanobacteria sont assez similaires, avec une proportion majoritaire de transferts vers les trois lignées d'Archaeplastida (42 et 29% respectivement), suivie par les transferts conjoints vers les Rhodophyta et Viridiplantae (16 et 28%), puis uniquement vers les Viridiplantae (19 et 16%) (Figure 18D). En revanche, les Bacteroidetes et Proteobacteria présentent une proportion de transferts de gène uniquement vers les Viridiplantae presque doublée par rapport à celle des deux groupes précédents, avec respectivement 36 et 38% des LGT identifiés. Les Bacteroidetes et Proteobacteria semblent donc avoir influencé davantage l'évolution des Viridiplantae, en comparaison des Chlamydia et cyanobactéries. La proportion est quant à elle inversée en ce qui concerne l'impact des différentes bactéries sur les Rhodophyta (Figure 18D), puisque ce sont les Chlamydia et Cyanobacteria qui présentent cette fois ci une proportion des LGT spécifiquement aux Rhodophyta supérieure par rapport aux Bacteroidetes et Proteobacteria (respectivement 9, 8, 5 et 5% des LGT identifiés).

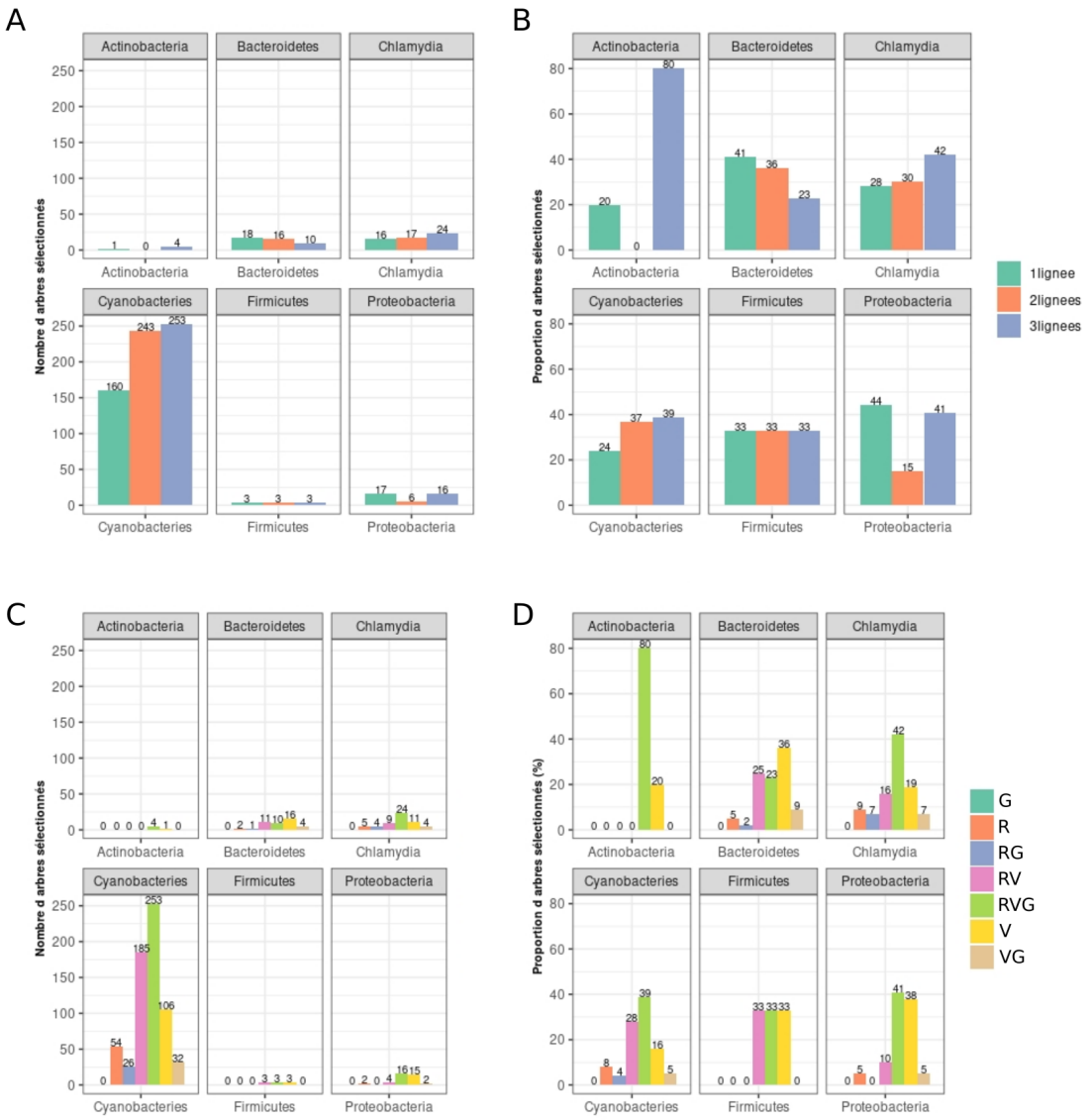


Figure 18: Analyse des tropismes de chaque sélection. Les panels A et B comptabilisent la quantité d'arbres qui branchent 1, 2 et 3 lignées d'Archaeplastida pour chaque pipeline en nombre (A) et en proportion (B). Les panels C et D détaillent davantage cette diversité en dénombrant les arbres en fonction des lignées d'Archaeplastida impliquées, en nombre (C) et en proportion (D). G = Glaucophyta; R = Rhodophyta, V = Viridiplantae.

Jusqu'à présent, nous avons évalué les différents tropismes phylogénétiques des LGT identifiés de façon binaire (présence ou absence d'une lignée d'Archaeplastida), sans prendre réellement en compte la diversité totale au sein des clans. Rappelons que les critères de sélection du pipeline automatique identifient chaque arbre pour lequel il existe une interaction phylogénétique entre au moins deux donneurs et trois accepteurs. Cependant, la diversité des organismes joue un rôle dans l'interprétation des LGT et notamment dans leur temporalité.

Nous avons donc analysé les proportions de donneurs et accepteurs au sein des clans sélectionnés pour les différents pipelines. Cyanobactéries mises à part, aucune différence

significative n'est observable entre les différents pipelines, que ce soit au niveau les donneurs (Figure 19B et E) ou accepteurs (Figure 19C et F), mais aussi en nombre d'espèces total incluses dans les clans (Figure 19A et D). Les distributions sont similaires, les groupes orthologues branchent donc de manière générale le même nombre de donneurs et d'accepteurs. Indirectement, ces résultats montrent aussi la robustesse du pipeline, puisque pour chaque condition contrôle testée, la diversité des clans est équivalente. Pris ensemble, les accepteurs montrent un profil de distribution similaire pour les différents pipelines. Cependant, l'évaluation séparée des lignées d'Archaeplastida révèle certaines différences (Figure 19 G à K). Cette particularité est vérifiée en analysant la distribution du nombre de Viridiplantae au sein des clans sélectionnés. En effet, on observe une proportion plus importante de Viridiplantae au sein de la sélection Bacteroidetes, tous tropismes confondus, en comparaison aux autres signaux étudiés (Figure 19G). Selon un test statistique de Kruskal-Wallis, ce signal est significativement plus important chez les Bacteroidetes par rapport aux autres groupes bactériens (Figure 19G). Cette particularité de l'évolution conjointe des Bacteroidetes et Viridiplantae peut être le signe de transferts de gènes plus continus au cours de l'évolution de ces organismes. De la même manière, nous avons aussi analysé la présence d'intrus dans les clans sélectionnés. Ici, nous appelons "intrus" les espèces non-cibles, que ce soit des eucaryotes ou des bactéries, faisant partie du sous arbre identifié par le pipeline, interrompant donc la monophylie des espèces cibles dans le clan. Les eucaryotes photosynthétiques présents issus de l'endosymbiose secondaire d'une algue rouge ou d'une algue verte ne sont pas considérés comme des intrus ici. Ces intrus peuvent révéler un partage de gène ancien, remontant à l'ancêtre commun de tous ces organismes et suivi de pertes multiples (et donc dans ce cas potentiellement invalider l'hypothèse du transfert endosymbiotique), un transfert plus tardif ou une contamination, mais aussi un artéfact phylogénétique. De ce fait, là encore, la diversité des organismes dans les clans, et particulièrement ici celle des intrus, peut aider à la différenciation des signaux liés à l'endosymbiose primaire du plaste, pour lesquels on s'attend à un faible nombre d'intrus, de ceux qui ne le sont pas. L'analyse du nombre d'espèces non cibles dans les clans, en fonction de chaque pipeline étudié, met en évidence une proportion plus importante d'intrus dans les clans sélectionnés pour le pipeline Bacteroidetes (Figure 19J), significativement différente des pipelines Chlamydia, Cyanobacteria et Proteobacteria. Ceci, tout comme la prévalence des Viridiplantae comme accepteurs, suggère un profil différent pour les Bacteroidetes.

En évaluant les distributions de donneurs et accepteurs séparément, il n'apparaît pas de différences, hormis pour les Viridiplantae et les intrus. Cependant, lorsqu'on examine conjointement les accepteurs et les donneurs de ces transferts de gènes dans les clans, des profils différents émergent en fonction des pipelines. Là encore, si on retire les résultats obtenus pour les pipelines Firmicutes et Actinobactéries, puisque le nombre de transferts de gènes identifiés est nettement inférieur aux autres groupes bactériens, nous pouvons distinguer deux profils différents, regroupant les Proteobacteria et les Bacteroidetes d'un

côté, et les Chlamydia et les cyanobactéries de l'autre. En effet, la Figure 20 visualise en partie le lien existant entre les différentes espèces cibles de chaque clan sélectionné par les pipelines. De manière générale, nous pouvons observer une corrélation entre les nombres de donneurs et d'accepteurs des transferts de gènes. Pour les pipelines chamydien et cyanobactérien, le nombre d'espèces d'Archaeplastida semble augmenter en même temps que le nombre de bactéries cibles dans le clan. A l'inverse, il semblerait que, pour les Bacteroidetes et les Proteobacteria, l'augmentation du nombre de donneurs dans le clan soit en lien avec la diminution du nombre d'eucaryotes photosynthétiques cibles. Pour résumer, un transfert de gène chez les Proteobacteria et Bacteroidetes concerne soit une diversité plus importante de bactéries, soit une diversité plus importante d'Archaeplastida, alors que pour les Chlamydia et cyanobactéries, ces transferts de gènes affectent plus intensément à la fois la diversité des bactéries et des Archaeplastida. Cette particularité pourrait témoigner d'une histoire évolutive différente en fonction des groupes bactériens étudiés, plutôt basée sur des transferts latéraux de gènes dispersés au cours de l'évolution d'un côté, et issus d'évènements endosymbiotiques de l'autre. De plus, une plus grande diversité des espèces cibles, notamment des Archaeplastida, au sein des clans identifiés implique nécessairement un transfert du gène chez l'ancêtre commun de ces organismes, remontant potentiellement à l'endosymbiose primaire du plaste. En fonction du tropisme de ces transferts et de la diversité des organismes impactés, nous pouvons imaginer quatre scénarios différents, en combinant deux dimensions, la profondeur et la densité. i) Dans le cas d'un scénario profond et dense, le gène est transféré chez le dernier ancêtre commun des plastes (LCA pour Last Common Ancestor) et conservé depuis, il sera donc visible dans plus d'une lignée d'Archaeplastida et une diversité d'accepteurs importante. ii) Pour le deuxième scénario, profond et rare, soit le gène a été transféré chez le LCA mais perdu ou remplacé depuis lors, soit le transfert a eu lieu plusieurs fois à la suite immédiate de l'émergence du LCA, dans ce cas, bien que la présence du gène soit visible dans plus d'une lignée d'Archaeplastida, la diversité représentée est moins importante. iii) Le scénario superficiel et dense représente quant à lui un gène transféré plus tardivement mais conservé complètement dans le sous-groupe d'accepteurs, et enfin iv) le dernier scénario, superficiel et rare, correspondant aux transferts multiples indépendants, sera visualisé par une faible diversité, notamment des accepteurs, au sein des clans, qui majoritairement concernent une seule lignée d'Archaeplastida.

La proportion de transferts de gènes vers une seule lignée d'Archaeplastida, leur impact privilégié sur les Viridiplantae, ainsi que les profils donneurs en fonction des accepteurs différents, peuvent témoigner de transferts latéraux de gènes dispersés au sein de l'évolution de ses organismes, sans lien direct avec l'endosymbiose primaire du plaste, particulièrement pour les Bacteroidetes. Cependant, une majorité des LGT, y compris provenant des Bacteroidetes, sont observables chez deux ou trois lignées d'Archaeplastida, impliquant dès

lors un timing plus ancien. Ainsi, bien qu'il semble y avoir des profils de diversité différents entre Chlamydia, cyanobactéries et autres groupes contrôles bactériens, cette analyse n'est pas encore suffisante pour confirmer ou réfuter MATH.

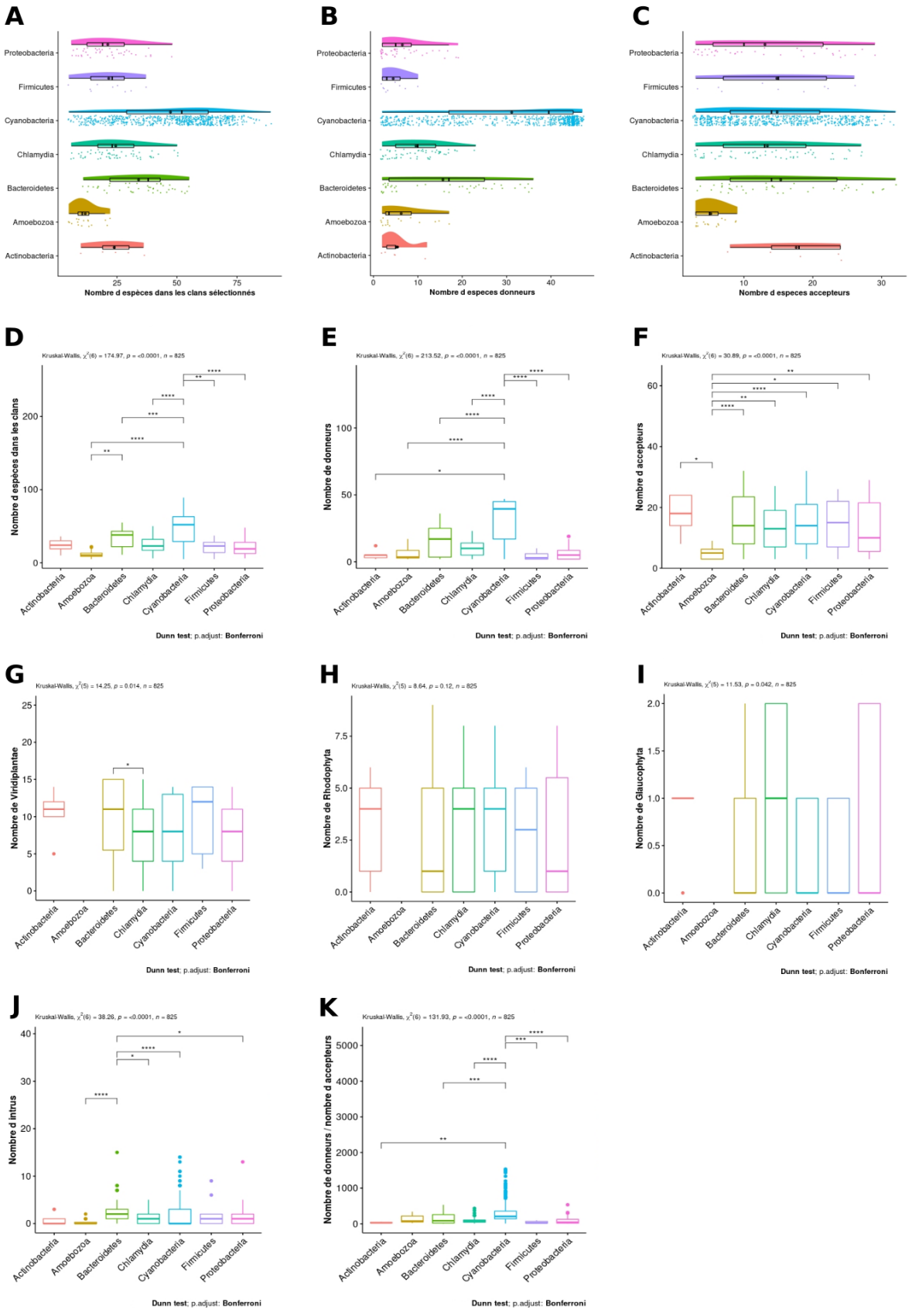


Figure 19: Analyse de la diversité des clans sélectionnés pour chaque pipeline. Pour chaque pipeline étudié, la distribution du nombre d'espèces dans les clans (A) a été analysée, puis détaillée en donneurs (B) et accepteurs (C). Un test statistique de Kruskal Wallis évalue les différences de diversité dans les pipelines. Plusieurs paramètres sont reportés pour évaluer la diversité des clans sélectionnés, à savoir le nombre d'espèces (D), le nombre de donneurs (E), et d'accepteurs (F), mais aussi la diversité en Viridiplantae (G), Rhodophyta (H), et Glaucophytes (I), mais aussi la proportion d'intrus (J). Le dernier paramètre analysé est le rapport entre donneurs et accepteurs dans les clans (K). Les paramètres spécifiques au test de Kruskal Willis pour chaque condition évaluée sont indiqués sur les graphes. Les différences significatives entre les pipelines sont représentées par des étoiles (*).

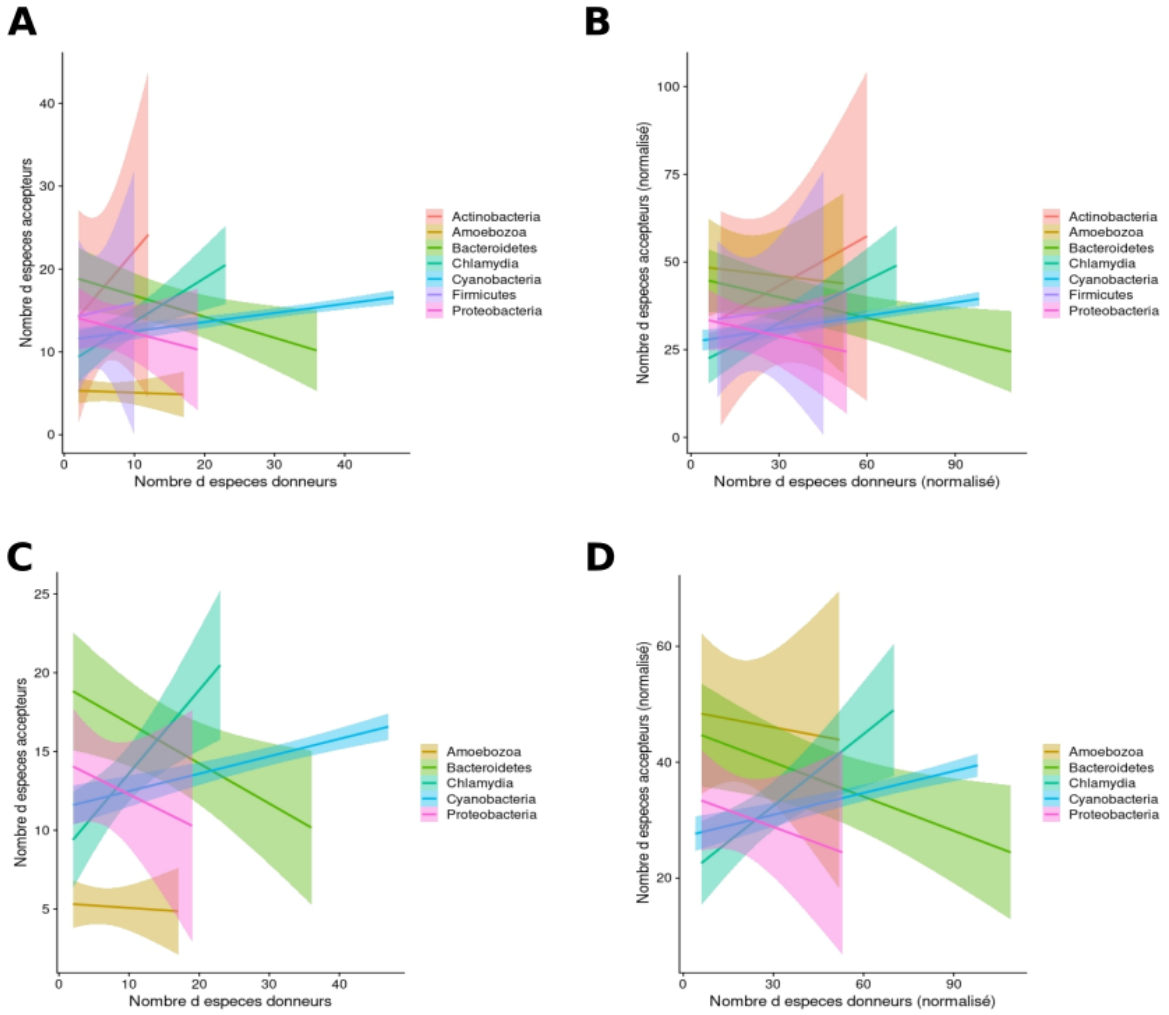


Figure 20: Analyse du nombre d'accepteurs dans les clans sélectionnés en fonction du nombre de donneurs. Pour chaque pipeline, est rapportée la diversité des accepteurs dans les clans en fonction du nombre de donneurs. En (A) sont représentées les données brutes, en nombre absolu d'espèces présentes dans les clans sélectionnés pour les donneurs et pour les accepteurs. En (B), ces données sont normalisées avec le nombre d'espèces donneurs et accepteurs entrant le pipeline. Pour les graphes (C) et (D), les pipelines Actinobacteria et Firmicutes ont été retiré, pour plus de clareté (ne contiennent que 5 et 9 clans sélectionnés respectivement).

d. Annotations fonctionnelles

Au-delà de l'analyse phylogénétique, qui permet de discerner la nature des signaux étudiés et d'éclaircir l'impact spécifique d'un groupe bactérien par rapport à un autre, une étude fonctionnelle des transferts de gènes identifiés peut apporter d'autres informations pour

évaluer la plausibilité de l'hypothèse du Ménage à Trois, notamment au sujet de la mise en place des flux biochimiques à l'origine de l'intégration métabolique de la cyanobactérie. On s'attend en effet, si le chlamydia était présent dans la vésicule d'inclusion, que le profil d'annotation des LGT identifiés laisse entrevoir des similarités avec le profil cyanobactérien (EGT), notamment de part la présence de transferts multiples au sein de mêmes voies métaboliques. De plus, selon MATH, le chlamydia fournit les pièces manquantes de la symbiose, apportant les protéines nécessaires à la connectivité du plaste. Le présence de transporteurs notamment pourrait appuyer cette hypothèse. Pour chaque sélection de chaque pipeline, nous avons donc annoté les séquences présentes dans les branchements d'intérêt grâce à EggNog Mapper et BlastKoala. Nous avons généralisé au clan entier lorsqu'au moins 50% des séquences présentes étaient annotées de la même manière (Table 1, annexe 8).

Table 1: Table récapitulative de l'annotation fonctionnelle des 57 arbres sélectionnés par le pipeline automatique chlamydien. L'annotation de chaque clan sélectionné, réalisée par eggNOG mapper et par BlastKOALA, a été généralisée si au moins 50% des séquences étaient annotées de la même manière. Le tropisme de chaque clan est indiqué en deuxième colonne. R : Rhodophyta, V : Viridiplantae, G : Glaucophyta.

Pipeline Chlamydia				
OG	**Tropisme**	**Identification protéine**	**N°Kegg**	**Fonction**
OG0000013	RVG	NLRC3, NOD3; NLR family CARD domain-containing protein 3	K22614	Signaling and cellular processes
OG0000134	R	aqpZ; aquaporin Z	K06188	**Transporters**
OG0000479	RVG	aroDE, DHQ-SDH; 3-dehydroquinate dehydratase / shikimate dehydrogenase [EC:4.2.1.10 1.1.1.25]	K13832	Amino acid metabolism
OG0000649	VG	ksgA; 16S rRNA (adenine1518-N6/adenine1519-N6)-dimethyltransferase [EC:2.1.1.182]	K02528	Ribosome biogenesis
OG0000674	V	TC.PIT; inorganic phosphate transporter, PiT family	K03306	**Transporters**
OG0000812	VG	K07146; UPF0176 protein	K07146	
OG0000904	V	rluB; 23S rRNA pseudouridine2605 synthase [EC:5.4.99.22]	K06178	Ribosome biogenesis
OG0000913	RVG	fabF, OXSM, CEM1; 3-oxoacyl-[acyl-carrier-protein] synthase II [EC:2.3.1.179]	K09458	Lipid metabolism - Metabolism of cofactors and vitamins
OG0001000	V	NSF, SEC18; vesicle-fusing ATPase [EC:3.6.4.6]	K06027	Membrane trafficking
OG0001059	V	E1.1.1.82; malate dehydrogenase (NADP+) [EC:1.1.1.82]	K00024	Carbohydrate metabolism - Energy metabolism - Amino acid metabolism
OG0001078	RV	ribBA; 3,4-dihydroxy 2-butanone 4-phosphate synthase / GTP cyclohydrolase II [EC:4.1.99.12 3.5.4.25]	K14652	Metabolism of cofactors and vitamins
OG0001293	VG	glgA; starch synthase [EC:2.4.1.21]	K00703	Carbohydrate metabolism
OG0001410	RVG	RP-L24, MRPL24, rplX; large subunit ribosomal protein L24	K02895	Ribosome
OG0001468	R	trpC; indole-3-glycerol phosphate synthase [EC:4.1.1.48]	K01609	Amino acid metabolism
OG0001851	RV	mraW, rsmH; 16S rRNA (cytosine1402-N4)-methyltransferase [EC:2.1.1.199]	K03438	Ribosome biogenesis
OG0001950	RV	trxB, TRR; thioredoxin reductase (NADPH) [EC:1.8.1.9]	K00384	Metabolism of other amino acids
OG0002167	RVG	PHYH; phytanoyl-CoA hydroxylase [EC:1.14.11.18]	K00477	Peroxisome
OG0002222	RVG	SAL; 3'(2'), 5'-bisphosphate nucleotidase / inositol polyphosphate 1-phosphatase [EC:3.1.3.7 3.1.3.57]	K15422	Carbohydrate metabolism - Energy metabolism
OG0002300	R	PTH1, pth, spoVC; peptidyl-tRNA hydrolase, PTH1 family [EC:3.1.1.29]	K01056	Translation factors
OG0002395	RVG	truA, PUS1; tRNA pseudouridine38-40 synthase [EC:5.4.99.12]	K06173	Transfer RNA biogenesis

OG0002498	RV	YARS, tyrS; tyrosyl-tRNA synthetase [EC:6.1.1.1]	K01866	Transfer RNA biogenesis
OG0002584	RV	pnp, PNPT1; polyribonucleotide nucleotidyltransferase [EC:2.7.7.8]	K00962	Messenger RNA biogenesis
OG0002591	RVG			
OG0003272	RVG	trpD; anthranilate phosphoribosyltransferase [EC:2.4.2.18]	K00766	Amino acid metabolism
OG0003309	RG	rnhB; ribonuclease HII [EC:3.1.26.4]	K03470	DNA replication proteins
OG0003312	RVG	tyrP; tyrosine-specific transport protein	K03834	**Transporters**
OG0003383	RV		K03319	**Transporters**
OG0003449	RVG	TC.AAA; ATP:ADP antiporter, AAA family	K03301	**Transporters**
OG0003873	RVG	ISA, treX; isoamylase [EC:3.2.1.68]	K01214	Carbohydrate metabolism
OG0003961	V	fabI; enoyl-[acyl-carrier protein] reductase I [EC:1.3.1.9 1.3.1.10]	K00208	Lipid metabolism - Metabolism of cofactors and vitamins
OG0004281	RV	K09858; SEC-C motif domain protein	K09858	
OG0004382	RVG			
OG0004493	RVG	Na H antiporter		**Transporters**
OG0004746	RVG	uhpC; MFS transporter, OPA family, sugar phosphate sensor protein UhpC	K07783	**Transporters**
OG0004766	RVG	ispE; 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase [EC:2.7.1.148]	K00919	Metabolism of terpenoids and polyketides
OG0004954	RVG			
OG0005053	V	gcpE, ispG; (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase [EC:1.17.7.1 1.17.7.3]	K03526	Metabolism of terpenoids and polyketides
OG0005097	RVG			
OG0005231	RVG	E2.6.1.83; LL-diaminopimelate aminotransferase [EC:2.6.1.83]	K10206	Amino acid metabolism
OG0005232	RVG	ispD; 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase [EC:2.7.7.60]	K00991	Metabolism of terpenoids and polyketides
OG0005255	RVG			
OG0005308	RV	CHS; chalcone synthase [EC:2.3.1.74]	K00660	Biosynthesis of other secondary metabolites
OG0005374	R	rlmH; 23S rRNA (pseudouridine1915-N3)-methyltransferase [EC:2.1.1.177]	K00783	Ribosome biogenesis
OG0005382	RVG		K03215	Ribosome biogenesis
OG0005581	RVG	ATS1; glycerol-3-phosphate O-acyltransferase [EC:2.3.1.15]	K00630	Lipid metabolism
OG0006000	V	kdsB; 3-deoxy-manno-octulosonate cytidylyltransferase (CMP-KDO synthetase) [EC:2.7.7.38]	K00979	Glycan biosynthesis and metabolism
OG0007168	RVG			
OG0008425	VG	UGP3; UTP---glucose-1-phosphate uridylyltransferase [EC:2.7.7.9]	K22920	Lipid metabolism
OG0008763	V			Ribosome biogenesis
OG0008957	V			
OG0008974	RV	ddl; D-alanine-D-alanine ligase [EC:6.3.2.4]	K01921	Metabolism of other amino acids - Glycan biosynthesis and metabolism
OG0009869	RG	apbE; FAD:protein FMN transferase [EC:2.7.1.180]	K03734	
OG0014617	V	dnaQ; DNA polymerase III subunit epsilon [EC:2.7.7.7]	K02342	DNA replication proteins
OG0017499	RG	wbpA; UDP-N-acetyl-D-glucosamine dehydrogenase [EC:1.1.1.136]	K13015	Carbohydrate metabolism - Glycan biosynthesis and metabolism
OG0017872	R		K01046	Glycerolipid metabolism
OG0024221	V	murB; UDP-N-acetylmuramate dehydrogenase [EC:1.3.1.98]	K00075	Carbohydrate metabolism - Glycan biosynthesis and metabolism

OG0028045	RG	queD, ptpS, PTS; 6-pyruvoyltetrahydropterin/6-carboxytetrahydropterin synthase [EC:4.2.3.12 4.1.2.50]	K01737	Metabolism of cofactors and vitamins

A première vue, les transferts de gènes identifiés impactent de multiples voies métaboliques et processus cellulaires. Cependant, nous pouvons observer une prédominance de transferts de transporteurs parmi les gènes provenant des Chlamydia. Dans un contexte endosymbiotique, la présence de transporteurs est nécessaire pour la mise en place des flux biochimiques et ainsi conduire à l'intégration métabolique. En effet, sur les 57 groupes orthologues sélectionnés par le pipeline automatique, 7 sont annotés comme étant des transporteurs, parmi lesquels nous pouvons retrouver les transporteurs clés de MATH (à savoir UhpC, un transporteur ATP et TyrP). Pour les Bacteroidetes, aucune protéine de ce type n'a été identifiée. En revanche, les Proteobacteria semblent en avoir transmis deux aux Archaeplastida. En approfondissant l'analyse de ces protéines en particulier, nous pouvons remarquer que ces deux transporteurs identifiés pour les Proteobacteria le sont aussi dans le pipeline chlamydien, ce qui apparaît incompatible. Or, l'analyse manuelle des arbres correspondants montre le branchement de deux Proteobacteria avec les Archaeplastida, puis directement avec les Chlamydia pour l'arbre correspondant à TyrP (Figure 21) et du transporteur ADP-ATP (Figure 23). L'analyse des arbres correspondants sélectionnés par le pipeline Chlamydien appuie l'identification d'un clan paralogue composé des deux Proteobacteria et Stramenopiles de l'arbre TyrP (Figure 22). En ce qui concerne le transporteur ADP-ATP, l'arbre sélectionné par le pipeline Chlamydia ne présente aucune Proteobacteria, malgré l'enrichissement avec un jeu de données bactérien composé de 27 de ces organismes (Figure 24, annexe 9). Il est fortement probable dans ces cas-ci que les gènes identifiés soient chlamydiens, transférés chez les Archaeplastida lors de l'endosymbiose primaire du plaste, et chez les deux Proteobacteria en question lors d'événement de transferts latéraux de gènes annexes. Dans un contexte endosymbiotique, la présence de transporteurs est nécessaire pour la mise en place des flux biochimiques et ainsi conduire à l'intégration métabolique. Cette proportion de transporteurs chlamydiens retrouvés chez les Archaeplastida et absente des autres groupes bactériens étudiés, appuie dès lors l'hypothèse d'une implication particulière de ces pathogènes lors de l'endosymbiose primaire du plaste.

La visualisation des transferts de gènes sélectionnés sur des cartes métaboliques permet de se rendre compte de l'impact de certains groupes sur le métabolisme des Archaeplastida. En effet, les Chlamydia semblent jouer un rôle plus prononcé dans certaines voies métaboliques, en comparaison des autres groupes bactériens, notamment de par le transfert de plusieurs gènes impactant les mêmes voies. Par exemple, la voie de biosynthèse du tryptophane, déjà étudiée par Cenci et al, pour laquelle 4 des 9 étapes sont catalysées par des enzymes d'origine chlamydienne, est aussi identifiée dans nos résultats. Cette particularité est aussi retrouvée pour les voies de synthèse des acides gras et des isoprénoïdes et semble propre aux gènes

affiliés aux Chlamydia. En ce qui concerne les autres groupes bactériens étudiés, les transferts de gènes sont dispersés au sein du métabolisme et non regroupés en voies métaboliques. Une exception existe cependant, puisque les Bacteroidetes impactent la voie de synthèse des acides gras sur deux enzymes : FabZ et FATA. Ainsi, une analyse succincte des fonctions des transferts de gènes sélectionnés par les pipelines permet de mettre en évidence un profil chlamydien différent des autres groupes bactériens, impactant plus spécifiquement les transporteurs et révèle des transferts multiples dans certaines voies métaboliques.

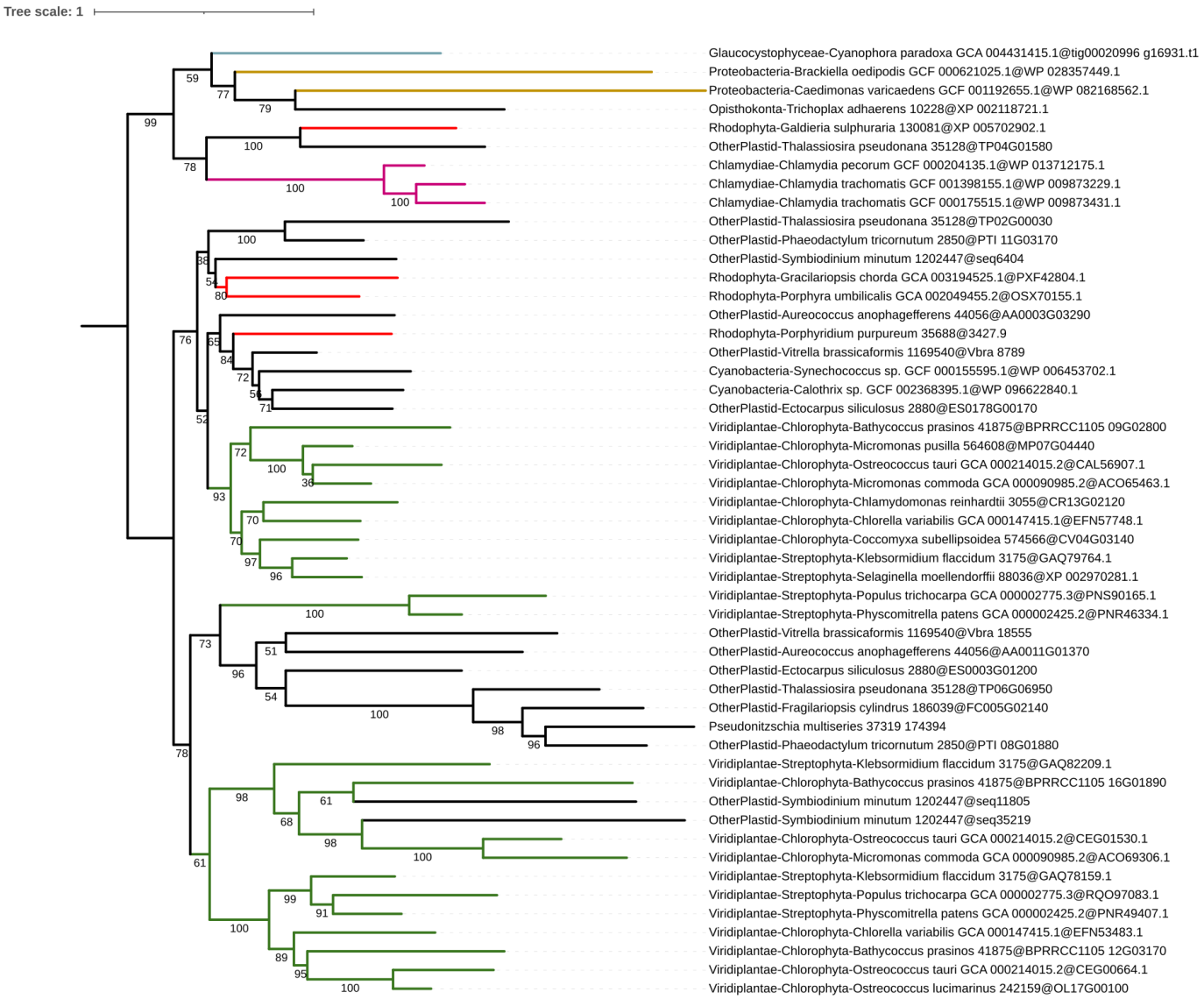


Figure 21: Arbre phylogénétique simple gène du transporteur TyrP identifié par le pipeline Proteobacteria. Cet arbre correspond à la reconstruction phylogénétique du groupe orthologue OG0003312, annoté comme ATP:ADP antiporter, sélectionné par le pipeline Proteobacteria. L'arbre a été obtenu par IQ-TREE, sous un modèle LG4X en ultrafast-bootstrap. Les espèces identifiées comme "other-plastid" ne sont pas considérées comme des intrus, et font donc partie intégrante des critères de sélection du clan. Les valeurs de bootstrap sont indiquées sur les branches. L'enracinement est en mid-point. L'échelle compte en nombre de

substitutions des acides aminés par site. En rouge : Rhodophyta et Rhodelphea, en vert : Viridiplantae et en bleu : Glaucophyta. En rose : Chlamydia, en jaune : Proteobacteria

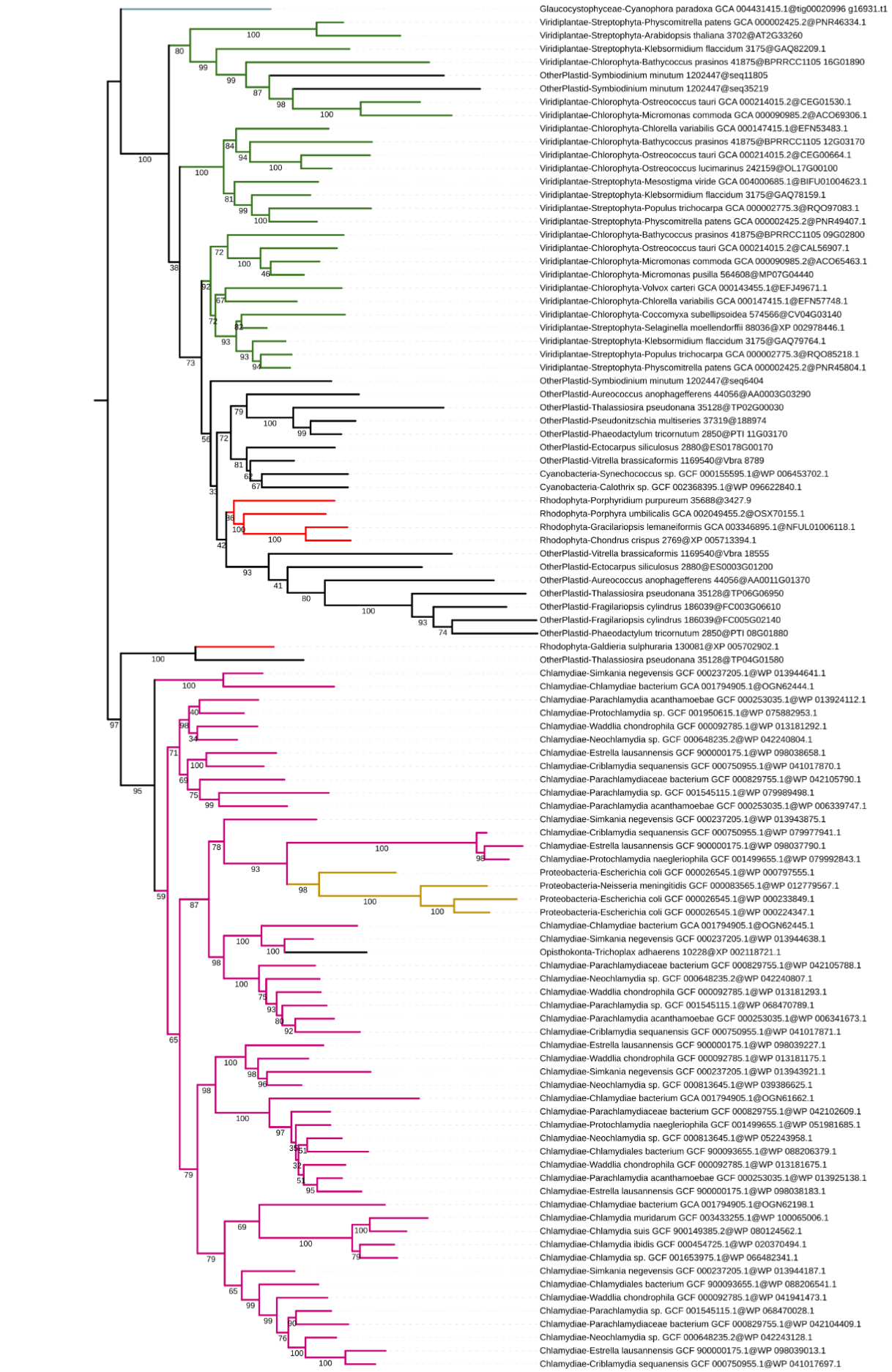

ee scale: 1
Glaucocystophyceae-Cyanophora paradoxa GCA 004431415.1@tig00020996 g16931.t1
Viridiplantae-Streptophyta-Physcomitrella patens GCA 000002425.2@PNR46334.1
Viridiplantae-Streptophyta-Arabidopsis thaliana 3702@AT2G33260
Viridiplantae-Streptophyta-Klebsormidium flaccidum 3175@GAQ82209.1
Viridiplantae-Chlorophyta-Bathycoccus prasinos 41875@BPRRCC1105 16G01890
OtherPlastid-Symbiodinium minutum 1202447@seq11805
OtherPlastid-Symbiodinium minutum 1202447@seq35219
Viridiplantae-Chlorophyta-Ostreococcus tauri GCA 000214015.2@CEG01530.1
Viridiplantae-Chlorophyta-Micromonas commoda GCA 000090985.2@ACO69306.1
Viridiplantae-Chlorophyta-Chlorella variabilis GCA 000147415.1@EFN53483.1
Viridiplantae-Chlorophyta-Bathycoccus prasinos 41875@BPRRCC1105 12G03170
Viridiplantae-Chlorophyta-Ostreococcus tauri GCA 000214015.2@CEG00664.1
Viridiplantae-Chlorophyta-Ostreococcus lucimarinus 242159@OL17G00100
Viridiplantae-Streptophyta-Mesostigma viride GCA 004000685.1@BIFU01004623.1
Viridiplantae-Streptophyta-Klebsormidium flaccidum 3175@GAQ78159.1
Viridiplantae-Streptophyta-Populus trichocarpa GCA 000002775.3@RQO97083.1
Viridiplantae-Streptophyta-Physcomitrella patens GCA 000002425.2@PNR49407.1
Viridiplantae-Chlorophyta-Bathycoccus prasinos 41875@BPRRCC1105 09G02800
Viridiplantae-Chlorophyta-Ostreococcus tauri GCA 000214015.2@CAL56907.1
Viridiplantae-Chlorophyta-Micromonas commoda GCA 000090985.2@ACO65463.1
Viridiplantae-Chlorophyta-Micromonas pusilla 564608@MP07G04440
Viridiplantae-Chlorophyta-Volvox carteri GCA 000143455.1@EFJ49671.1
Viridiplantae-Chlorophyta-Chlorella variabilis GCA 000147415.1@EFN57748.1
Viridiplantae-Chlorophyta-Coccomyxa subellipsoidea 574566@CV04G03140
Viridiplantae-Streptophyta-Selaginella moellendorffii 88036@XP 002978446.1
Viridiplantae-Streptophyta-Klebsormidium flaccidum 3175@GAQ79764.1
Viridiplantae-Streptophyta-Populus trichocarpa GCA 000002775.3@RQO85218.1
Viridiplantae-Streptophyta-Physcomitrella patens GCA 000002425.2@PNR45804.1
OtherPlastid-Symbiodinium minutum 1202447@seq6404
OtherPlastid-Aureococcus anophagefferens 44056@AA0003G03290
OtherPlastid-Thalassiosira pseudonana 35128@TP02G00030
OtherPlastid-Pseudonitzschia multiseries 37319@188974
OtherPlastid-Phaeodactylum tricornutum 2850@PTI 11G03170
OtherPlastid-Ectocarpus siliculosus 2880@ES0178G00170
OtherPlastid-Vitrella brassicaformis 1169540@Vbra 8789
Cyanobacteria-Synechococcus sp. GCF 000155595.1@WP 006453702.1
Cyanobacteria-Calothrix sp. GCF 002368395.1@WP 096622840.1
Rhodophyta-Porphyridium purpureum 35688@3427.9
Rhodophyta-Porphyra umbilicalis GCA 002049455.2@OSX70155.1
Rhodophyta-Gracilariopsis lemaneiformis GCA 003346895.1@NFUL01006118.1
Rhodophyta-Chondrus crispus 2769@XP 005713394.1
OtherPlastid-Vitrella brassicaformis 1169540@Vbra 18555
OtherPlastid-Ectocarpus siliculosus 2880@ES0003G01200
OtherPlastid-Aureococcus anophagefferens 44056@AA0011G01370
OtherPlastid-Thalassiosira pseudonana 35128@TP06G06950
OtherPlastid-Fragilariopsis cylindrus 186039@FC003G06610
OtherPlastid-Fragilariopsis cylindrus 186039@FC005G02140
OtherPlastid-Phaeodactylum tricornutum 2850@PTI 08G01880
Rhodophyta-Galdieria sulphuraria 130081@XP 005702902.1
OtherPlastid-Thalassiosira pseudonana 35128@TP04G01580
Chlamydiae-Simkania negevensis GCF 000237205.1@WP 013944641.1
Chlamydiae-Chlamydiae bacterium GCA 001794905.1@OGN62444.1
Chlamydiae-Parachlamydia acanthamoebae GCF 000253035.1@WP 013924112.1
Chlamydiae-Protochlamydia sp. GCF 001950615.1@WP 075882953.1
Chlamydiae-Waddlia chondrophila GCF 000092785.1@WP 013181292.1
Chlamydiae-Neochlamydia sp. GCF 000648235.2@WP 042240804.1
Chlamydiae-Estrella lausannensis GCF 900000175.1@WP 098038658.1
Chlamydiae-Criblamydia sequanensis GCF 000750955.1@WP 041017870.1
Chlamydiae-Parachlamydiaceae bacterium GCF 000829755.1@WP 042105790.1
Chlamydiae-Parachlamydia sp. GCF 001545115.1@WP 079989498.1
Chlamydiae-Parachlamydia acanthamoebae GCF 000253035.1@WP 006339747.1
Chlamydiae-Simkania negevensis GCF 000237205.1@WP 013943875.1
Chlamydiae-Criblamydia sequanensis GCF 000750955.1@WP 079977941.1
Chlamydiae-Estrella lausannensis GCF 900000175.1@WP 098037790.1
Chlamydiae-Protochlamydia naegleriophila GCF 001499655.1@WP 079992843.1
Proteobacteria-Escherichia coli GCF 000026545.1@WP 000797555.1
Proteobacteria-Neisseria meningitidis GCF 000083565.1@WP 012779567.1
Proteobacteria-Escherichia coli GCF 000026545.1@WP 000233849.1
Proteobacteria-Escherichia coli GCF 000026545.1@WP 000224347.1
Chlamydiae-Chlamydiae bacterium GCA 001794905.1@OGN62445.1
Chlamydiae-Simkania negevensis GCF 000237205.1@WP 013944638.1
Opisthokonta-Trichoplax adhaerens 10228@XP 002118721.1
Chlamydiae-Parachlamydiaceae bacterium GCF 000829755.1@WP 042105788.1
Chlamydiae-Neochlamydia sp. GCF 000648235.2@WP 042240807.1
Chlamydiae-Waddlia chondrophila GCF 000092785.1@WP 013181293.1
Chlamydiae-Parachlamydia sp. GCF 001545115.1@WP 068470789.1
Chlamydiae-Parachlamydia acanthamoebae GCF 000253035.1@WP 006341673.1
Chlamydiae-Criblamydia sequanensis GCF 000750955.1@WP 041017871.1
Chlamydiae-Estrella lausannensis GCF 900000175.1@WP 098039227.1
Chlamydiae-Waddlia chondrophila GCF 000092785.1@WP 013181175.1
Chlamydiae-Simkania negevensis GCF 000237205.1@WP 013943921.1
Chlamydiae-Neochlamydia sp. GCF 000813645.1@WP 039386625.1
Chlamydiae-Chlamydiae bacterium GCA 001794905.1@OGN61662.1
Chlamydiae-Parachlamydiaceae bacterium GCF 000829755.1@WP 042102609.1
Chlamydiae-Protochlamydia naegleriophila GCF 001499655.1@WP 051981685.1
Chlamydiae-Neochlamydia sp. GCF 000813645.1@WP 052243958.1
Chlamydiae-Chlamydiales bacterium GCF 900093655.1@WP 088206379.1
Chlamydiae-Waddlia chondrophila GCF 000092785.1@WP 013181675.1
Chlamydiae-Parachlamydia acanthamoebae GCF 000253035.1@WP 013925138.1
Chlamydiae-Estrella lausannensis GCF 900000175.1@WP 098038183.1
Chlamydiae-Chlamydiae bacterium GCA 001794905.1@OGN62198.1
Chlamydiae-Chlamydia muridarum GCF 003433255.1@WP 100065006.1
Chlamydiae-Chlamydia suis GCF 900149385.2@WP 080124562.1
Chlamydiae-Chlamydia ibidis GCF 000454725.1@WP 020370494.1
Chlamydiae-Chlamydia sp. GCF 001653975.1@WP 066482341.1
Chlamydiae-Simkania negevensis GCF 000237205.1@WP 013944187.1
Chlamydiae-Chlamydiales bacterium GCF 900093655.1@WP 088206541.1
Chlamydiae-Waddlia chondrophila GCF 000092785.1@WP 041941473.1
Chlamydiae-Parachlamydia sp. GCF 001545115.1@WP 068470028.1
Chlamydiae-Parachlamydiaceae bacterium GCF 000829755.1@WP 042104409.1
Chlamydiae-Neochlamydia sp. GCF 000648235.2@WP 042243128.1
Chlamydiae-Estrella lausannensis GCF 900000175.1@WP 098039013.1
Chlamydiae-Criblamydia sequanensis GCF 000750955.1@WP 041017697.1

Figure 22: Arbre phylogénétique simple gène du transporteur TyrP identifié par le pipeline Chlamydia. (page précédente) Cet arbre correspond à la reconstruction phylogénétique du groupe orthologue OG0003312, annoté comme TyrP, sélectionné par le pipeline Chlamydia. L'arbre a été obtenu par IQ-TREE, sous un modèle LG4X en ultrafast-bootstrap. Les espèces identifiées comme "other-plastid" ne sont pas considérées comme des intrus, et font donc partie intégrante des critères de sélection du clan. Les valeurs de bootstrap sont indiquées sur les branches. L'enracinement est en mid-point. L'échelle compte en nombre de substitutions des acides aminés par site. En rouge : Rhodophyta et Rhodelphea, en vert : Viridiplantae et en bleu : Glaucophyta. En rose : Chlamydia, en jaune : Proteobacteria

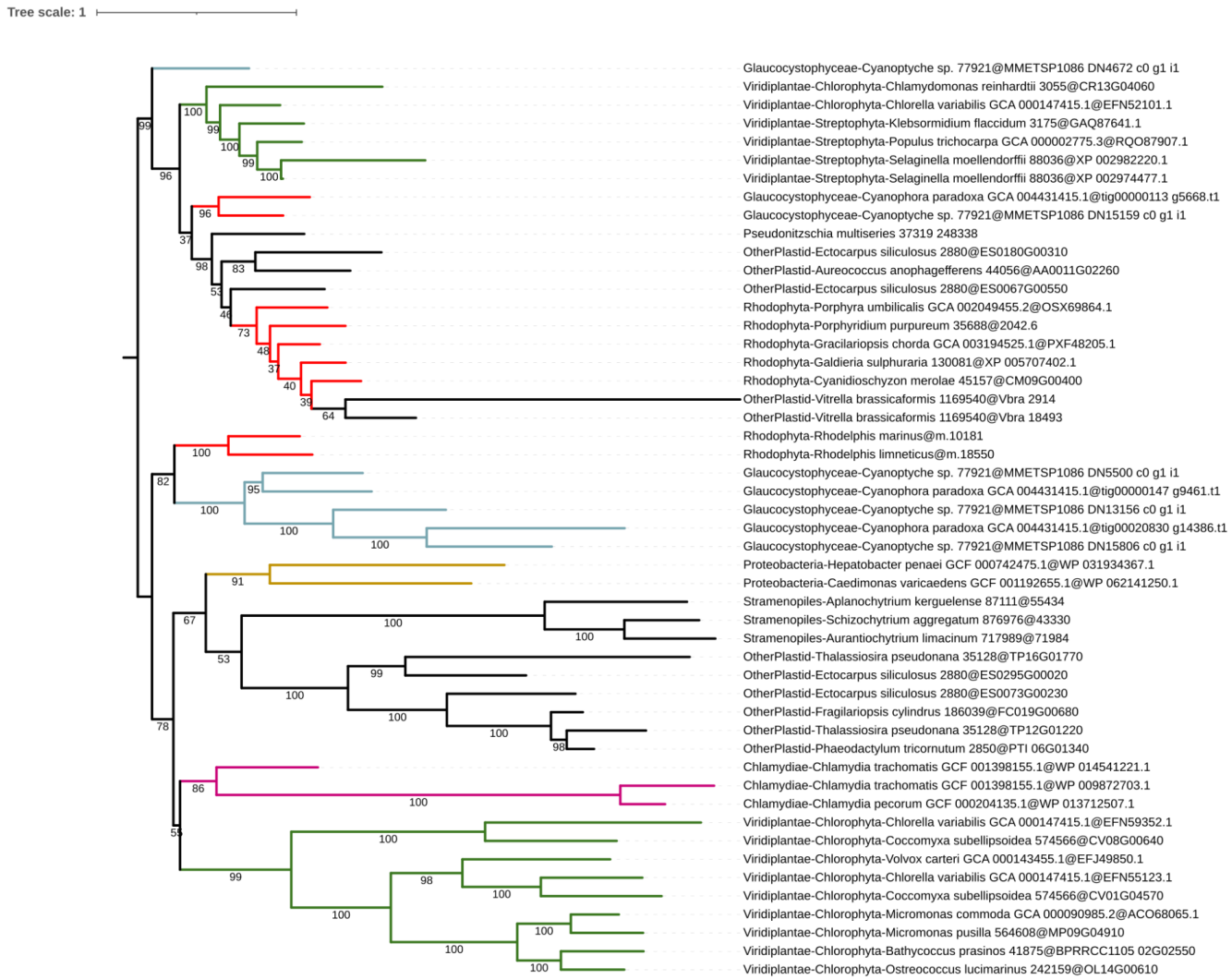


Figure 23: Arbre phylogénétique simple gène du transporteur ATP:ADP antiporter identifié par le pipeline Proteobacteria. Cet arbre correspond à la reconstruction phylogénétique du groupe orthologue OG0003449, annoté comme ATP:ADP antiporter, sélectionné par le pipeline Proteobacteria. L'arbre a été obtenu par IQ-TREE, sous un modèle LG4X en ultrafast-bootstrap. Les espèces identifiées comme "other-plastid" ne sont pas considérées comme des intrus, et font donc partie intégrante des critères de sélection du clan. Les valeurs de bootstrap sont indiquées sur les branches. L'enracinement est en mid-point. L'échelle compte en nombre de substitutions des acides aminés par site. En rouge : Rhodophyta et Rhodelphea, en vert : Viridiplantae et en bleu : Glaucophyta. En rose : Chlamydia, en jaune : Proteobacteria

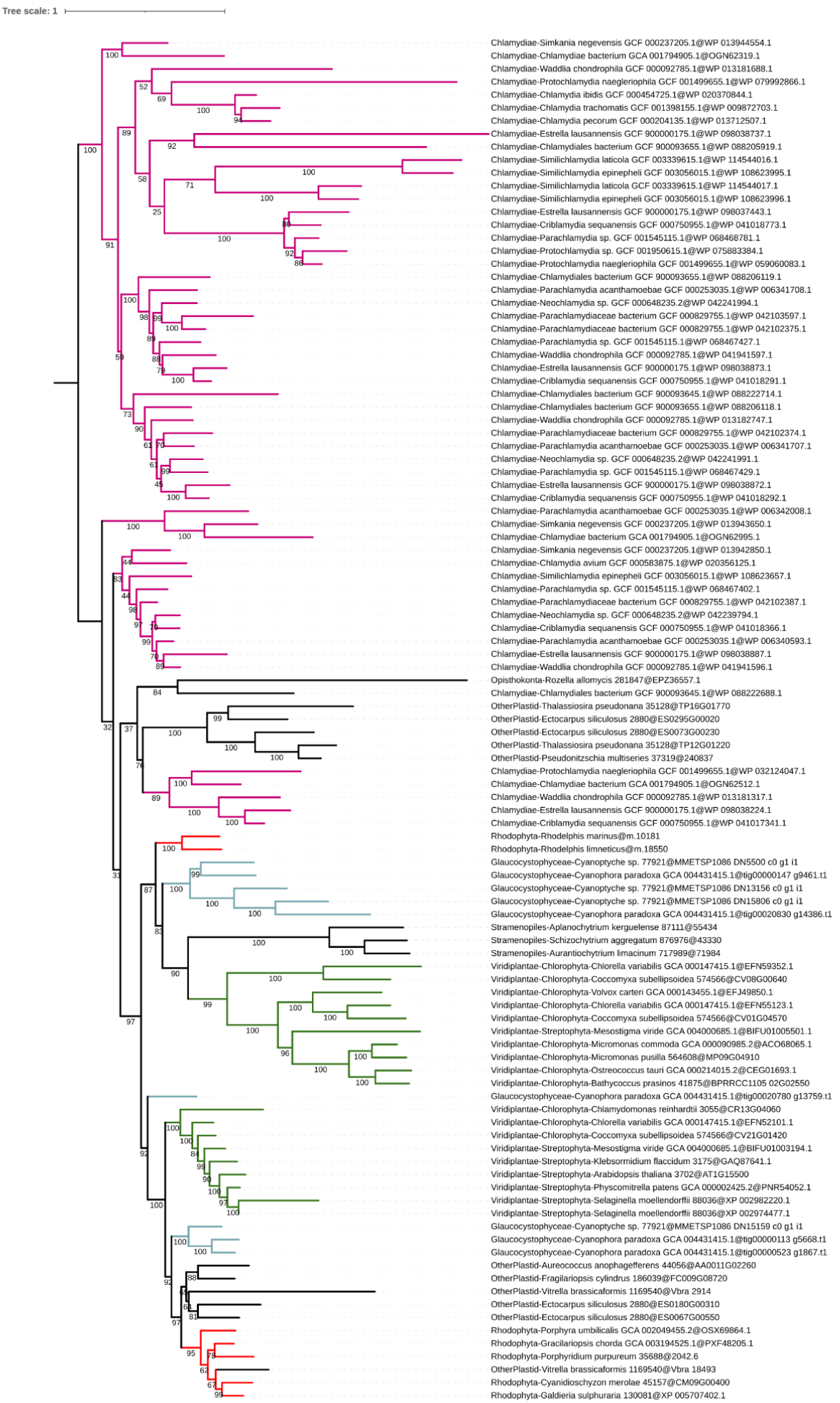

Tree scale: 1
Chlamydiae-Simkania negevensis GCF 000237205.1@WP 013944554.1
Chlamydiae-Chlamydiae bacterium GCA 001794905.1@OGN62319.1
Chlamydiae-Waddlia chondrophila GCF 000092785.1@WP 013181688.1
Chlamydiae-Protochlamydia naegleriophila GCF 001499655.1@WP 079992866.1
Chlamydiae-Chlamydia ibidis GCF 000454725.1@WP 020370844.1
Chlamydiae-Chlamydia trachomatis GCF 001398155.1@WP 009872703.1
Chlamydiae-Chlamydia pecorum GCF 000204135.1@WP 013712507.1
Chlamydiae-Estrella lausannensis GCF 900000175.1@WP 098038737.1
Chlamydiae-Chlamydiales bacterium GCF 900093655.1@WP 088205919.1
Chlamydiae-Similichlamydia laticola GCF 003339615.1@WP 114544016.1
Chlamydiae-Similichlamydia epinepheli GCF 003056015.1@WP 108623995.1
Chlamydiae-Similichlamydia laticola GCF 003339615.1@WP 114544017.1
Chlamydiae-Similichlamydia epinepheli GCF 003056015.1@WP 108623996.1
Chlamydiae-Estrella lausannensis GCF 900000175.1@WP 098037443.1
Chlamydiae-Criblamydia sequanensis GCF 000750955.1@WP 041018773.1
Chlamydiae-Parachlamydia sp. GCF 001545115.1@WP 068468781.1
Chlamydiae-Protochlamydia sp. GCF 001950615.1@WP 075883384.1
Chlamydiae-Protochlamydia naegleriophila GCF 001499655.1@WP 059060083.1
Chlamydiae-Chlamydiales bacterium GCF 900093655.1@WP 088206119.1
Chlamydiae-Parachlamydia acanthamoebae GCF 000253035.1@WP 006341708.1
Chlamydiae-Neochlamydia sp. GCF 000648235.2@WP 042241994.1
Chlamydiae-Parachlamydiaceae bacterium GCF 000829755.1@WP 042103597.1
Chlamydiae-Parachlamydiaceae bacterium GCF 000829755.1@WP 042102375.1
Chlamydiae-Parachlamydia sp. GCF 001545115.1@WP 068467427.1
Chlamydiae-Waddlia chondrophila GCF 000092785.1@WP 041941597.1
Chlamydiae-Estrella lausannensis GCF 900000175.1@WP 098038873.1
Chlamydiae-Criblamydia sequanensis GCF 000750955.1@WP 041018291.1
Chlamydiae-Chlamydiales bacterium GCF 900093645.1@WP 088222714.1
Chlamydiae-Chlamydiales bacterium GCF 900093655.1@WP 088206118.1
Chlamydiae-Waddlia chondrophila GCF 000092785.1@WP 013182747.1
Chlamydiae-Parachlamydiaceae bacterium GCF 000829755.1@WP 042102374.1
Chlamydiae-Parachlamydia acanthamoebae GCF 000253035.1@WP 006341707.1
Chlamydiae-Neochlamydia sp. GCF 000648235.2@WP 042241991.1
Chlamydiae-Parachlamydia sp. GCF 001545115.1@WP 068467429.1
Chlamydiae-Estrella lausannensis GCF 900000175.1@WP 098038872.1
Chlamydiae-Criblamydia sequanensis GCF 000750955.1@WP 041018292.1
Chlamydiae-Parachlamydia acanthamoebae GCF 000253035.1@WP 006342008.1
Chlamydiae-Simkania negevensis GCF 000237205.1@WP 013943650.1
Chlamydiae-Chlamydiae bacterium GCA 001794905.1@OGN62995.1
Chlamydiae-Simkania negevensis GCF 000237205.1@WP 013942850.1
Chlamydiae-Chlamydia avium GCF 000583875.1@WP 020356125.1
Chlamydiae-Similichlamydia epinepheli GCF 003056015.1@WP 108623657.1
Chlamydiae-Parachlamydia sp. GCF 001545115.1@WP 068467402.1
Chlamydiae-Parachlamydiaceae bacterium GCF 000829755.1@WP 042102387.1
Chlamydiae-Neochlamydia sp. GCF 000648235.2@WP 042239794.1
Chlamydiae-Criblamydia sequanensis GCF 000750955.1@WP 041018366.1
Chlamydiae-Parachlamydia acanthamoebae GCF 000253035.1@WP 006340593.1
Chlamydiae-Estrella lausannensis GCF 900000175.1@WP 098038887.1
Chlamydiae-Waddlia chondrophila GCF 000092785.1@WP 041941596.1
Opisthokonta-Rozella allomycis 281847@EPZ36557.1
Chlamydiae-Chlamydiales bacterium GCF 900093645.1@WP 088222688.1
OtherPlastid-Thalassiosira pseudonana 35128@TP16G01770
OtherPlastid-Ectocarpus siliculosus 2880@ES0295G00020
OtherPlastid-Ectocarpus siliculosus 2880@ES0073G00230
OtherPlastid-Thalassiosira pseudonana 35128@TP12G01220
OtherPlastid-Pseudonitzschia multiseries 37319@240837
Chlamydiae-Protochlamydia naegleriophila GCF 001499655.1@WP 032124047.1
Chlamydiae-Chlamydiae bacterium GCA 001794905.1@OGN62512.1
Chlamydiae-Waddlia chondrophila GCF 000092785.1@WP 013181317.1
Chlamydiae-Estrella lausannensis GCF 900000175.1@WP 098038224.1
Chlamydiae-Criblamydia sequanensis GCF 000750955.1@WP 041017341.1
Rhodophyta-Rhodelphis marinus@m.10181
Rhodophyta-Rhodelphis limneticus@m.18550
Glaucocystophyceae-Cyanoptyche sp. 77921@MMETSP1086 DN5500 c0 g1 i1
Glaucocystophyceae-Cyanophora paradoxa GCA 004431415.1@tig00000147 g9461.t1
Glaucocystophyceae-Cyanoptyche sp. 77921@MMETSP1086 DN13156 c0 g1 i1
Glaucocystophyceae-Cyanoptyche sp. 77921@MMETSP1086 DN15806 c0 g1 i1
Glaucocystophyceae-Cyanophora paradoxa GCA 004431415.1@tig00020830 g14386.t1
Stramenopiles-Aplanochytrium kerguelense 87111@55434
Stramenopiles-Schizochytrium aggregatum 876976@43330
Stramenopiles-Aurantiochytrium limacinum 717989@71984
Viridiplantae-Chlorophyta-Chlorella variabilis GCA 000147415.1@EFN59352.1
Viridiplantae-Chlorophyta-Coccomyxa subellipsoidea 574566@CV08G00640
Viridiplantae-Chlorophyta-Volvox carteri GCA 000143455.1@EFJ49850.1
Viridiplantae-Chlorophyta-Chlorella variabilis GCA 000147415.1@EFN55123.1
Viridiplantae-Chlorophyta-Coccomyxa subellipsoidea 574566@CV01G04570
Viridiplantae-Streptophyta-Mesostigma viride GCA 004000685.1@BIFU01005501.1
Viridiplantae-Chlorophyta-Micromonas commoda GCA 000090985.2@ACO68065.1
Viridiplantae-Chlorophyta-Micromonas pusilla 564608@MP09G04910
Viridiplantae-Chlorophyta-Ostreococcus tauri GCA 000214015.2@CEG01693.1
Viridiplantae-Chlorophyta-Bathycoccus prasinos 41875@BPRRCC1105 02G02550
Glaucocystophyceae-Cyanophora paradoxa GCA 004431415.1@tig00020780 g13759.t1
Viridiplantae-Chlorophyta-Chlamydomonas reinhardtii 3055@CR13G04060
Viridiplantae-Chlorophyta-Chlorella variabilis GCA 000147415.1@EFN52101.1
Viridiplantae-Chlorophyta-Coccomyxa subellipsoidea 574566@CV21G01420
Viridiplantae-Streptophyta-Mesostigma viride GCA 004000685.1@BIFU01003194.1
Viridiplantae-Streptophyta-Klebsormidium flaccidum 3175@GAQ87641.1
Viridiplantae-Streptophyta-Arabidopsis thaliana 3702@AT1G15500
Viridiplantae-Streptophyta-Physcomitrella patens GCA 000002425.2@PNR54052.1
Viridiplantae-Streptophyta-Selaginella moellendorffii 88036@XP 002982220.1
Viridiplantae-Streptophyta-Selaginella moellendorffii 88036@XP 002974477.1
Glaucocystophyceae-Cyanoptyche sp. 77921@MMETSP1086 DN15159 c0 g1 i1
Glaucocystophyceae-Cyanophora paradoxa GCA 004431415.1@tig00000113 g5668.t1
Glaucocystophyceae-Cyanophora paradoxa GCA 004431415.1@tig00000523 g1867.t1
OtherPlastid-Aureococcus anophagefferens 44056@AA0011G02260
OtherPlastid-Fragilariopsis cylindrus 186039@FC009G08720
OtherPlastid-Vitrella brassicaformis 1169540@Vbra 2914
OtherPlastid-Ectocarpus siliculosus 2880@ES0180G00310
OtherPlastid-Ectocarpus siliculosus 2880@ES0067G00550
Rhodophyta-Porphyra umbilicalis GCA 002049455.2@OSX69864.1
Rhodophyta-Gracilariopsis chorda GCA 003194525.1@PXF48205.1
Rhodophyta-Porphyridium purpureum 35688@2042.6
OtherPlastid-Vitrella brassicaformis 1169540@Vbra 18493
Rhodophyta-Cyanidioschyzon merolae 45157@CM09G00400
Rhodophyta-Galdieria sulphuraria 130081@XP 005707402.1

Figure 24: Arbre phylogénétique simple gène du transporteur ATP:ADP antiporter identifié par le pipeline Chlamydia. (page précédente) Cet arbre correspond à la reconstruction phylogénétique du groupe orthologue OG0003449, annoté comme ATP:ADP antiporter, sélectionné par le pipeline Chlamydia. L'arbre a été obtenu par IQ-TREE, sous un modèle LG4X en ultrafast-bootstrap. Les espèces identifiées comme "other-plastid" ne sont pas considérées comme des intrus, et font donc partie intégrante des critères de sélection du clan. Les valeurs de bootstrap sont indiquées sur les branches. L'enracinement est en mid-point. L'échelle compte en nombre de substitutions des acides aminés par site. En rouge : Rhodophyta et Rhodelphea, en vert : Viridiplantae et en bleu : Glaucophyta. En rose : Chlamydia, en jaune : Proteobacteria

e. Inventaire de la littérature et corrélation avec nos résultats

L'estimation du nombre de gènes chlamydiens retrouvés chez les Archaeplastida varie entre 30 et 100 en fonction des études et de la stringence des protocoles utilisés. En regroupant les publications de (Ball et al., 2013; Becker et al., 2008; Cenci et al., 2018, 2017; Huang and Gogarten, 2007; Moustafa et al., 2008) nous avons créé un inventaire de la littérature qui se distribue sur 125 de nos groupes orthologues initiaux. Nos méthodes, quant à elles, identifient 26 gènes chlamydiens chez les Archaeplastida lors de l'analyse manuelle des arbres et 57 grâce au pipeline automatique. Au total, ce sont 73 gènes différents qui ont été sélectionnés d'une manière ou d'une autre par nos méthodes. Parmi eux, 51 ont aussi été identifiés dans ces précédentes études (Figure 9B). En fonction des différentes analyses menées, certains arbres sont plus ou moins convaincants pour appuyer l'hypothèse du Ménage à Trois. Nous avons classé la totalité des gènes chlamydiens retrouvés chez les Archaeplastida selon la robustesse du signal, c'est-à-dire en fonction du nombre de méthodes pour lesquelles ils ont été retrouvés (annexe 7). Ainsi, 19 gènes sélectionnés par les 2 pipelines mis en œuvre dans cette étude, validés en analyse manuelle par 2 ou 3 personnes ont aussi été retrouvés dans la littérature. Ceux-ci constituent donc un noyau robuste de gènes ayant une histoire évolutive commune et pouvant appuyer l'implication des Chlamydia dans l'endosymbiose primaire du plaste.

f. Identification des LGT spécifiques entre Chlamydia et Glaucophytes

Comme dit précédemment, 51 gènes de l'inventaire de la littérature sont aussi retrouvés par nos méthodes (Figure 9B). Pour les 74 autres, une analyse manuelle des arbres correspondants a été réalisée, permettant la validation de nos méthodes, mais aussi une mise à jour conservative de l'inventaire des gènes MATH. Ainsi, cette analyse confirme l'invalidation d'un certain nombre de gènes précédemment identifiés dans la littérature, ce qui indique que leur rejet par notre pipeline était justifié. Leur sélection initiale s'explique probablement par l'indisponibilité de certaines données génomiques et protéomiques à l'époque de ces précédentes études, bien que leur disqualification puisse aussi être due aux nouveaux outils d'analyse utilisés ici.

L'analyse manuelle de ces arbres révèle aussi l'existence de gènes montrant une interaction phylogénétique entre des Chlamydia et des Glaucocystophyceae uniquement. Ces arbres ne sont en effet pas reconnus par nos méthodes, puisque les Glaucocystophyceae sont absentes du dataset initial utilisé pour créer les groupes orthologues. De ce fait, pour identifier spécifiquement les transferts de gènes entre ces organismes, un nouveau clustering a été réalisé. A partir des séquences des 33 Chlamydia et de *Cyanophora paradoxa*, 27087 groupes orthologues ont été créés par OrthoFinder (nommé OFCh). Comme certains clusters sont aussi présents dans OF57, l'ensemble des groupes orthologues des 57 eucaryotes utilisés pour les précédents pipelines, nous les avons retirés du dataset OFCh. Seuls ceux présentant au moins 2 Chlamydia et 1 *C. paradoxa* (795 clusters) ont alors été enrichis avec les cyanobactéries, puis les eucaryotes et le reste des bactéries, de la même façon que décrit plus haut (Figure 7). En sortie du pipeline automatique, 60 arbres ayant une relation phylogénétique entre au moins 2 Chlamydia et 1 Glaucocystophyceae sont identifiés, dont 7 sont aussi rapportés dans la littérature. Nous avons dû ici changer les critères de sélection puisque seules 2 Glaucocystophyceae sont présentes dans le dataset. Cependant, au vu de la qualité des données de *Cyanophora paradoxa*, ainsi que la moindre proportion des glaucophytes disponibles et intégrées dans notre protocole, le choix des critères de sélection mériterait d'être optimisé en fonction de l'identification de ces transferts de gènes en particulier. En effet, une analyse manuelle rapide et succincte, effectuée ici par une seule personne et non pas trois comme précédemment, estime une validation d'à peine 30% des arbres sélectionnés. Les 60 arbres identifiés nécessitent donc des analyses plus approfondies pour confirmer les transferts de gènes correspondants.

Conclusions et Discussion

Les différentes problématiques posées par ce projet de thèse visaient l'évaluation du signal chlamydien chez les Archaeplastida, non seulement d'un point de vue existence d'un signal lié à l'endosymbiose primaire du plaste, mais aussi d'un point de vue de la contribution particulière du pathogène, différente des autres acteurs bactériens. Le protocole bioinformatique nous a donc permis de répondre aux différents sujets, et ainsi tester l'hypothèse du Ménage à Trois. L'analyse manuelle des arbres générés en premier lieu replace les transferts de gènes ainsi identifié dans un contexte endosymbiotique, cependant, c'est l'automatisation complète du pipeline qui permet de mettre en perspective le signal chlamydien au sein de l'histoire évolutive globale, aussi bien bactérienne qu'eucaryote. Il est important ici de garder en tête, que le but de ce projet n'était pas d'établir une liste exhaustive des gènes MATH, mais bel et bien de tester les prédictions phylogénétiques de cette hypothèse. Ce faisant, il est probable, et même certain, que des transferts de gènes endosymbiotiques ont pu ne pas être identifiés par nos méthodes, mais aussi que d'autres soient en réalité de faux positifs.

Les différents pipelines mis en œuvre permettent de comparer le signal d'un point de vue des donneurs mais aussi des accepteurs de ces transferts de gènes liés à l'endosymbiose primaire du plaste. Concernant les hôtes eucaryotes, les Chlamydia ont joué un rôle particulier dans l'évolution des Archaeplastida en comparaison des Fungi et Amoebozoa. Nous comptons en effet, 1 LGT identifié entre Chlamydia et Fungii, et 16 entre Chlamydia et Amoebozoa. Pour ces derniers, la fenêtre de temps durant laquelle ces eucaryotes étaient potentiellement en contact avec les Chlamydia peut être considérée comme plus importante que pour les Archaeplastida. Combinée à l'accès possible des membranes à l'infection, ceci laisse entrevoir une contribution théoriquement plus importante des pathogènes. Or, le nombre de LGT identifiés est nettement moindre par rapport à ceux retrouvés chez les Archaeplastida, et la congruence du signal n'est pas aussi nette. De fait, les amibes étant censé être un témoin positif de l'impact chlamydien sur l'évolution des eucaryotes, on peut en déduire que ces pathogènes ont eu un rôle particulier chez les Archaeplastida en comparaison des autres eucaryotes.

Du côté bactérien, 97 LGT sont identifiés pour les 5 principaux groupes bactériens. L'étude confirme ainsi 656 LGTs cyanobactériens dans le génome des Archaeplastida, 57 LGTs

chlamydiens, 44 LGTs de Bacteroidetes, 39 LGTs pour l'ensemble des Proteobacteria, 9 LGTs de Firmicutes et 5 LGTs d'Actinobacteria. Ces résultats montrent la contribution principale des Chlamydias dans la diversité des LGTs bactériens. Bien qu'encourageants, ils ne valident pas pour autant le modèle MATH puisque la typologie des transferts est analogue chez toutes les bactéries, montrant dans tous les cas l'existence d'évènements très anciens contemporains à l'endosymbiose et ne permettant pas de distinguer les Chlamydias des autres groupes bactériens sur ce plan.

L'analyse plus fine de ces LGT, autant d'un point de vue métabolique et fonctionnel que d'un point de vue de la diversité des organismes impactés par ce transfert, laisse cependant entrevoir des différences entre donneurs bactériens. En effet, comme on pouvait le prévoir, aucun groupe contrôle bactérien ne montre, malgré l'analyse de 97 LGTs distincts, l'existence de transferts latéraux multiples affectant une même voie métabolique. Par contre les Chlamydias présentent cette caractéristique pour 4 métabolismes différents ; ceux de l'amidon/glycogène (2 LGTs), des isoprénoides (3/4 LGTs), de la ménaquinone (2 LGTs) et du tryptophane (3/4 LGT). A l'exception de la voie de synthèse de la ménaquinone, qui sera discutée par la suite, chacun des gènes chlamydiens impactant ces voies métaboliques, et déjà rapportés dans la littérature, sont aussi identifiés par nos méthodes. De plus, deux de ces voies sont associées à 3 des 7 transporteurs plastidiaux retrouvés dans les LGT chlamydiens alors qu'aucun transporteur plastidial n'est retrouvé chez l'ensemble des 97 LGTs bactériens non chlamydiens. Bien que deux transporteurs soient identifiés pour le pipeline protéobactérien, une analyse approfondie des arbres correspondants redéfinit ces transferts comme chlamydiens et non pas proteobactériens. Ces phylogénies montrent en effet un branchement de deux Proteobacteria seulement avec les Archaeplastida (malgré l'enrichissement avec les 36 espèces sélectionnées) et les trois Chlamydia présentes alors dans le jeu de données, qui sont donc considérées comme étant des intrus mais en proportion suffisamment basse pour valider les critères de sélection de PhySortR. En toute logique, ces deux transporteurs identifiés avec le pipeline protéobactérien le sont aussi avec le pipeline chlamydien, montrant alors une diversité plus importante des pathogènes branchée avec les Archaeplastida. Selon les paramètres retenus dans notre étude (calibrée sur base de l'analyse manuelle), un arbre est identifié par le pipeline si 2 donneurs branchent avec 3 accepteurs minimum. En ajustant ces paramètres pour identifier des arbres ayant un branchement d'au moins 3 donneurs et 3 accepteurs, nous retirons 6 LGT chlamydiens, 5 Bacteroidetes, 1 actinobactérien, 4 Amibes, 3 Firmicutes, 11 protéobacteriens et 16 cyanobactériens. Si ceci

n'induit pas de changement majeur sur les conclusions apportées précédemment, il a pour effet de retirer les deux transporteurs protéobactérien initialement identifiés, appuyant donc l'absence de transporteurs bactériens autres que cyanobactériens et chlamydiens chez les Archaeplastida. La connectivité précoce d'origine procaryote (l'essentiel étant clairement d'origine eucaryote) du plaste apparait donc sous le contrôle exclusif de ces deux groupes.

Comme dit précédemment, le pipeline bioinformatique a été conçu de sorte à tester l'hypothèse du Ménage à Trois et non pas à établir une liste exhaustive des gènes MATH. De fait, nous avons accepté quelques limites dans notre protocole. L'un d'eux notamment réside dans le regroupement en familles de gènes des différents protéomes eucaryotes. En effet, suite à l'analyse manuelle des arbres sélectionnés par le pipeline semi-automatique et de l'inventaire de la littérature, il est apparu évident que certains groupes orthologues distincts auraient dû être regroupés ensemble quand d'autres, largement multigéniques, auraient nécessité une découpe plus stricte. C'est le cas pour les gènes de la voie de synthèse de la ménaquinone. Chez les plantes vertes, le PHYLLO encode quatre des enzymes intervenant dans cette voie métabolique. Or chez les bactéries, au contraire, les gènes correspondants sont séparés les uns des autres. Le regroupement en familles de gène étant basé sur les eucaryotes, un seul groupe orthologue est identifié pour une majorité de la voie Men, qui sera alors enrichi avec les séquences bactériennes dans la suite du protocole. Le caractère multigénique de l'arbre généré entraîne donc nécessairement une non-sélection du gène, bien que son origine chlamydienne fut démontrée dans Cenci et al.

Concernant la ménaquinone toujours, nous avons mis en place une analyse annexe afin d'identifier une potentielle voie de synthèse alternative de cette vitamine K chez les Gloeobacterales. Après avoir déterminé la distribution des voies de synthèse de la ménaquinone, cette analyse s'est basée sur le regroupement en familles de gènes des 48 cyanobactéries, puis sur l'identification des gènes communs aux seuls détenteurs de cette potentielle voie alternative. Cependant, l'approche bioinformatique de ce projet annexe s'est avérée non concluante, puisque très peu de gènes candidats ont été identifiés, et nécessiterait tout du moins d'être repensée et approfondie. D'un point de vue fonctionnel, il était prévu de faire suivre cette analyse bioinformatique de la voie de synthèse de la ménaquinone par la création et la caractérisation d'une banque de mutants de *Gloeobacter violaceus*. Suite à la crise sanitaire de 2020-2021, ce projet a cependant été reporté.

L'ensemble des résultats présentés précédemment se base sur la sélection des organismes entrant le protocole et sur le regroupement en familles de gènes des protéomes eucaryotes. Or

dans ce jeu de données initial, aucun glaucophyte n'était présent. La réorientation du pipeline sur l'identification des LGT entre les Chlamydia et les glaucophytes permet donc de pallier cette absence. Les 60 LGT identifiés sont cependant à prendre avec précaution. En effet, seuls deux glaucophytes ont des données d'assez bonne qualité pour intégrer le pipeline, les critères de sélection des arbres ont donc été adaptés en ce sens, induisant la sélection de chaque arbre ayant au minimum un glaucophyte branché avec au moins deux Chlamydia. Après une analyse manuelle très succincte de ces arbres, il s'avère que la majorité seraient à reconsidérer.

L'automatisation du pipeline mis en place dans la deuxième partie du projet permet de diminuer les difficultés d'interprétation des arbres phylogénétiques. Il a été optimisé, certes, sur l'analyse manuelle des arbres chlamydiens, mais surtout pour établir une comparaison des différents signaux bactériens. De fait, certains paramètres seraient à revoir pour pouvoir différencier les EGT des ERGT particulièrement mais aussi différencier les différents scénarios possibles. Nous avons en effet imaginé plusieurs scénarios phylogénétiques en fonction de la "profondeur" et de la "rareté" du signal retrouvé dans le diversité archaeplastidienne. Rappelons ici qu'il s'agit de décrire différents scénarios possibles dans le but final de tester l'hypothèse du Ménage à Trois et non pas de lister les gènes MATH de façon exhaustive. Ainsi, deux mêmes topologies, selon les critères établis, peuvent peuvent être catégorisées dans le même scénario sans pour autant être reliées à l'endosymbiose primaire du plaste. En effet, si l'on peut facilement associer le scénario profond et dense avec un potentiel gène MATH (EGT ou ERGT) et, à l'inverse, associer le scénario superficiel et rare avec un transfert de gène (LGT) tardif, l'interprétation des deux autres scénarios intermédiaires (profond et rare ou superficiel et dense) est plus sujette aux questionnements. La diversité des espèces impactées par le transfert de gène ainsi que les tropismes phylogénétiques éclaireront sur la possibilité des transferts multiples d'un côté, versus un transfert chez l'ancêtre commun mais perdu chez les descendants de l'autre. Ceci reflète finalement les difficultés de l'analyse manuelle des arbres, auquel cas le scénario dense et profond correspondrait aux arbres validés par les observateurs, et le scénario superficiel et rare aux arbres rejetés par le pipeline. Les arbres répondant aux deux autres scénarios, quant à eux, seraient catégorisés en analyse manuelle comme incertains.

A la suite de ce projet de thèse, nous pouvons identifier trois critères pour lesquels le signal chlamydien chez les Archaeplastida peut être considéré comme différent des autres signaux bactériens: i) le nombre de LGT identifié légèrement supérieur, ii) l'impact multiple de ces

LGT sur les voies métaboliques et iii) la présence de transporteurs. Ces trois caractéristiques, combinées à l'analyse manuelle des arbres qui replace les LGT dans un contexte endosymbiotique, appuient l'hypothèse d'un ménage à trois lors de l'endosymbiose primaire du plaste, qu'il serait utile de comparer à la totalité de la contribution bactérienne.

Dans ce doctorat, je me concentre sur la façon dont le cyanobionte a pu devenir un plaste en utilisant un point de vue fonctionnel, car je crois que la première exigence d'une telle symbiose devrait être métabolique. Cela permettra donc de créer un lien entre les organismes et une pression sélective pour maintenir une relation symbiotique. Je suis donc d'accord pour dire qu'une vision plus complète de l'organellogenèse devrait incorporer une vision plus génomique de l'intégration des cyanobiontes. En effet, l'idée de prérequis génomique, telle que défendue dans le "shopping bag model" (Howe, et al., 2008), pourrait aider à définir comment les liens ont pu être créés. Cette hypothèse est un bon complément, qui ne rejette pas nécessairement l'hypothèse MAT. En outre, notre vision du transfert latéral de gènes (c'est-à-dire le transfert de gènes endosymbiotiques et le transfert de gènes liés à l'endosymbiose) qui s'est produit au cours du processus d'endosymbiose est particulièrement conforme au "Shopping bag model". Le point de divergence étant que notre vision est principalement fonctionnelle alors que leur vision est centrée sur la préadaptation. Une deuxième hypothèse à prendre en compte pour comprendre le transfert latéral de gènes est la capacité d'incorporer des gènes provenant directement du symbiote. En effet, il a été proposé que le besoin de plus d'un endosymbiont facilite leur acquisition une fois que l'endosymbionte meurt et que l'ADN est libéré. Cette hypothèse, appelée "The limited transfer window hypothesis" (Barbrook et al., 2006), pourrait être explorée plus en profondeur. Cependant, cette hypothèse est difficilement testable ici, et ne peut être discutée que théoriquement. En effet, si l'hypothèse MAT est correcte, le nombre de cellules de chlamydia dans sa vésicule d'inclusion devrait être élevé et que rien n'empêche, du fait du non-couplage de la division cellulaire entre l'hôte, chlamydia et le cyanobionte, la présence de plusieurs cyanobactéries à l'intérieur de la vésicule d'inclusion.

Futures Recherches et Perspectives

A la suite de ce projet, 6 mois supplémentaires m'ont été accordés pour finaliser l'analyse du signal chlamydien chez les Archaeplastida. Ces 6 mois permettront d'analyser la contribution bactérienne totale dans l'évolution des Archaeplastida et d'approfondir l'aspect fonctionnel des LGT identifiés. En effet, pour l'instant, seuls 4 signaux bactériens ont été étudiés, et bien que l'on puisse déjà observer des différences importantes dans les profils phylogénétiques, topologiques et de diversité, l'analyse de la contribution bactérienne totale remettra les chlamydia dans un contexte d'évolution plus complet. Combinés à une analyse fonctionnelle là aussi plus approfondie des LGT identifiés, ainsi qu'à une analyse de la diversité impactée par ces transferts de gènes (autant du côté donneur de LGT que du côté accepteur), l'ensemble des résultats du projet replace le chlamydia dans l'évolution de manière générale, et plus particulièrement dans le contexte de l'endosymbiose primaire du plaste. D'un point de vue théorique et phylogénétique, la conceptualisation plus poussée et plus détaillée des EGT et ERGT, de leur profil topologique et de diversité, permettrait de distinguer les différents transferts de gènes et de statuer sur le rôle chlamydien (symbionte transitoire ou simple source de ERGT). D'un point de vue fonctionnel, une proportion importante des LGT identifiés ne l'ont pas été dans les précédentes publications sur le sujet, et de fait laissent entrevoir l'impact chlamydien sur de nouvelles voies métaboliques.

Cependant, ce projet a été réalisé en tirant parti de toutes les nouvelles ressources disponibles, que ce soit au niveau génomique et protéomique, qu'au niveau méthodologique. Ainsi, une fois finalisée, et dans l'état actuel des ressources, cette étude marque l'aboutissement du test de l'Hypothèse du Ménage à Trois de manière bioinformatique. Peut-être serait-il possible d'identifier des preuves environnementales de ce Ménage à Trois?

Matériels et Méthodes

La totalité des travaux de calculs présentés dans cette étude a été réalisée sur un cluster de calcul IBM/Lenovo Flex sous CentOS 6.6. Le système est composé d'un gros nœud de calcul (x440) et de 11 autres plus petits (x240), pour un total de 228 cœurs physiques, 3 To de RAM et 160 To de stockage partagé.

1. Sélection des données génomiques et protéomiques

a. Données eucaryotes et bactériennes

L'analyse présentée ici se base sur une sélection des données génomiques et protéomiques optimisée pour maximiser la représentation de la diversité tout en réduisant au minimum les contaminations. Nous avons pris parti d'utiliser des jeux de données déjà disponibles au laboratoire, pour lesquels la qualité des données a été vérifiée dans de précédents projets, mais aussi le recours à des outils automatiques pour la création et l'évaluation de notre sélection, respectivement TQMD (Léonard et al., 2021) et 42 (Van Vlierberghe et al., 2021). TQMD (ToRQuEMaDA: tool for retrieving queried Eubacteria, metadata and dereplicating assemblies) permet, à partir des bases de données publiques notamment, de produire des listes de génomes procaryotes dérépliqués et représentatifs de la diversité. Les jeux de données bactériens nécessaires pour cette étude (détaillés ensuite) ont donc été créés à partir de TQMD, et sont publiés dans Léonard et al., 2021 (publication pour laquelle je suis co-auteur). 42 quant à lui permet non seulement d'évaluer la qualité des données génomiques et protéomiques (que ce soit par l'estimation des contaminations ou dans une moindre mesure par la complétude des données), mais aussi d'enrichir des groupes orthologues avec des séquences d'organismes supplémentaires. Cette deuxième fonction de l'outil sera utilisée dans la deuxième partie du protocole.

Ainsi, nous avons sélectionné au départ de notre étude 57 protéomes eucaryotes (http://hdl.handle.net/2268/254874; https://doi.org/10.6084/m9.figshare.13550267.v2), pour lesquels un clustering en groupes orthologues était déjà disponible au laboratoire (créé par OrthoFinder (Emms and Kelly, 2015) sous un paramètre d'inflation de 1.5), ainsi que les différents datasets générés par TQMD et publiés dans Léonard et al., 2021. A ce listing, nous avons manuellement sélectionné et ajouté 14 génomes d'Archaeplastida, comprenant 2 Glaucocystophyceae et 2 Rhodelphea (Gawryluk et al., 2019), 10 Amoebozoa, ainsi que 2 cyanobactéries considérées comme basales *Gloeobacter violaceus* et *Pseudanabaena biceps*. Tous les génomes et protéomes sélectionnés ont été évalués par 42, dont le fichier de configuration est disponible en annexe 5 (pour plus de détails voir : (Cornet and Baurain, 2022; et Van Vlierberghe et al., 2021), afin de s'assurer de la qualité des données.

Au total, nous avons choisi 72 eucaryotes (dont 54 eucaryotes photosynthétiques), 10 Amoebozoa, 33 Chlamydia, 48 Cyanobacteria, 37 Bacteroidetes, 36 Proteobacteria, 22 Firmicutes et 20 Actinobacteria. En fonction de l'orientation du pipeline décrit par la suite, et donc en fonction du signal étudié, un sous-ensemble de cette sélection entrera dans le protocole bioinformatique (Figure 12). De ce fait, pour assurer une représentation bactérienne globale, nous avons combiné deux autres jeux de données bactériens, sans distinction de taxa, provenant aussi de TQMD: le premier contenant 49 espèces, publié dans Léonard et al., 2021, et le deuxième de 92 espèces déjà disponible au laboratoire.

b. Répartition des génomes et protéomes sélectionnés dans les différentes analyses

Le test de l'Hypothèse du Ménage à Trois comporte une partie d'identification du signal chlamydien chez les Archaeplastida, mais aussi une seconde partie de comparaison de ce signal avec d'autres groupes. Pour chaque contrôle, le pipeline est orienté sur le crible des LGT spécifiques aux groupes étudiés. Cette réorientation est mise en œuvre notamment par l'échantillonnage des génomes et protéomes intégrant le pipeline. Ainsi, chaque condition se voit attribuer un jeu de données en partie spécifique. Les 72 eucaryotes sélectionnés précédemment, les 48 cyanobactéries ainsi qu'une majorité des deux listes de 49+92 bactéries sont communs à tous les contrôles. Cependant, les jeux de données bactériennes spécifiques varient d'un pipeline à l'autre. Le crible de LGT entre Chlamydia et Archaeplastida nécessite l'introduction des 33 Chlamydia dans le protocole, et donc l'adaptation de la liste bactérienne générale en retirant ce groupe taxonomique. Il en va de même pour les autres groupes contrôles : les datasets des organismes cibles testés sont ajoutés au pipeline, tout en veillant à les retirer de la sélection bactérienne générale (de 49+92 espèces). Tous les génomes et protéomes sélectionnés, ainsi que leur entrée dans le pipeline en fonction du signal étudié, sont répertoriés dans la table annexe 9.

2. Identification du signal chlamydien chez les Archaeplastida

a. Pipeline semi-automatique : démarche méthodologique générale

Le pipeline semi-automatique décrit ici se divise en trois étapes (Figure 7): 1) la répartition et sélection des 57 protéomes eucaryotes en groupes orthologues; 2) l'enrichissement et filtration des groupes orthologues sélectionnés et 3) la sélection des LGT après reconstruction phylogénétique Ce pipeline est dit semi-automatique puisqu'il est suivi d'une analyse manuelle des arbres générés. Après séparation des données génomiques et protéomiques en familles de gènes ou groupes orthologues (OG ou clusters), le signal chlamydien est directement mis en compétition avec le signal cyanobactérien. Chaque cluster retenu est ensuite aligné avec MAFFT (Katoh and Standley, 2013) et enrichi avec les données eucaryotes et bactériennes sélectionnées précédemment La reconstruction phylogénétique

permet alors de vérifier si le signal chlamydien tient toujours. Enfin, une analyse manuelle des arbres sélectionnés par le pipeline permet de confirmer ou infirmer le transfert de gène et ensuite, si le transfert est validé, de déterminer si le contexte est endosymbiotique. L'ensemble du pipeline est représenté dans la Figure 7.

b. Tri des groupes orthologues

A partir des 57 eucaryotes, un clustering en familles de gène a été réalisé par OrthoFinder (paramètre d'inflation fixé à 1.5, (Emms and Kelly, 2015)) (appelé OF57). Parmi les groupes orthologues (OG) contenant moins de 3000 séquences, classify-mcl.pl (https://metacpan.org/dist/Bio-MUST-Tools-Mcl) n'a retenu que ceux contenant au moins 2 Viridiplantae et/ou Rhodophyta. Une deuxième étape de réduction des groupes orthologues intervient ensuite, après l'enrichissement avec les séquences chlamydiennes. Classify-ali.pl ne retient alors que les OG présentant au moins un chlamydia.

c. Enrichissements et filtration des groupes orthologues

Chaque groupe orthologue sélectionné à l'étape précédente est enrichi avec les protéomes des 33 Chlamydia et 48 Cyanobacteria. Cette étape est réalisée par 42, dont le fichier de configuration est disponible en annexe 6 (pour plus de détails voir : (Cornet and Baurain, 2022; et Van Vlierberghe et al., 2021). Classify-ali.pl (https://metacpan.org/dist/Bio-MUST-Core) nous a ensuite permis de filtrer une seconde fois les groupes orthologues sur la présence d'au moins un Chlamydia. Ces OG sélectionnés par classify-ali.pl ont ensuite été nettoyés par prune-outliers.pl (https://metacpan.org/dist/Bio-MUST-Apps-FortyTwo, –min-threshold=0, max-threshold=0.9, evalue=1e-02) et HmmCleaner (Di Franco et al., 2019) et alignés avec MAFFT v7.453 (Katoh and Standley, 2013). ali2phylip.pl (min=0.3, max=05) a ensuite réduit la proportion de sites manquants dans les alignements. La reconstruction phylogénétique des arbres correspondant à chaque OG a été réalisée par RAxML (Stamatakis, 2014), sous un modèle PROTGAMMALG4X, et en ultra-fast bootstrap. L'analyse automatique des arbres ainsi générés, réalisée par clans-label.pl, repère alors les arbres ayant une interaction phylogénétique entre au moins un Chlamydia et un Archaeplastida. Cet outil analyse chaque bipartition de chaque arbre et permet l'identification des clans composés d'au minimum un Chlamydia et d'un Archaeplastida, sans interruption par d'autres espèces.

Une deuxième phase d'enrichissement intervient alors avec 42, sur les alignements des arbres d'intérêt, avec les protéomes et génomes sélectionnés précédemment non encore inclus dans les alignements. À ce stade, tous les groupes orthologues ont reçu, ou ont eu la possibilité de recevoir, les séquences homologues des 286 organismes sélectionnés pour la première étape de cette étude. Comme décrit précédemment, nous avons éliminé les séquences présentant une faible similarité et réduit la proportion de sites manquants avec prune-outliers.pl (treshold

sélectionné à 0.2, evalue=1e-02) et ali2phylip.pl (min=0.3, max=0.5, mask=BMGE) (https://metacpan.org/dist/Bio-MUST-Core).

Plusieurs outils et paramètres ont été testés dans la mise au point du pipeline. En ce qui concerne le filtrage des alignements, nous avons comparé les résultats obtenus avec HmmCleaner combiné à ali2phylip.pl (min=0.3, max=0.5), ali2phylip.pl en appliquant un retrait des blocs par BMGE (Criscuolo and Gribaldo, 2010) (min=0.3, max0.5, bmge-mask=loose), et ali2phylip.pl en réduction des proportions de sites manquants uniquement (min=0.3, max=0.5). De même, la reconstruction phylogénétique des arbres a été réalisée par RAxML (PROTGAMMALG4X, ultrafast-bootstrap) en comparaison à IQ-TREE (Nguyen et al., 2015)(LG4X, ultrafast-bootstrap).

d. Reconstruction phylogénétique et sélection des LGT

Après les étapes successives d'enrichissement et de filtration des groupes orthologues, la reconstruction phylogénétique a été réalisée par RAxML, avec le modèle PROTGAMMALG4X et bootstrap ultrarapide (Stamatakis, 2014). clans-label.pl a ensuite identifié les arbres avec une interaction phylogénétique entre au moins une Chlamydia et au moins une Archaeplastida.

e. Analyse manuelle des arbres

Les arbres sélectionnés ont été automatiquement colorés par format-tree.pl (trouvé dans Bio::MUST::Core) et visualisés par iTOL v4 (Letunic and Bork, 2019). L'analyse manuelle des arbres s'est ensuite effectuée par trois personnes indépendantes. Plusieurs critères d'évaluation ont été définis dans un premier temps, de sorte à permettre la classification des arbres en trois catégories : 1) en faveur de MATH, c'est-à-dire montrant une interaction phylogénétique qui indique un contexte endosymbiotique ; 2) pas en faveur de MATH et 3) incertain, impossible à conclure. Ces critères comprennent i) la qualité et la diversité du donneur (Chlamydia) dans les sous-arbres d'intérêt, c'est à dire l'abondance des organismes dans le clan, mais aussi la présence de multiple séquence ii) la qualité et la diversité des accepteurs (Archaeplastida et autres organismes photosynthétiques) dans les sous-arbres d'intérêt, iii) la présence d'intrus dans les sous-arbres d'intérêt et enfin iv) la topologie générale de l'arbre, afin d'identifier de potentiels paralogue ou familles multigéniques, mais aussi d'évaluer la diversité totale des donneurs et accepteurs.

f. Inventaire des gènes chlamydiens dans la littérature

Pour créer un inventaire des gènes de Chlamydia chez les Archaeplastida et le comparer à nos propres données, nous avons récupéré les séquences publiées dans Ball et al., 2013; Becker et al., 2008; Cenci et al., 2018, 2017; Huang and Gogarten, 2007; Moustafa et al., 2008. Nous avons dérépliqué ces séquences avec un cd-hit à 95% (Li and Godzik, 2006) et identifié les

groupes orthologues dans notre pipeline correspondant à cet inventaire grâce à BLAST (Altschul et al., 1997).

3. Spécificité du signal chlamydien chez les Archaeplastida

a. Automatisation du pipeline

Alors que l'analyse manuelle est nécessaire pour évaluer et replacer le transfert de gène identifié par le pipeline semi-automatique dans un contexte MATH, la comparaison du signal chlamydien chez les Archaeplastida au signal d'autre bactériens nécessite l'automatisation complète du protocole. En nous basant sur l'analyse manuelle, nous avons calibré ce pipeline automatique de sorte qu'il mime au mieux l'application manuelle des critères d'évaluation des arbres précédemment cités (Figure 7).

Après une première filtration des groupes orthologues sur la présence d'au moins 2 Viridiplantae et/ou Rhodophyta, tous les OG sont enrichis en Chlamydia et en cyanobactéries avec 42. Les clusters sélectionnés par classify-ali.pl ont ensuite été nettoyés et alignés par cleanOG.pl (–min-threshold=0, max-threshold=0.9, evalue=1e-02) et MAFFT v7.453 (Katoh and Standley, 2013). Nous avons utilisé 42 pour enrichir tous les groupes orthologues avec le reste des génomes et protéomes eucaryotes et bactériens sélectionnés pour ce pipeline. Après une seconde filtration des clusters par cleanOG.pl et ali2phylip.pl (min=0.3, max=0.5, mask-bmge=loose), IQ-TREE (Nguyen et al., 2015, LG4X, ultrafast-bootstrap) s'est chargé de la reconstruction phylogénétique des arbres. Nous avons ensuite utilisé Treeshrink (Mai and Mirarab, 2018) pour supprimer les longues branches et Treemer (Menardo et al., 2018) pour dérépliquer phylogénétiquement les arbres en conservant au moins une séquence par espèce. Chaque séquence identifiée par Treeshrink et Treemer a ensuite été retirée des fichiers d'alignements et IQ-TREE a effectué la reconstruction phylogénétique finale avec un modèle LG4X et en ultrafast-bootstrap. PhySortR (Stephens et al., 2016) a alors automatiquement sélectionné chaque arbre présentant un clan d'intérêt, en autorisant 10% d'intrus et avec un minimum de 30% de toutes les espèces cibles dans l'arbre se trouvant dans le sous-arbre d'intérêt. Enfin, classif-ali.pl a identifié les clans présentant une diversité minimale de 2 donneurs et 3 accepteurs https://metacpan.org/dist/Bio-MUST-Core).

b. Contrôles du signal chlamydien chez les Archaeplastida

Pour évaluer le signal chlamydien chez les Archaeplastida, 2 types de contrôle sont proposés : d'abord le contrôle du signal chlamydien chez les Archaeplastida par rapport aux autres groupes bactériens et ensuite, le contrôle du signal chlamydien spécifiquement chez les Archaeplastida par rapport aux autres groupes eucaryotes, appelés respectivement contrôle "donneur" et "accepteur".

Pour le côté donneur de ce contrôle, nous avons comparé le signal chlamydien chez les Archaeplastida avec les Proteobacteria, Bacteroidetes, Firmicutes et Actinobacteria. Dans le

même esprit, nous avons également quantifié le signal des cyanobactéries chez les Archaeplastida, afin d'évaluer la sensibilité de nos méthodes. Pour chaque cas, nous avons réorienté le pipeline automatique sur l'identification des transferts de gènes entre chaque groupe bactérien et les Archaeplastida. Le protocole reste exactement le même que celui décrit précédemment, mais la combinaison de génomes et protéomes entrant dans le pipeline diffère pour chaque condition (annexe 9). Ainsi, au lieu d'ajouter les 33 protéomes de Chlamydia et de filtrer les clusters sur la présence de ces organismes, nous avons enrichi les groupes orthologues avec le dataset correspondant au pipeline contrôle étudié et filtré les clusters sur la présence de ces espèces cibles. Nous avons également ajusté l'ensemble des données bactériennes en retirant les espèces cibles de la combinaison des deux listes de 49 et 92 espèces.

Pour le côté accepteur de l'analyse, nous avons comparé le signal chlamydien chez les Archaeplastida au signal chlamydien chez les amibes et les champignons. Le pipeline a également été réorienté sur l'identification des transferts de gènes correspondant en modifiant le filtre du clustering. Après le clustering effectué par OrthoFinder (Emms and Kelly, 2015), nous avons filtré les groupes orthologues sur la présence d'au moins 2 Fungi ou 2 Amoebozoa. En ce qui concerne le pipeline amoebozoa, cette étape a été précédée par l'enrichissement de tous les clusters avec les génomes et protéomes d'amoebozoa. Le reste du protocole reste exactement le même que celui décrit pour le signal chlamydien chez les Archaeplastida.

c. Concaténation et congruence du signal

Pour chaque condition de réorientation du pipeline, nous avons évalué la congruence du signal identifié. A partir des séquences des espèces cibles présentes dans les sous-arbres d'intérêt, l'ensemble des transferts de gènes identifiés par les différents pipelines ont subi deux types d'analyse : une analyse de la congruence du signal par concaténation en super-matrice, et une analyse par concaténation en super-arbre.

Après filtration et alignements des séquences par ali2phylip.pl implémenté du filtre BMGE (min=0.3, max=0.5, bmge-mask=loose, https://metacpan.org/dist/Bio-MUST-Core, Criscuolo and Gribaldo, 2010,) et MAFFT v7.453 (Katoh and Standley, 2013), les supermatrices ont été créées par SCaFoS v1.30k (Roure et al., 2007), en utilisant la distance évolutive minimale comme critère de sélection parmi les séquences paralogues d'un même OTU (threshold d'élimination complète à 25%), le pourcentage maximal de sites manquants pour une séquence complète fixé à 10 et le nombre maximal d'OTUs manquants fixé à 25 (sauf pour les Firmicutes, fixé à 22, et pour les Actinobacteria, fixé à 20). De plus, les espèces dont la fréquence des séquences était inférieure à 10% ont été retirées de l'alignement. IQ-TREE a alors permis la reconstruction phylogénétique des supermatrices avec les modèles LG4X, C20 et C60 et en ultrafast-bootstrap (Nguyen et al., 2015).

En utilisant split-matrix.pl (https://metacpan.org/dist/Bio-MUST-Core) pour rediviser les supermatrices en gènes individuels, nous avons ré-échantillonné chaque ensemble de groupes orthologues, pour une taille correspondant à un tiers des supermatrices. Pour chacun des 100 réplicats créés par jack-ali-dir.pl (https://metacpan.org/dist/Bio-MUST-Core) pour chaque pipeline (sauf pour les cyanobactéries pour lesquelles nous avons réalisé 2 ensembles de réplicats, l'un correspondant à un tiers de la supermatrice initiale et l'autre avec une longueur fixée à 4500 AA), SCaFoS v1.30k a recréé les supermatrices en utilisant les mêmes paramètres que ceux décrits précédemment et IQ-TREE a permis la reconstruction phylogénétique des arbres, avec un modèle LG4X et en ultrafast-bootstrap.

En parallèle, nous avons également réalisé une analyse de la congruence du signal basée sur la création de super-arbres. ASTRAL-III (v5.7.5) (Mirarab et al., 2014) a produit un super-arbre pour chaque pipeline à partir d'arbres phylogénétiques simples gènes générés par IQ-TREE (LG4X, bootstrap ultra-rapide) reprenant uniquement les espèces cibles d'intérêt pour chaque condition.

d. Enracinement des arbres issus des concaténations

Les arbres issus des concaténations sont uniquement composés d'espèces cibles, et ne possèdent pas d'outgroup. L'enracinement de ces arbres est donc manuel, sur les espèces donneurs les plus basales présentes. Nous avons donc d'abord reconstruit la phylogénie d'espèce pour chaque jeu de données bactérien.

Pour chaque sélection de TQMD (Léonard et al., 2021), nous avons récupéré les protéomes correspondant et utilisé 42 pour récupérer leurs protéines ribosomales (Cornet and Baurain, 2022; Van Vlierberghe et al., 2021). La taxonomie de ces protéines ont été étiquetées en calculant le dernier ancêtre commun (best hit BLAST) dans les alignements correspondants (en excluant les auto-appariements), à condition qu'ils aient un bit-score ≥80 et qu'ils se situent dans un intervalle de 99 % du bit-score du premier résultat (algorithme MEGAN-like (Cornet et al., 2018)). Les supermatrices correspondant à chaque sélection de TQMD ont été assemblées à partir des protéines ribosomales. Brièvement, les séquences ont été alignées avec MAFFT v7.453 (Katoh and Standley, 2013), puis les alignements ont été filtrés en utilisant ali2phylip.pl (https://metacpan.org/dist/Bio-MUST-Core), implémenté avec le filtre BMGE (Criscuolo & Gribaldo, 2010) (min=0.3, max=0.5, bmge-mask=loose). Cette étape a permis de réduire la proportion de sites manquants dans les alignements. Ensuite, nous avons utilisé Scafos v1.30k (Roure et al., 2007) pour créer les six supermatrices différentes, en utilisant la distance évolutive minimale comme critère de sélection des séquences (threshold fixé à 25%), le pourcentage maximal de sites manquants pour une "séquence complète" fixé à 10 et le nombre maximal d'OTUs manquants fixé à 25, sauf pour les Firmicutes (22) et les Actinobacteria (20). Enfin, IQ-TREE (Hoang et al., 2018; Nguyen et al., 2015) a été utilisé pour reconstruire l'arbre phylogénomique associé à chaque supermatrice, en utilisant le

modèle LG4X avec des bootstraps ultrarapides. Les arbres ont été automatiquement annotés et colorés à l'aide de format-tree.pl (également de Bio::MUST::Core), puis visualisés avec iTOL v4 (Letunic and Bork, 2019).

e. Comparaisons des différentes sélections

La comparaison des différentes sélections les unes par rapport aux autres permet d'éventuellement identifier des divergences au sein des alignements, ou des arbres, et ainsi d'affiner le rôle des Chlamydia lors de l'endosymbiose primaire du plaste. Pour chaque arbre sélectionné de pipeline, nous avons inventorié plusieurs paramètres : le nombre de séquences et d'espèces total dans les clans, puis séparé en donneurs - accepteurs et intrus et la topologie des clans (à savoir donc le nombre de lignées d'Archaeplastida présentes). Ces données ont ensuite été visualisées avec différents packages R. La distribution de chaque paramètre au sein des différentes sélections s'est effectué grâce au package Rainclouds plot (Allen et al., 2021). La significativité des différences entre les différents pipelines a été évaluée par un test de Kruskal-Wallis (les paramètres spécifiques à chaque conditions sont inscrits sur les figures correspondantes) (figures 7 et 18).
La visualisation des données a nécessité de manière générale ggplot2 (Wickham, 2009), pour la création des diagrammes et des graphes. L'intersection des différentes sélections chlamydienne à quant à elle été évaluée par le package UpSetR (Conway et al., 2017).

f. Annotations fonctionnelles

Pour chaque clan sélectionné de chaque pipeline, nous avons récupéré les séquences des espèces cibles. Ces différentes listes de séquences ont ensuite été annotées par EggNog mapper (Huerta-Cepas et al., 2017) et BlastKoala (Kanehisa et al., 2016). Nous avons attribué une annotation à l'ensemble des clans sélectionnés par le pipeline lorsqu'au moins 50% des séquences étaient annotées de la même manière.

Nous avons également inféré la localisation de chaque protéine identifiée par le pipeline. En se basant uniquement sur les séquences d'Archaeplastida de chaque cluster sélectionné, et sur les annotations de localisation présentes dans (Van Vlierberghe et al., 2021), annotate.pl (https://doi.org/10.6084/m9.figshare.18544955.v2) a retrouvé, pour chacune de nos séquences, l'annotation correspondante. Là encore, nous avons attribué une localisation au clan sélectionné si au moins 50% des séquences avaient la même.

4. Identification des LGT spécifiques entre Chlamydia et Glaucophyta

Afin de pallier le manque de données de glaucophytes dans le clustering initial réalisé à partir des 57 eucaryotes, nous avons réorienté le pipeline sur l'identification spécifique des transferts de gènes entre les Chlamydia et ces organismes. Pour ce faire, à partir des 33

protéomes de Chlamydia sélectionnés, auxquels nous avons ajouté le transcriptome disponible de *Cyanophora paradoxa* (Price et al., 2019), un deuxième regroupement en familles de gènes a été réalisé par OrthoFinder (paramètre d'inflation = 1.5, (Emms and Kelly, 2015)). Après une première sélection des groupes orthologues sur la présence d'au moins 1 Chlamydia et 1 Glaucystophyceae, nous avons retiré de ce jeu de données tous les OG déjà identifiés dans le pipeline principal. La suite du protocole reste la même que celle précédemment décrite. Les clusters sélectionnés par classify-mcl.pl sont alors enrichis avec les cyanobactéries et le dataset de 57 eucaryotes avant d'être nettoyés par prune-outlier.pl (threshold sélectionné à 0.2 et evalue = 1e-05, https://metacpan.org/dist/Bio-MUST-Apps-FortyTwo) et alignés avec MAFFT v7.453 (Katoh and Standley, 2013). Les protéomes des 15 Archaeplastida supplémentaires (sauf *Cyanophora paradoxa* qui fait déjà partie du clustering), et les protéomes des 49+92 espèces bactériennes sélectionnés par TQMD (Léonard et al., 2021) (desquelles nous avons retiré les cyanobactéries et les Chlamydia) sont ensuite ajoutés à chaque alignement par 42. Les séquences sont alors nettoyées une seconde fois par prune-oulier.pl et ali2phylipp.pl implémenté avec le filtre BMGE (min=0.3, max=0.5, bmge-mask=loose, (Criscuolo and Gribaldo, 2010)), puis IQ-TREE (Nguyen et al., 2015) effectue la reconstruction phylogénétique de chaque groupe orthologue avec un modèle LG4X et un ultrafast-bootstrap. Treeshrink (Mai and Mirarab, 2018) et Treemer (Menardo et al., 2018) assurent ensuite le retrait des longues branches et la déréplication des séquences en prenant soin de laisser au minimum une séquence par espèce présente dans l'arbre. PhySortR (Stephens et al., 2016) se charge finalement d'analyser les arbres générés par IQ-TREE (LG4X, ultrafast-bootstrap) et de sélectionner ceux présentant une association phylogénétique d'au moins 1 Chlamydia avec au moins un Archaeplastida (Glaucophyta), autorisant 10% d'intrus ainsi que la présence minimale de 30% des espèces cibles totales dans le clan identifié. Classify-ali.pl termine finalement en distinguant les clans pour lesquels au moins 3 Chlamydia branchent avec au moins 1 glaucophyte.

English version

Introduction

1. Primary endosymbiosis of the plastid

a. Endosymbiosis and evolution

The acquisition of photosynthesis by eukaryotes marks a major turning point in the evolution of life on Earth. The ability of organisms to use electrons from water to drive the electron transport chains necessary for the reduction of CO2 and its incorporation into organic matter through light, leads to a significant atmospheric change resulting from the release of oxygen from the photolysis of water. This change has led to an important evolution of the terrestrial landscape.

Since the work of Lynn Margulis (Sagan, 1967), the scientific community agrees on the bacterial origin of the energetic organelles of eukaryotic cells. The endosymbiotic theory proposes the internalization of bacteria, which form the mitochondria and the plastid, through their degeneration and metabolic integration. Endosymbiosis is an evolutionary phenomenon described as the internalization of living organisms by others, all leading, depending on the degree of integration, to a new organelle, in the latter case we speak of organellogenesis (once targeting to the organelle of host encoded proteins is achieved). In cases of so-called "primary" endosymbiosis (to be distinguished from "secondary" or "tertiary" endosymbiosis) a bacterium is internalized by another organism. The mitochondrion originates from an endosymbiotic event produced about 2 billion years ago, between an Alphaproteobacteria and a proto-eukaryote. The plastid, on the other hand, originated from the internalization of a cyanobacterium by a unicellular heterotrophic eukaryote, and is thought to have occurred about 1.6 billion years ago (Chan et al., 2011; Rodríguez-Ezpeleta et al., 2005; Strassert et al., 2021; Yoon et al., 2004). Each of these events marked terrestrial evolutionary history, including the advent of eukaryotes and then the spread of photosynthesis in eukaryotes.

b. Primary plastid endosymbiosis and Archaeplastida

Several plastid endosymbiotic events have shaped the diversity and evolution of photosynthetic eukaryotes (Figure 1) (Archibald, 2009; Cenci et al., 2015). The earliest and most important involves the establishment of a symbiosis between a single-celled heterotrophic eukaryote and a photosynthetic cyanobacterium. The degeneration of the latter, during the process of metabolic integration, will lead to the formation of a new organelle called the plastid (chloroplasts in plants), allowing the spread of photosynthesis within eukaryotes (McFadden, 2014; Reyes-Prieto et al., 2007; Rodríguez-Ezpeleta et al., 2005). Internalized as prey by the eukaryote, most likely by phagocytosis, the ancestral cyanobacterium is then in a vacuole undergoing acidification (Ball et al., 2015). Unlike

secondary endosymbioses where remnants remain of an internalizing vacuole that may or may not be subsequently fused to the ER (van Dooren et al., 2001), the integrity of the phagocytosis membrane was not preserved during primary plastid endosymbiosis. Note, however, that the outer leaflet of the plastid outer membrane has distinct characteristics from the inner leaflet bringing them closer to eukaryotes and prokaryotes respectively (Joyard et al., 2010). These properties are shared with the mitochondria. It is not known how these endosymbionts escape phagocytosis. Symbiosis between the different organisms is only possible by establishing communicating links between the vacuole and the cytoplasm of the host cell. Two cases are to be distinguished here constituted on the one hand by the transient endosymbioses very common in all eukaryotes (Rhizobium, Frankia, zooxanthellae, insect endosymbionts...) which will not be intended to give rise to organelles (Gil and Latorre, 2019; Poole et al., 2018) and on the other hand primary and secondary plastid and mitochondrial endosymbioses of exceptionally low or even unique frequency (Archibald, 2015). In the case of transient endosymbioses, the most recent data suggest the establishment of a non-selective membrane permeability by the interaction of envelopes with host-synthesized antimicrobial peptides (Masson et al., 2016; Mergaert, 2018; Mergaert et al., 2017). This non-selective permeability allowing small molecules to pass through in a non-specific manner can of course prove cytotoxic to the symbiont and could explain in some cases its cell death (e.g. in the case of Rhizobium) (Kim et al., 2015). In a recent review, my host laboratory proposes that endosymbioses based on membrane energy metabolism requiring strict control of their permeability cannot allow for non- or low-selective permeability (under revision). The latter must therefore set up selective connectivity via transporters, and this from the earliest stages of endosymbiosis. The limiting stage of organelle-generating endosymbiosis corresponds to the establishment of selective communication between the different partners involved. Thus, the ancestral cyanobacterium, or cyanobiont, can benefit from the resources of its host, just as the eukaryote can benefit from the photosynthetic capacities of the bacterium, in an environment of controlled and regulated exchanges. The monopolization and targeting of transporters on the envelope of the symbiont is therefore a key element of the metabolic integration of the plastid.

This major evolutionary event, known as primary plastid endosymbiosis, gave rise to the group Archaeplastida (Adl et al., 2005), or Plantae if we take the classification of Thomas Cavalier-Smith (Cavalier-Smith, 1998), among which we can distinguish three different lineages: Glaucophyta (blue algae), Rhodophyta (red algae) and Viridiplantae or Chloroplastida (plants and green algae) (Figure 1). From a phylogenetic point of view, studies diverge, but seem to agree on the monophyly of the Archaeplastida, and thus on the uniqueness of the primary plastid endosymbiosis in these organisms (Irisarri et al., 2022; Moreira et al., 2000; Rodríguez-Ezpeleta et al., 2005). However, the order of branching of the different lineages diverges among studies. Most place the separation of glaucophytes first,

followed by Rhodophyta and Viridiplantae. However, this topology still does not reach consensus, as some publications identify the Rhodophyta as the first to diverge (Hackett et al., 2007; Moreira et al., 2000; Nozaki et al., 2009; Rodríguez-Ezpeleta et al., 2005; Sato, 2019).

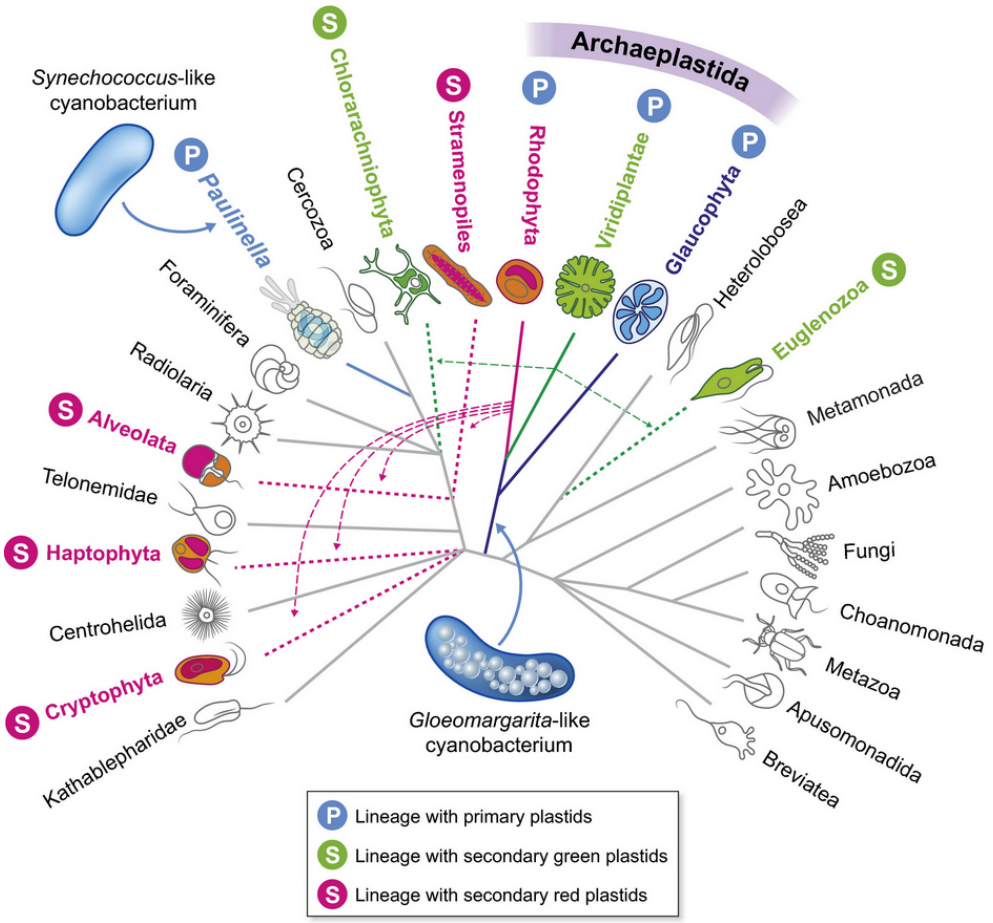


Figure 1: Distribution of photosynthesis in the evolution of eukaryotes. (taken from Ponce-Toledo et al., 2019). Photosynthetic lineages from primary endosymbiosis are indicated by colored solid lines, those from secondary endosymbiosis by dashed lines. The two known primary endosymbiosis events are marked by the blue arrows, giving rise to Archaeplastida and Paulinella. The pink and green arrows indicate secondary endosymbiosis of Rhodophyta (pink) or Viridiplantae (green). Non-photosynthetic eukaryotes are shown in gray (New Phytologist, Volume: 224, Issue: 2, Pages: 618-624, First published: 28 May 2019, DOI: (10.1111/nph.15965))

c. Other photosynthetic eukaryotes

Archaeplastida represent a large number of photosynthetic eukaryotes, and are in any case the most visible in the current terrestrial landscape. However, the spread of photosynthesis does not stop at the primary endosymbiosis of the plastid. A majority of photosynthetic eukaryotes, especially algae, are indeed derived from secondary and tertiary endosymbioses (although some may be the result of kleptoplasty (Bodył, 2018)) between Archaeplastida and other eukaryotes (Figure 1) (Archibald, 2009; Keeling, 2010; McFadden,

2001). Thus, various eukaryotic lineages have arisen as a result of internalization of red or green algae by other eukaryotes. We can count at least three main secondary endosymbiotic events, whose impact on the diversity and evolution of organisms is considerable. The internalization of two green algae led to the appearance of chlorarachniophytes and euglenes, while the endosymbiosis of at least one red alga led to the emergence of haptophytes, cryptomonadales, heterokonts, dinoflagellates and apicomplexans (Figure 1). The situation regarding red endosymbiosis appears to be much more complex, since the phylogenetic distribution of organisms with red secondary plastids is hardly compatible with a single endosymbiosis followed by a simple vertical evolution. In addition, the progressive loss of photosynthesis first, then of plastidial DNA and finally of the plastidial compartment, which is accompanied by the loss of the nuclear genes necessary for the maintenance of the different stages, is complex. These phenomena have been particularly well studied in Apicomplexa and their sister lineages. It is very difficult for these reasons to count the precise number of endosymbiotic events in particular for the red algae-derived endosymbioses. Unlike Archaeplastida, the plastids of organisms from secondary or tertiary endosymbiosis have three to four membranes. These secondary endosymbiosis species are considered complex algae. More recently, another primary endosymbiosis event leading to the acquisition of photosynthesis by eukaryotes has taken place. Totally independent of the previously described primary plastid endosymbiosis, *Paulinella chromatophora* is currently the only representative of this association (Figure 1). The metabolic integration of the two chromatophores is *a priori* not yet complete, in fact, the study of this organism would shed light on the mechanisms set up during primary plastid endosymbiosis (Gabr et al., 2020; Marin et al., 2005; Nowack et al., 2008).

d. The ancestral cyanobacteria

Although the endosymbiotic origin of the plastid during a single event is widely accepted (Moreira et al., 2000; Rodríguez-Ezpeleta et al., 2005; Sato, 2021, 2019), the identification of the ancestral cyanobacterium that causes endosymbiosis is still subject to debate. Studies indeed diverge, but seem to oppose two different origins for this cyanobiont: a very basal origin of the ancestral cyanobacterium, or an origin among more recent organisms, similar to the species in sub-section IV (Criscuolo and Gribaldo, 2011; Dagan et al., 2013; Ochoa de Alda et al., 2014; Ponce-Toledo et al., 2017; Turner et al., 1999). The latter hypothesis is based on sequence similarities between the different species, a debated process with respect to the study of primary plastid endosymbiosis. In addition, the similarity of the ancestral cyanobacterium to the Nostocales or sub-section IV cyanobacteria may be due in part to the G+C composition of the sequences. In 2011, Criscuolo and Gribaldo investigated the phylogenetic signal of cyanobacteria within the Archaeplastida, and suggested a basal ancestral cyanobacterium. Their work indeed shows a phylogenetic

branching of Archaeplastida directly after the divergence of *Gloeobacter violaceus*. These results were refined in 2017, when Ponce-Toledo et al.. placed *Gloeomargarita lithophora* at the base of the Archaeplastida for each of the conditions and datasets tested.

e. Endosymbiotic gene transfers

Like the mitochondrion, the plastid is derived from a bacterium and has its own genome. As previously mentioned, the successful integration of the plastid relies largely on the establishment of communication links between the different partners. The establishment of the active connectivity of the plastid required the establishment of an targeting mechanism to the organelle of proteins synthesized by the host. Eventually, these targeting systems allow the establishment of more complex processes such as gene transfer between the genomes of the endosymbiont and the eukaryotic host.

Endosymbiosis in general is accompanied by massive gene loss (Cavalier-Smith, 2003; McCutcheon and Moran, 2011). This is due to (i) the isolation of the endosymbiont within its host, relative to the original population of free-living organisms, prohibits repair of mutations by recombination (ii) a decrease in selective pressure to keep some functions. This concept, stated in another form by Mueller, and called Mueller's ratchet, results in the presence of many inactive pseudogenes encoding functions that are no longer needed in the intracellular environment (Moran, 1996). The fate of these pseudogenes will be irrevocably lost after deletion of the corresponding DNA segment. The concept of isolation from the natural population does not, however, apply to transient endosymbionts, capable of escaping to recombine with their free-living partners such as Rhizobium, Frankia, zooxanthellae or Wolbachia. However, it applies to many insect endosymbionts whose numbers of genes essential for symbiosis are limited to at most a few dozen functions (McCutcheon and Moran, 2011). These endosymbionts, which can also be termed transient because they are not intended, like mitochondria or plastids, to remain eternally associated with their hosts, exhibit a high degree of convergent evolution in their genome structures, resulting in tiny genomes that are as, if not more, exotic and irregularly organized than those of mitochondrial genomes (Clayton et al., 2016). **The absence of transporters in the membranes of these endosymbionts, the absence of a hybrid proteome within them, the absence of an targeting system, and finally the absence, or rarity, of horizontal gene transfers from the endosymbiont to the nuclear genome of the insect, demonstrate that despite the remarkable simplification of these genomes, this simplification does not correlate in any way with, and is not to be compared with, the emergence of a true organelle**.

The number of genes required for symbiosis in plastids and mitochondria, unlike in insect endosymbionts, is several hundred. However, the necessary presence in organelle-generating endosymbionts of systems for targeting proteins to the organelle has made it possible to share the coding of the functions necessary for symbiosis with the host genome. This allowed the

organelle genome to be further reduced. Primary plastid endosymbiosis is thus accompanied by a massive reduction in the cyanobiont genome (Cavalier-Smith, 2003). Compared to present-day bacteria, the plastid genome is less than 5% of the cyanobacterial genome size (Green, 2011; Raven and Allen, 2003). This reduction can therefore be explained on the one hand by irreversible gene loss, due to Mueller ratchet, and on the other hand by gene transfer to the host cell nucleus. This proportion of endosymbiotic gene transfer (EGT) can be explained by the host's need to control the organelle and the availability of its resources and to protect as many essential functions as possible from the harmful effect of Mueller's pawl by allowing the repair of mutations by the host's recombination processes involving, in particular, its sexual cycle. This particular form of lateral gene transfer (LGT), since the transfer takes place between organisms of different species, is characterized by a return of the produced proteins to the organelle, here the plastid. Most LGTs are involved in photosynthesis and organelle-specific maintenance functions (machinery and metabolism). Several factors can explain why gene transfers are preferred over protein translation within the plastid, for instance it could be related to stress response (Allen, 2017), or the balance of energy costs required for both phenomena. Indeed, for the latter, according to Steven Kelly, 2021, from an energetic point of view, it is more advantageous for the cell to transfer the gene into its own nucleus and then to translate the protein in the cytoplasm, rather than leaving this function to the plastid. In green plants, more than 1000 proteins are present in the plastid, for less than 100 genes in the plastid genome (Vries and Archibald, 2018). Depending on the study, between 600 and 5000 genes are thought to have been transferred from the ancestral cyanobacterium to the host eukaryotic genome during primary plastid endosymbiosis (Martin et al., 2002; Price et al., 2012). However, cyanobacteria were not the only contributors to the photosynthetic eukaryotic genome. Surprisingly, the second most important source of gene transfer in Archaeplastida comes from Chlamydia (Price et al., 2012; Qiu et al., 2013; Stephens et al., 1998). Several studies indeed establish a listing of Chlamydial genes found in Archaeplastida, going so far as to question the primary plastid endosymbiosis as involving only the host and the ancestral cyanobacterium in favor of a view centered on at least three distinct organisms involved in a common symbiosis (Ball et al, 2015, 2013; Becker et al., 2008; Cenci et al., 2017, 2016; Huang and Gogarten, 2007; Moustafa et al., 2008).

2. Ménage à Trois Hypothesis

a. Description and evolution of Chlamydia

A recent hypothesis challenges our view of the mechanism of photosynthetic acquisition in eukaryotes by invoking the involvement of a third partner during primary plastid endosymbiosis. This hypothesis, called the Ménage à Trois Hypothesis (MATH),

proposes the direct involvement of a Chlamydia-like pathogen (Ball et al., 2015; Facchinelli et al., 2013).

Chlamydiae are obligate intracellular pathogens for which there are 16 different families that can be generally grouped into two main ones: Chlamydiaceae and environmental Chlamydia. The latter are mainly isolated from environmental samples. Unlike Chlamydiaceae, which specifically infect metazoans, environmental Chlamydia have a broader spectrum of infection (Omsland et al., 2014). The hosts for these pathogens are quite variable among various invertebrates, fish, mammals, and protists. Their obligate intracellular lifestyle results in significant genome reduction (Henrissat et al., 2002) and the development of survival mechanisms within the host cell. The secretion of virulence factors by the type 3 secretion system, associated with the translation and targeting of specific transporters, allows these pathogens not only to ensure their protection against the host cell's defenses, but also to monopolize its energy metabolism, notably via the import of ATP and glucose-6-phosphate. The exit of Chlamydia from the host cell results in host cell death via cell lysis (AbdelRahman and Belland, 2005; Omsland et al., 2014).

The human pathogenicity of some Chlamydia, particularly *Chlamydia trachomatis*, has accelerated studies on these organisms. In the late 1990s, the sequencing of the *C. trachomatis* genome was published (Stephens et al., 1998). Among the 894 genes that make up its genome, 35 are identified as being of eukaryotic origin. The first analyses deduced a phenomenon of horizontal gene transfer between these pathogens and their eukaryotic hosts. However, phylogenetic studies show a proximity of these genes not to the target animals of Chlamydiaceae in general and *C trachomatis* in particular but to photosynthetic eukaryotes (Ball et al., 2013; Becker et al., 2008; Collingro et al., 2011; Huang and Gogarten, 2007; Moustafa et al., 2008). The evolution of the phagotrophic to phototrophic state of eukaryotes, following primary plastid endosymbiosis, results in a thickening of the cell wall, with the result that the membranes are no longer exposed to Chlamydia infection. In fact, no infection of Archaeplastida by Chlamydia has been reported to date. Surprisingly, chlamydial genes were found in Viridiplantae first, and then in Rhodophyta when the genome of *Cyanidioschyzon merolae* became available. Thus, these LGTs between pathogens and photosynthetic eukaryotes impact two of the three lineages that make up the Archaeplastida. This, combined with the inability of Chlamydia to infect plant cells, thus leads to the determination of the timing of these LGTs in the common ancestor of Archaeplastida.

Following these first discoveries, several studies focused on the relationship between Chlamydia and Archaeplastida, first listing gene transfers between the different groups. In 2007, Huang and Gogarten are the first to evoke the potential involvement of Chlamydia in primary plastid endosymbiosis. The following year, Moustafa et al. and Becker et al. made similar proposals based on similar observations. Although the first studies on the subject outline the contours of the Ménage à Trois hypothesis, it is the study of the metabolism of reserve polysaccharides in Archaeplastida that suggests a chlamydial contribution to the very

heart of the symbiotic process. Indeed, a previous study of the evolution of the metabolism of reserve polysaccharides in Archaeplastida allowed us to reconstruct the probable state of this metabolism at the initiation of the endosymbiotic process (Deschamps et al., 2008). This reconstruction is based on three hypotheses: 1°) the monophyly of Archaeplastida, 2°) the exclusively cytosolic compartmentalization of this metabolism, 3°) the very rapid loss of reserve polysaccharide metabolism in the symbiont. The very strong argument supporting these hypotheses can be found in numerous reviews (Ball et al., 2013; Baum, 2013; Cenci et al., 2017; Facchinelli et al., 2013; McFadden, 2014). We reproduce its originally proposed form in the accompanying diagram (Figure 2).

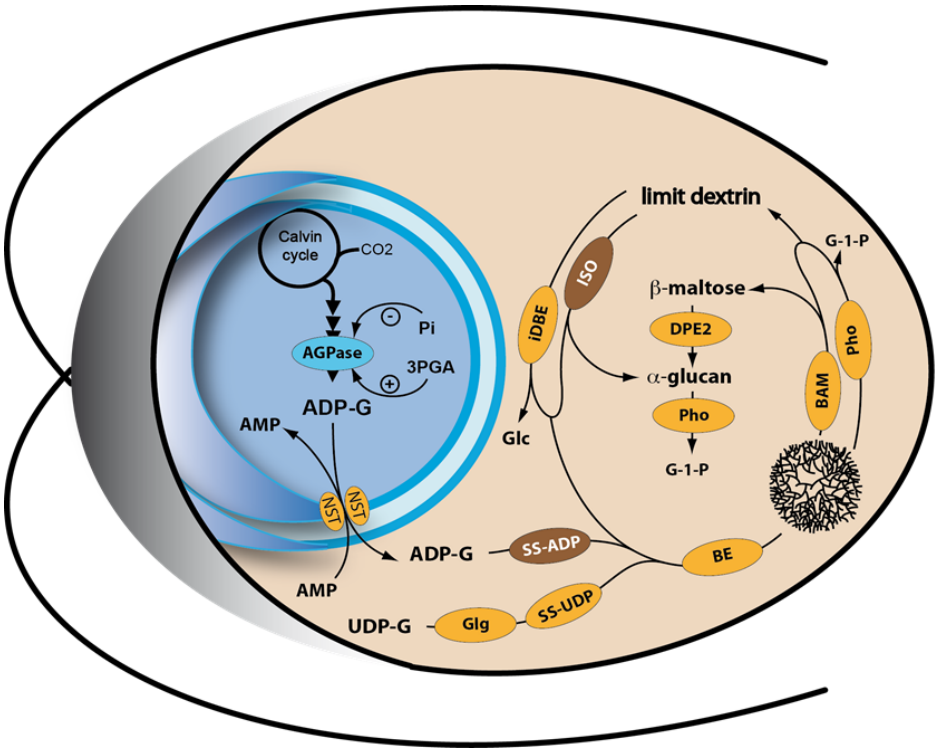


Figure 2: Endosymbiotic flows in the ancestor of Archaeplastida. (from Deschamps et al., 2008). In the cyanobiont, ADP-glucose pyrophosphorylase (AGPase) responds to the availability of photosynthetic carbon by synthesizing ADP-glucose (ADPG) which in cyanobacteria is normally destined for glycogen synthesis. The nucleotide-sugar is transported by a nucleotide-sugar/triose-phosphate (NST) translocator that originates from the host endomembrane system, as proposed in Weber et al. 2006. ADP-glucose is polymerized into glycogen without any interference with host pathways. The synthesis involves a soluble starch/glycogen synthase (SSADP) requiring ADP-glucose. The chains produced are branched by a branching enzyme (BE) and then incorporated into glycogen. Independent of photosynthesis and cyanobiont, the host is still able, like most eukaryotes, to supply glucose storage through the use of a soluble starch/glycogen synthase requiring UDP-glucose, initiated by glg (glycogenin). Mobilization of glucose from starch will be entirely dependent on host requirements through host enzymes that include phosphorylases (Pho), b-amylases (BAM), and a maltose-specific α-1,4 glucanotransferase (DPE2). The phylogenetic origin of each enzyme is represented either by a blue color (proven cyanobacterial origin), a brown color (bacterial origin), or a beige color (host origin). The cyanobiont is represented in blue (from Mol Biol Evol, Volume 25, Issue 3, March 2008, Pages 536-548, https://doi.org/10.1093/molbev/msm280)

The metabolic flux begins in the symbiont with the synthesis of ADP-glucose, a sugar nucleotide solely and exclusively dedicated to glycogen synthesis in bacteria. This enzyme in cyanobacteria is very finely coupled to the Calvin cycle and to photosynthesis through substrates and regulatory effectors (3-PGA) and inhibitors (Pi). It is now found in all the plastids of green algae and terrestrial plants and presents a phylogeny rooting it just next to *Gloeomargarita lithophora* (Deschamps, unpublished). Its maintenance in the symbiont would not make sense if the ADP-glucose produced was not used. However, it was proposed that the symbiont had become unable to accumulate glycogen, as is the case in all endosymbionts documented to date. To resolve this paradox, the authors of this hypothesis proposed the existence of a host-encoded ADP-glucose transporter responsible for the efflux of photosynthetic carbon into the cytosol. Indeed, at the time of this study, it was shown that all transporters exporting various sugars from plastids to the cytosol of green and red algae, green plants and diatoms belong to the same family called PPT. Their origin is also unquestionably monophyletic (Colleoni et al., 2010; Weber et al., 2006). They are derived from transporters of the eukaryotic endomembrane system, including transporters of GDP-mannose, which is a structural analogue of ADP-glucose. In addition, some of these green plant transporters had also been shown to have the property of integrating into mitochondrial membranes when expressed in yeast, without requiring a targeting sequence (Loddenkötter et al., 1993). The proposal of the presence of such an ancient transporter, although not demonstrated, was therefore plausible, although highly speculative. This speculation was further strengthened by the *in vitro* demonstration of the efficiency of current GDP-mannose transporters for ADP-glucose transport. Once in the cytosol of the host, the use of this sugar nucleotide was a priori impossible since this metabolite does not exist in any eukaryote. However, despite the fact that all the genes involved in the eukaryotic metabolism of cytosolic glycogen are present in the reconstruction of the metabolism of the reserve polysaccharides of the common ancestor of Archaeplastida, it also required the presence of two additional bacterial genes. These had to express two activities in the cytosolic compartment of the host, which is the only place where the substrates and products of the corresponding enzymes were found. Initially, the genes encoding these functions were thought to be of cyanobacterial origin, based on old rudimentary phylogenetic analyses and insufficiently representative sequence samples. These genes encoded on the one hand a glycogen synthase that uses ADP-glucose, and not UDP-glucose as in eukaryotes, and on the other hand an isoamylase that we know today to be involved in the evolutionary transition from glycogen to starch. In their reconstruction, Deschamps et al. had correctly seen that glycogen synthase, since it incorporated the entire surplus of ADP-glucose due to photosynthesis of the endosymbiont into the carbon reserves of the host, materialized the molecular reality of the symbiotic link between the symbiont and its host. An unexpected

consequence of the proposed symbiotic flux was that it resolved the asynchrony between photosynthetic carbon supply and demand at the time of endosymbiosis, whereas there was no metabolic integration and cross-regulation between the partners to resolve this asynchrony. Indeed, the symbiont could continue to divert a fraction of the photosynthetic flux to its reserves without any negative impact on its physiology (as all free-living cyanobacteria do) and at a time when the eukaryotic host would not have an immediate need for it, whereas the eukaryotic host could mobilize the extra reserves available in its cytosol using its own glycogen mobilization network, for example in the dark, or at any other time when the symbiont could not provide the necessary carbon. The only negative impact on the endosymbiotic partnership would have been the disappearance of the glycogen pool in the bacterial compartment. Nevertheless, we know that free cyanobacteria mutants lacking glycogen are perfectly viable in continuous light but suffer from ATP deficiency in the dark. It should be noted that it is now established that the three lineages resulting from endosymbiosis host, in the internal envelope of their plastids, ATP importing proteins of indisputable chlamydial phylogeny.

The reconstruction proposal described above was initially well received. Their authors later realized that phylogenies pointing to a cyanobacterial origin of the enzymes had to be rediscussed in the light of additional data. Indeed, a cyanobacterial origin can still be rejected for these genes while a strong phylogenetic signal links them to Chlamydia, even if a small group of Proteobacteria cannot, in particular for glycogen synthase, be totally rejected as a possible source. Uncertainties of this type are nevertheless common in single gene phylogenies. To explain the presence of Chlamydial enzymes, Ball et al., 2013, proposed that these enzymes were effectors secreted by the Chlamydial type III secretion system into the host cytosol to manipulate carbon flux. Several groups of microbiologists studying the infection cycle of Chlamydia in animals have found this proposal appealing and proved in vitro and in vivo that all enzymes of Chlamydial glycogen metabolism are, indeed, metabolic effectors of their infection cycle (Ball et al., 2013; Gehre et al.; Lu et al., 2013). Such a hypothesis of involvement during endosymbiosis, if verified, would result in all three genomes being interdependent in establishing symbiotic flows. This very quickly termed "ménage à trois" hypothesis (MATH) explains in detail the reasons for the long conservation of a chlamydial pathogen/symbiont in the cytosol of the common ancestor of the Archaeplastida and the importance of the resulting phylogenetic signal in the genomes of its current descendants.

b. Mechanism of action involved in MATH

Historically, there are two different versions of the Ménage à Trois Hypothesis. The first proposes the prior infection of the heterotrophic eukaryote by Chlamydia, followed by

phagocytosis of the cyanobacterium (Facchinelli et al., 2013). Immediate escape of the symbiont from the vacuole is then required to initiate endosymbiosis. The second version involves simultaneous internalization of both bacteria, which would then be protected in the chlamydial inclusion vesicle (Facchinelli et al., 2013). In this second version, the proximity between the bacterial partners would lead to a facilitation of conjugation and gene transfer phenomena, thus postponing the escape of the future plastid to a more distant future. As analyses and sequencing data became available, the second version of the hypothesis was favored and will be detailed in the remainder of this manuscript. The reason for this choice lies in the interpretation made by Gehre et al. of a set of experimental results concerning the infection of human or animal cells by Chlamydiaceae. Their approaches were based on the first reported use of mutants of intracellular bacteria obligate for Chlamydia on the one hand and on the examination of the consequences of "gene silencing" constructs for the host cell on the other hand. These experiments aimed at understanding the mechanisms of glycogen accumulation in the inclusion vesicle of the pathogen. This study resulted in proposing a detailed model of how current metabolic fluxes involving reserve polysaccharides and chlamydial effectors work as shown in Figure 3. The study of *Chlamydia trachomatis* and *Chlamydia muridarum* was favored because these pathogens are the only ones where glycogen accumulation in the vesicle reaches levels sufficient for quantification by cytological staining. The "environmental" Chlamydiales differ from the Chlamydiaceae in two aspects of glycogen metabolism. The first is the effector nature of all the enzymes involved, including ADP-glucose pyrophosphorylase, whereas the latter is not effector in Chlamydiaceae. The second is a marked preference of the Chlamydiaceae glycogen synthase for UDP-glucose, whereas the corresponding enzymes of other Chlamydiales are unable to use this substrate and show absolute selectivity for ADP-glucose.

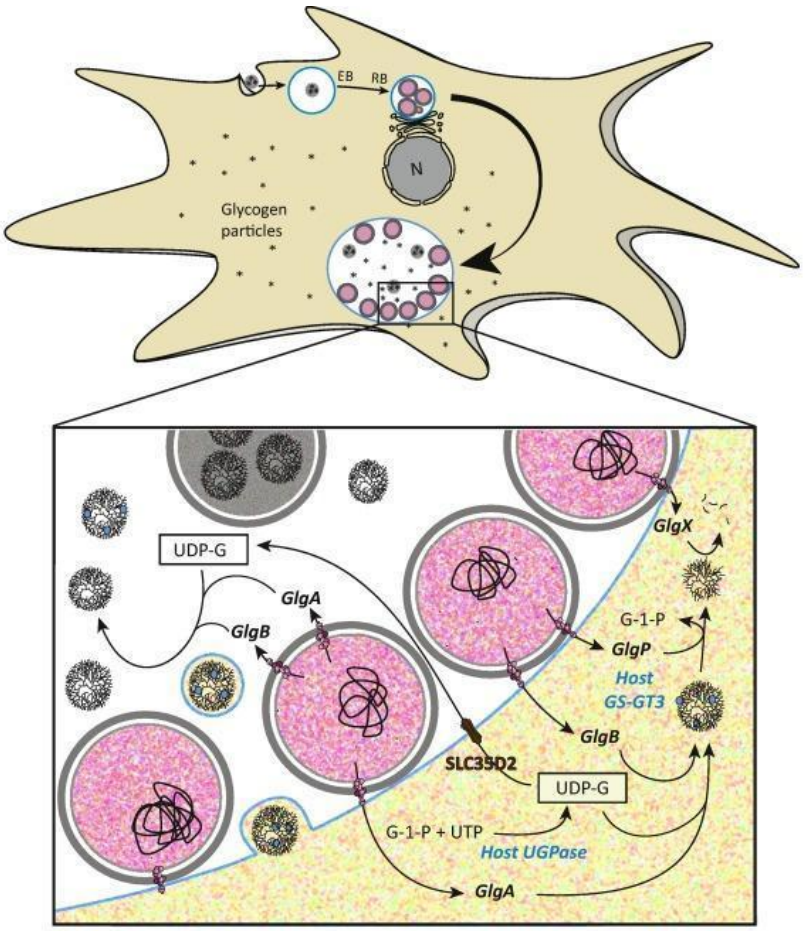


Figure 3: Glycogen metabolism in *Chlamydia trachomatis*. (Taken from Cenci et al., 2017) The reticulated bodies (RBs) of *C. trachomatis* are shown in pink. These actively replicating bacteria secrete, as effector proteins, the enzymes of bacterial glycogen metabolism into the inclusion vesicle and cytosol via the T3SS (shown in pink on inclusion-facing (white) or cytosol-facing (beige) bacterial envelopes). GlgA (glycogen synthase) elongates glucose chains by α-1,4 linkages from activated sugar-nucleotides (UDP-Glc or ADP-Glc), and GlgB (branching enzyme) introduces the α-1,6 branches into glycogen. GlgC (ADP-glucose pyrophosphorylase), the enzyme responsible for the synthesis of ADP-Glc, a bacterial-specific substrate, is not secreted by pathogens and remains within the bacteria. Glycogen synthesis in the inclusion occurs by two pathways. A minor pathway is the budding import of host glycogen vesicles from the cytosol (represented by black particles with bound host glycogen synthase (GS-GT3) represented by blue circles). The major pathway of glycogen synthesis in the inclusion is dependent on chlamydial effectors. Chlamydial glycogen synthase is able to utilize UDP-Glc which is imported into the inclusion by SLC35D2, a human UDP-Glc transporter recruited to the inclusion membrane. In the cytosol, glycogen metabolism involves both the host and chlamydial effector enzymes. Host cytosolic UDP-glucose pyrophosphorylase defines the only enzyme responsible for UDP-Glc synthesis used in both the inclusion and cytosol. Bacterial GlgC (ADP-glucose pyrophosphorylase) is thought to direct the synthesis of a portion of glycogen in elementary bodies (EBs, shown in gray) when RBs redifferentiate into EBs. Chlamydial effectors of glycogen degradation (GlgP (glycogen phosphorylase) and GlgX (glycogen disconnection enzyme)) are also secreted into the cytosol and inclusion vesicle. GlgP degrades the outer chains of glycogen grains while GlgX disconnects the remaining particle to allow further digestion by GlgP. This process produces short malto-oligosaccharides (represented by small black lines) that can only be used by pathogens. Unlike most prokaryotic glycogen synthases, the *C. trachomatis* GlgA enzyme efficiently utilizes UDP-Glc and ADP-Glc. (Cenci et al., 2017)

Extrapolating the experimental results obtained in Chlamydiaceae to the other Chlamydiales, (Cenci et al., 2017) proposed for all Chlamydiales the metabolic fluxes detailed in Figure 4 (panel A). Note that in Chlamydiales the UDP-glucose translocator would necessarily have been substituted by an ADP-glucose transporter whose origin we have already discussed. It is therefore easy to imagine the simultaneous internalization of a cyanobacterium and a pathogen (Figure 4 panel B). In the manner of the biotic interactions characterizing the infection of plant cells by *Agrobacterium tumefasciens*, it is possible to propose that Chlamydiales have evolved mechanisms to manipulate the metabolic flux of cyanobacteria to their advantage. In terms of carbon metabolism, the transmission by conjugation of genes encoding key transporters would have allowed the efflux of photosynthetic carbon while mitigating the disorders caused by this efflux. Such transporters are UhpC, a chlamydial protein responsible for carbon efflux in the form of Glucose-6-P still present today in glaucophyte plastids where it is the only possible responsible for photosynthetic carbon efflux, and NTT, an ATP importer present in all primary plastids that would have allowed the cyanobacteria to survive the ATP deficiency caused by this efflux in the dark. This export of glucose-6-P would have fed glycogen synthesis in the lumen of the chlamydial inclusion to the benefit of the pathogen. It would only be in the case of excess synthesis of ADP-glucose in the inclusion that the latter would be exported from the vesicle to the benefit of the host, thanks to the reversal of the flow of ADP-glucose import on the inclusion membrane (Figure 4 panel B). Phylogenetic dissection of tryptophan metabolism by Cenci et al. 2016, highlighted a significant contribution of Chlamydiales to this metabolism impacting 3/4 distinct enzymes out of the 7 that are responsible for the synthesis of this amino acid. The role of tryptophan seems to be central for the replication of Chlamydia. Indeed, two anti-chlamydial defense mechanisms have been documented in eukaryotes, one in mouse cells and the other in human cells. These two totally different mechanisms nevertheless lead to the same final effect: the selective deficiency of tryptophan in the infected cell, indicating a fierce competition between the host and the pathogen for this amino acid. Based on these observations, Cenci et al. (2016) proposed that the chlamydial genes required for tryptophan biosynthesis found today are the remnants of an ancient interaction consisting of the transfection of an entire chlamydial operon in the cyanobacterium and generating tryptophan overproduction. This proposal is particularly reinforced by the presence in the inner membrane of primary plastids of a chlamydial tryptophan transporter. Such a transporter would have allowed the efflux of tryptophan into the chlamydial inclusion, explaining the purpose of the gene transfer involved in tryptophan metabolism. Considering that only half a dozen or less chlamydial transporters have been counted on the inner membrane of the plastid, it is difficult to accept that the presence of this transporter is due to a mere coincidence rather than a remnant of an ancient biotic interaction. The importance of conjugative gene transfer in biotic interactions at the heart of the MATH hypothesis led scientists proposing this hypothesis to prefer the simultaneous arrival of endosymbiotic

partners in a common inclusion, thereby facilitating conjugative transfer. According to the authors, this interaction would have been of the same type as the one induced today by *Agrobacterium*, but differs in that it involves a pathogenic bacterium and a cyanobacterium in an intracellular environment, and not a pathogenic bacterium and plant cells in an extracellular environment. This transient endosymbiotic interaction does not, again according to these authors, signal the onset of primary plastid endosymbiosis, but it would have ensured its success by bringing the cyanobacterium to an unprecedented level of pre-adaptation. Indeed, this autotrophic bacterium, originally free and autonomous and therefore poorly endowed with transporters, is now enriched by a chlamydian arsenal of proteins on its envelope ensuring its connectivity in the intracellular environment. It will be the accidental escape of the cyanobiont from the chlamydial vesicle by hemifusion between the inclusion membrane and the outer membrane of the cyanobacterium that signs, according to the authors of MATH, the beginning of the plastid endosymbiosis proper. The connectivity between the different partners and their different metabolisms thus define a primordial aspect of the establishment of endosymbiosis.

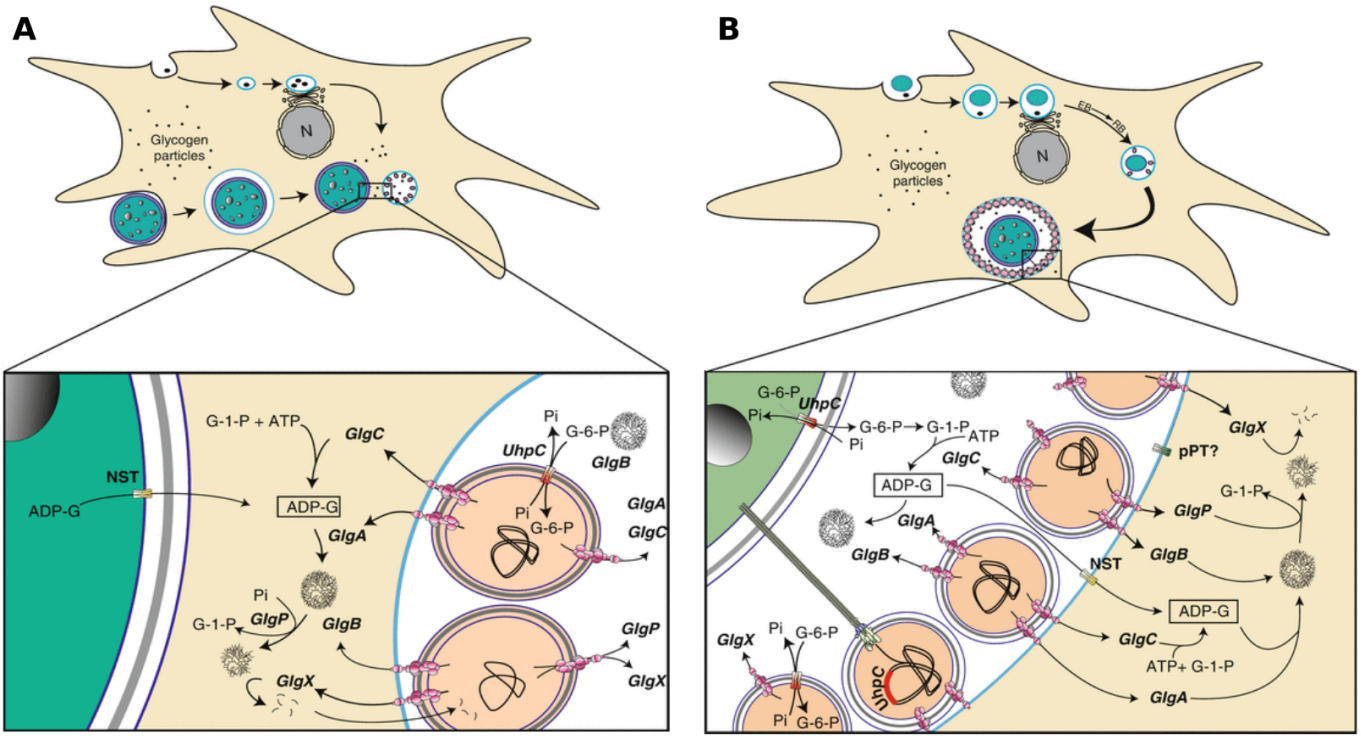


Figure 4: Two alternative scenarios explaining the Ménage à Trois hypothesis and their metabolic implication (taken from (Ball et al., 2015). (A) The classic Ménage à Trois Hypothesis (first NST scenario). Panel (A) summarizes the scenario previously detailed in Figure 3. The large cyanobiont (blue-green with starch granules displayed) is shown entering the host independently of the chlamydial pathogen (to scale), a single thick black dot. A section showing the tripartite interaction is enlarged and boxed. Chlamydial reticulate bodies are shown attached by their TTS (type 3 secretion system, shown in pink) to the inclusion vesicle membrane (light blue). Glycogen particles are shown in black inside the inclusion vesicle and in the host cytosol. Only enzymes of chlamydial origin are shown and abbreviated by their genetic symbols: GlgA glycogen synthase, GlgB branching enzyme, GlgC ADP-glucose pyrophosphorylase, GlgP glycogen (maltodextrin) phosphorylase , and GlgX GlgX direct DBE type. The NST (ADP-glucose transporter) is highlighted on the cyanobiont

membrane which was achieved independently of a non-existent TIC-TOC protein import mechanism. (B) The modified Ménage à Trois Hypothesis (first UhpC scenario). The cyanobiont is shown entering with a chlamydial elementary body. The elementary body differentiates into reticulate bodies attached by their TTS to the modified phagocytic vacuole, preventing phagocytosis of the cyanobiont. T4SS (type four secretion system shown in green, blue, and black) responsible for conjugative DNA transfer transfers UhpC and NTT (ATP import protein) genes from chlamydia ad minima to be expressed in the cyanobiont. UhpC is shown in red on the inner membrane of the cyanobiont, while NST is shown in yellow on the inclusion vesicle. The low-affinity ADP-glucose transporter (NST) transports excess carbon assimilation to the cytosol where it will be metabolized.

c. Menaquinone synthesis in Archaeplastida

Depending on the study, and the stringency of the methods applied, estimates of the number of chlamydial genes in Archaeplastida range from 30 to 100 (Ball et al., 2013; Becker et al., 2008; Collingro et al., 2011; Huang and Gogarten, 2007; Moustafa et al., 2008). These transfers are distributed among photosynthetic eukaryotes, impacting up to the three lineages derived from plastid endosymbiosis. These genes do not appear to be randomly distributed, but rather restricted to certain metabolic pathways. The first metabolic pathway identified as being impacted would be the glycogen pathway described above. These early studies, as well as the aforementioned analysis of tryptophan metabolism, appear to favor the pathway of sophisticated biotic interactions via conjugative transfers of genes encoding entire metabolic pathways and associated membrane transporters, with metabolic connectivity to the host provided by secretion of metabolic effectors (Cenci et al., 2017). If such transfers existed, it is logical to expect to find traces of them in present-day plastid genomes. Indeed, recent studies show the impact of conjugative transfers of chlamydial genes to plastids on menaquinone (vitamin K) synthesis pathways (Cenci et al., 2018).

Vitamins K are a group of lipid-soluble vitamins composed of a naphthoquinone ring attached to a poly-isoprenyl chain of variable length and saturation. In vertebrates, they have chelating properties and are involved in several molecular processes including, but not limited to, blood clotting, bone metabolism, and cell signaling (Vos et al., 2012). In photosynthetic organisms, these vitamins are associated with photosystem I, participating in electron transport (Reumann, 2013). Two forms are mainly found: vitamin K1, also called phylloquinone, present in majority in plants (Oostende et al., 2008), green algae (Lefebvre-Legendre et al., 2007) and cyanobacteria (Collins and Jones, 1981), and vitamin K2 or menaquinone, in archaea, bacteria (Collins and Jones, 1981), Rhodophyta (Yoshida et al., 2003) and diatoms (Ikeda et al., 2008). Unlike bacteria and archaea, animals and most protists do not possess the pathway to synthesize these vitamins, yet they are necessary for cell survival, and therefore must have an external supply (Li, 2016). The main difference between the different molecules lies in the saturation of the poly-isoprenyl chain (Figure 5A); the presence of double bonds indeed impacts their diffusion properties and thus their biochemical capabilities. These membrane-bound polyunsaturated isoprenoid quinones undergo a reversible two-step reduction, which gives them the ability to conduct electrons

between different protein complexes, such as those involved in photosynthesis or respiration (Nowicka and Kruk, 2010). Under aerobic conditions, ubiquinone and plastoquinone are the predominant molecules. However, menaquinone (but not phylloquinone) takes over during the presence of low amounts of oxygen, both for anaerobic respiration in bacteria and for oxidative phosphorylation under microaerophilic conditions in bacteria and metazoans (Li, 2016; Sharma et al., 2012). Reduction of the unsaturated isoprenoid chain of menaquinone forms phylloquinone, whose side chain is partially saturated.

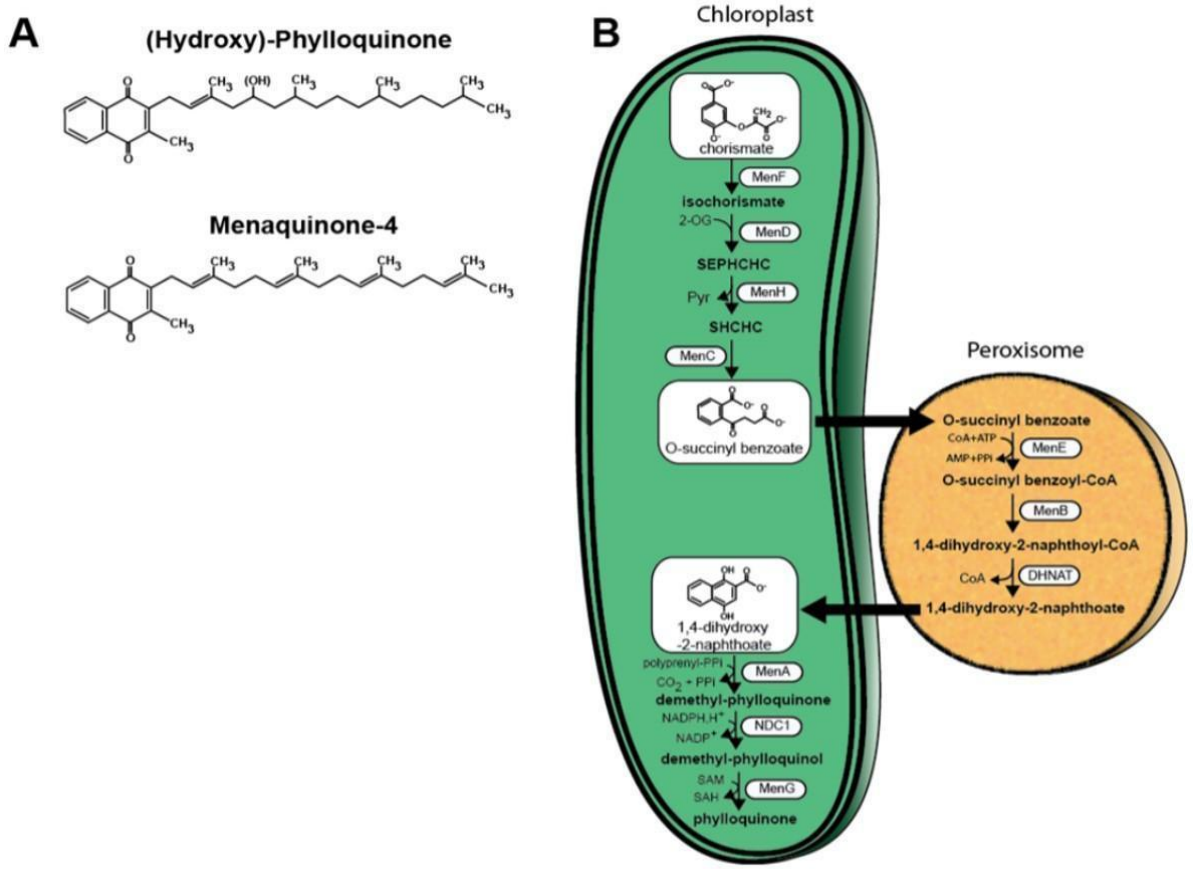


Figure 5: Structure and synthesis of Vitamin K in Viridiplantae (Adapted from Cenci et al., 2018). A, Structure of Vitamin K1 (phylloquinone) and Vitamin K2 (menaquinone). B, Synthesis of phylloquinone in plants. Chorismate is converted in four steps in the chloroplast to O-succinyl benzoate (OSB), by a protein called "Phyllo" representing the fusion of the protein products of the bacterial genes MenF, MenD, MenC and MenH of the menaquinone synthesis pathway. OSB is then diffused to the peroxisome where it is converted to DHNA (1,4-dihydroxy-2-naphthoate) in three steps catalyzed by MenE, MenB, and DHNAT, and then returns to the plastid. The polyprenylated isoprenoid chain is added to DHNA by MenA, and then the DHNA ring is reduced by Ndc1 (NADPH dehydrogenase) and methylated by MenG.

Two metabolic pathways allow for the conversion of chorismate to menaquinone or phylloquinone in Bacteria and Archaea (Men and Futalosin pathways) (Zhi et al., 2014). Cyanobacteria and plants, on the other hand, have only the classical Men pathway (Figure 5B). In plants, the first four steps of this synthetic pathway are carried out in the plastid by the protein "Phyllo" corresponding to the fusion of the four enzymes required for the conversion of chorismate to o-succinyl benzoate, which is then transported into the

peroxisome to form the naphthoquinone ring, before returning to the plastid where it is prenylated by MenA, reduced by Ndc1, and methylated by MenG before being associated with the PSI (Eugeni Piller et al., 2011; Reumann, 2013) (Figure 5B).

From a phylogenetic point of view, the origin of the Men pathway remains unclear. As this metabolic pathway is present in bacteria but absent in most eukaryotes, it may have been passed on to the common ancestor of Archaeplastida during primary plastid endosymbiosis. An initial analysis from 2008 raises the possibility of a non-cyanobacterial origin of the majority of the enzymes involved (Gross et al., 2008). A recent study published at the end of 2018 takes up the topic using newly available genomic data, and shows that all of the 7 genes constituting the cluster responsible for menaquinone synthesis, encoded within the plastid genome, notably in Cyanidiophytina, are of chlamydial origin (Cenci et al., 2018). This observation is supported by single gene phylogenies and concatenation of MenF and MenD, and supports the hypothesis of pathogen involvement during primary plastid endosymbiosis. It is also important to note that MenF and MenD could be affiliated with the phyllo protein found in red and green algae in the host nuclear genome. However, the phylogenetic argument supporting the affiliation is complex and would deserve to be reviewed in light of the different evolutionary patterns of the nuclear and plastid genomes. This analysis is all the more important because if the affiliation were to be confirmed, it would place the transfer of the Men cluster before the diversification of red and green algae. Given the absence of the Men pathway in eukaryotes other than Archaeplastida and their derivatives, and the rarity of LGTs to the plastid genome, it is plausible that the entire cluster responsible for menaquinone synthesis was transmitted at once from Chlamydia to the plastid ancestor. This observation is of considerable importance in that it suggests the reality of the conjugative transfers required in the biotic interactions at the heart of MATH. It was all the more unexpected to observe such relics since most plastids have reduced the size of their genome to the point of leaving only the genes responsible for coding components of the electron transport chains of photosystems and the proteins necessary for their translation, which are precisely not involved in the conjugative transfers and biotic interactions of MATH. According to the authors of MATH, it would not be a coincidence that they could be observed in the Cyanidiophytina since the plastid genomes of these organisms are the largest among the Archaeplastida, having conserved the greatest number of original cyanobacterial functions not included in the photosystems or the translation apparatus.
Interestingly, this study also shows the absence of part of the Men pathway, upstream of DHNAT, in Glaucophytes, one of the three groups derived from primary plastid endosymbiosis, and *Galdieria sulphuraria* (Cyanidiophytina), whereas biochemical analyses performed in *Cyanophora paradoxa* attest to the presence of hydroxyphylloquinone and thus to the ability of these organisms to synthesize it. Some cyanobacteria show the same

characteristics: Gloeobacteria and *Gloeomargarita lithophora* have neither the Men nor the Futalosin pathway, but still synthesize vitamin K

This biochemical similarity leads to the question of a possible common origin of the menaquinone synthesis pathway in these organisms. Identifying this alternative pathway first in Gloeobacteria and *G. lithophora*, then in Glaucophytes, would allow, if they are identical, to hypothesize the selection of this metabolic pathway during the separation of Archaeplastida, and thus to specify the nature of the cyanobacterium at the origin of the plastid.

3. A Controversial Hypothesis

a. MATH, a supported and controversial hypothesis

The MAT hypothesis is supported by molecular, biochemical and phylogenetic analyses. The involvement of Chlamydia during primary plastid endosymbiosis would therefore allow us to draw mechanisms underlying endosymbiosis that were previously unknown. From a functional point of view, the presence of Chlamydia in the inclusion vesicle is justified by the protection of the cyanobacterium but also by the transfer of key genes allowing the establishment of the biochemical flows of endosymbiosis. Chlamydia would thus act as a connector between the two other partners, allowing the export of photosynthetic resources from the cyanobacterium to the host cell without the latter suffering. The involvement of a third partner would also explain the rarity of the event during evolution.
However, this hypothesis is controversial within the scientific community. Some studies question the timing and relevance of LGTs, while others question the interpretation of single gene phylogenies, or the actual impact of Chlamydia compared to that of other bacterial groups (Dagan et al., 2013; Deschamps, 2014; Domman et al., 2015).

b. Criticisms and methodological aspects

Because of the age of the plastid signal, whether from a purely cyanobacterial point of view or in a MATH context, the phylogenetic signal that can be visualized in single gene trees is considerably weakened and subject to significant phylogenetic artifacts. The difficulty in identifying EGTs, due to the loss of signal but also to the acceleration of the evolution of these genes following their adaptation to a new environment, leads to the need for maximum caution in the interpretation of the generated phylogenies. Moreover, the identification of these EGTs is method-dependent. The phylogenetic reconstruction performed by maximum likelihood seems to overestimate the proportion of transferred genes compared to Bayesian methods, which are *a priori* less sensitive to phylogenetic artefacts. By manually re-analyzing the gene transfers identified by Becker et al., 2008, Moreira and Deschamps, 2014, indeed

confirm only 17 of the original 55 phylogenies suggesting chlamydial gene transfer to Archaeplastida.

The use of selection tools to automate protocols and reduce interpretation bias, and thus to quantify the endosymbiotic signal as a whole, however, has some biases in itself. Indeed, these tools analyze the bipartitions of the trees and select those presenting a branching of interest, in our case for example a branching between Chlamydia and Archaeplastida. Several aspects of the analysis are therefore neglected by these approaches, such as taking into account the complete topology of the tree or the taxonomic diversity. A manual selection approach also has some disadvantages and biases, especially in the variability of individual interpretation, but allows confirmation of a gene transfer of endosymbiotic origin by taking into account the tree as a whole. The discrepancy between the selections made by the two approaches, and even by different observers, can be surprising. Indeed, following the automatic identification of gene transfers between diatoms and Archaeplastida, two different teams investigate these results by a manual approach. Combined, the manual tree analysis of the two studies recovers only 10% of the transfers identified in automatic (Moreira and Deschamps, 2014).

It is important to keep in mind that most of the studies performed on the Ménage à Trois hypothesis and the identification of gene transfers between Chlamydia and Archaeplastida date from the early 2000s (Becker et al., 2008; Huang and Gogarten, 2007; Moustafa et al., 2008). The results obtained are therefore in agreement with the genomic, proteomic and transcriptomic data available at the time. However, the taxonomic diversity taken into account in phylogenetic analyses can have a significant impact on the results obtained. Technical advances in sequencing have considerably increased the quantity (and sometimes, but not always, the quality) of available data. However, there is still an imbalance in the diversity present in the databases, especially in NCBI. Indeed, taking into account only the RefSeq prokaryote of March 2021, 93% of the available data are represented by three main phyla: Proteobacteria, Firmicutes and Actinobacteria (Léonard et al., 2021). The other 50 bacterial phyla represent only 7% of the database. This marked disproportion can also be seen in the phylogenetic analyses that have been set up, and therefore requires a sampling effort that is representative of the real diversity of life.

c. The predominant role of Chlamydia and the endosymbiotic gene transfer

One of the main criticisms brought against the Ménage à Trois Hypothesis lies in the actual involvement of Chlamydia during primary plastid endosymbiosis compared to other organisms (Dagan et al., 2013). Some indeed acknowledge Chlamydian gene transfers to Archaeplastida, but put them in perspective with the gene transfers identified for other bacterial groups, putting aside the functional aspect of the hypothesis. Dagan et al. 2013

assess the proportions of gene transfers based on different prokaryotic phyla. According to their results, Chlamydia comes in 6th place, behind Cyanobacteria, Proteobacteria, Firmicutes, Actinobacteria and Bacteroidetes. Thus, before taking into account the involvement of pathogens in the evolution of Archaeplastida, the impact of other groups should also be investigated, the chlamydial signal becoming only a part of the overall bacterial contribution.

It is useful to consider here the significance of bacterial lateral transfers in general that seem to be associated with the endosymbiotic process. The establishment of symbiosis between the cyanobacterium and the eukaryotic host cell is accompanied by a degeneration of the bacterium, resulting in a large number of gene transfers from the cyanobacterial genome. These EGTs (Endosymbiotic Gene Transfer), by definition, come from the symbiont and add new functions to the eukaryotic nuclear genome (Figure 6). However, it is possible, for various reasons, that a cyanobacterial version of an existing eukaryotic gene is preferred to it, thus effecting the "great replacement" of eukaryotic stem genes. Still in an endosymbiotic context, other bacterial sources contribute to the establishment of the symbiosis and the new eukaryotic genome. These ERGT (Endosymbiotic Related Gene Transfer) , for the most part, concern functions present in cyanobacteria but whose genes are not found as such in the eukaryotic nucleus, which has preferred a bacterial version of different origin. A plausible explanation for the preference for "immigrants" genes rather than the genes of the cyanobacterial symbiont could be found in an examination of the conditions that are required by the plastid translocation system(s) of proteins synthesized in the host cytosol. It is useful to recall here that proteins destined for the plastid are not conformed in the form of sequential folding of microdomains at the exit of the ribosome tunnel as are most bacterial and eukaryotic cytosolic proteins; nor do they follow the pathways of protein translocation and conformation to other compartments (e.g. dry pathway) (Jarvis and Soll, 2002). However, depending on their final destination, they must interact with specific chaperones that keep them denatured when they leave the cytosol and allow them to be transported and to interact with various proteins of the external and internal translocons of the plastid as well as with other chaperones that preside over their correct final conformation *in situ* (Jarvis and Soll, 2002). The primary sequences of cyanobacterial genes have not evolved over hundreds of millions of years to "bend" to this exercise. It is therefore possible that, by chance, a sequence may indeed follow the correct path, without having to undergo mutations, in which case the plastid gene will very quickly be replaced by an authentic cyanobacterial nuclear version of that gene. On the other hand, it is possible that this replacement is only possible if the sequence undergoes a number of mutations that delay the final selection of a sequence duplicated in the host nucleus or even prohibit its replacement by the original cyanobacterial sequence. In this case, the replacement of the original gene will be done more rapidly by a foreign sequence requiring fewer mutations and which will compete with the cyanobacterial gene even if the transfer of this foreign sequence was originally of much lower frequency

than that of the symbiont genes. Since the endosymbiotic partnership does not necessarily aim to reproduce the original proteome of the cyanobacterium but only to ensure the sustainability of the symbiotic relationship for the benefit of the host, this lateral transfer will be selected. **It is useful to insist here on the exclusive dependence of this selection on the physicochemical properties of the protein concerned. Two enzymes of the same synthesis or degradation pathway of the same organism will not show the same adaptability in this respect. In other words, the occurrence of two such replacements from the same bacterial source in the same metabolic pathway would be an extraordinarily rare or even non-existent event**. On the other hand, cyanobacterial genes will behave differently since it is this proteome that guides and orders the process of gene migration to the nucleus. If the MATH proposal is incorrect, then one would expect 1) that the chlamydian phylogenetic signal in the Archaeplastida genome would be of the same nature and importance as, or even less than, the other bacterial signals as asserted by Dagan et al., 2) that multiple transfers within the same metabolic pathway would therefore be non-existent.

On the other hand, if the MATH proposal is correct, as chlamydia itself becomes a symbiont, its ERGTs should by definition be considered as EGTs and their properties will approximate those of the cyanobacterial EGT signal. In particular one can expect on the one hand that ancient conjugative transfers will result in multiple signals in one pathway and on the other hand that metabolic effectors at the core of the symbiotic process will result in the same outcome.

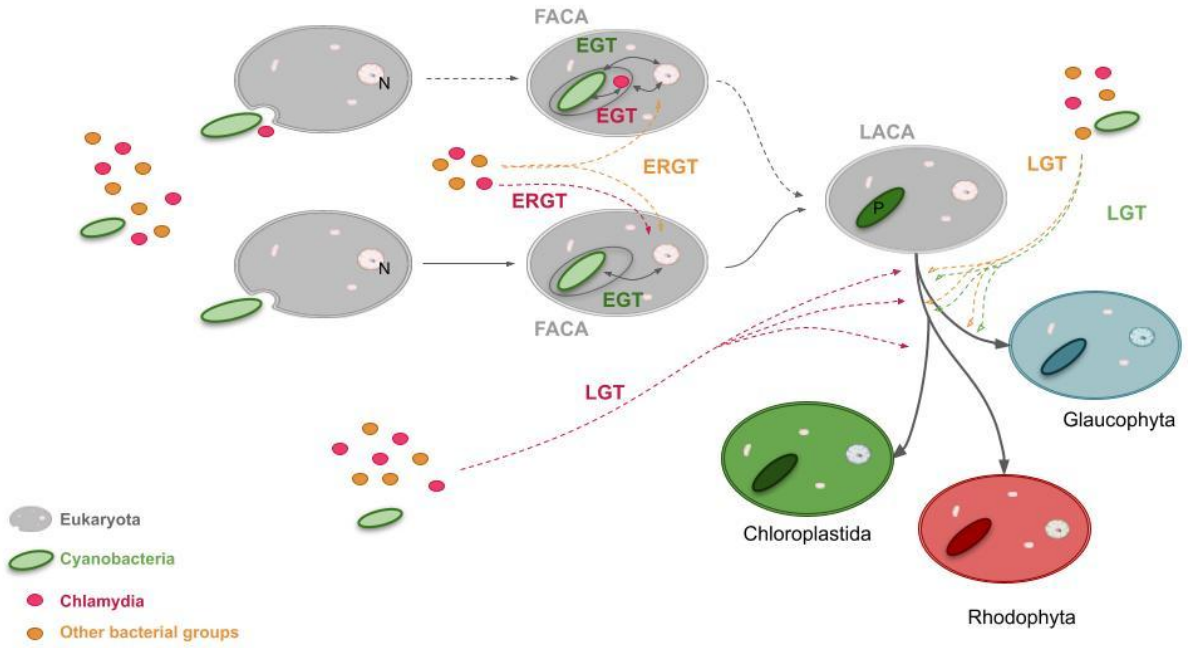


Figure 6: Horizontal gene transfers involved in the evolution of Archaeplastida. The emergence of the three lineages of Archaeplastida (Glaucophyta, Rhodophyta and Viridiplantae) follows the primary endosymbiosis of

the plastid, and the internalization of a photosynthetic cyanobacterium by a heterotrophic eukaryote (represented by the solid arrows). The first common ancestor of Archaeplastida (FACA) thus presents an internalized cyanobacterium, in the process of degeneration. At this stage, the genome of the symbiont is reduced and a large number of genes are transferred into the nuclear genome of the host cell (EGT for Endosymbiotic Gene Transfer). Bacterial contributions are multiple during the evolution of Archaeplastida (hatched colored arrows), and can occur in an endosymbiotic context, which is then called ERGT (Endosymbiotic Related Gene Transfer) but also in a later context of evolution through LGT (Lateral Gene Transfer) taking place when the plastid is integrated into the last common ancestor of Archaeplastida (LACA) and its descendants. Recently, the Threesome Hypothesis proposes the involvement of a chlamydial pathogen during plastid endosymbiosis (gray hatched arrows). Chlamydia is then a transient symbiont. Gene transfer from pathogens can therefore be of the three types described: EGT, ERGT and LGT, depending on the context and timing of transfer. Unlike EGTs, LGTs and ERGTs can be characterized as "external" contributions to the symbiosis and thus involve several bacterial sources (either at the taxonomic level or within a species group itself). Gene transfers related to the endosymbiotic context (EGT and ERGT) are more likely to be found in a large diversity of Archaeplastida since they were made during the establishment of the symbiosis before the last common ancestor of Archaeplastida.

Objectives of the project

The aim of this project is to test the Ménage à Trois Hypothesis (MATH), and thus to clarify the potential role of a Chlamydia partner in primary plastid endosymbiosis. To this end, the study will be organized in two complementary parts. First, it will examine the nature of the Chlamydial phylogenetic signal in Archaeplastida and verify that it can be interpreted as a remnant of MATH. Then, in a second step, we will have to compare this signal with other signals, chosen as controls, in order to check the specificity of the interaction, both from the point of view of the Chlamydia donor (bacterial controls) and the Archaeplastida host (eukaryotic controls). Thus we can identify two main issues to this project:

- Is there a chlamydial signal in Archaeplastida specifically related to primary plastid endosymbiosis? If so, is this signal already known from the literature and does it confirm the metabolic foundations of MATH?
- Is this chlamydial signal in Archaeplastida then different compared to other bacterial signals?

A bioinformatics protocol was therefore set up to screen lateral gene transfers (LGT) between Chlamydia and Archaeplastida, then a manual analysis of the trees generated for the selected genes allows us to reposition these transfers within the evolution of eukaryotes, thus answering the first part of the problem posed. The complete automation of the protocol then allows to quantify the signals of different control groups, both bacterial and eukaryotic, and thus to verify the uniqueness of the contribution of Chlamydia to Archaeplastida.

Results

1. Identification of the chlamydial signal in Archaeplastida

a. General approach

The general approach of the project is broadly broken down into three phases: (i) screening of LGTs between Chlamydia and Archaeplastida to establish the phylogenetic signal, (ii) manual analysis of the generated trees to verify that the identified signal may be a remnant of a primary endosymbiosis of the tripartite plastid, and (iii) full automation of the protocol, mimicking the manual analysis, in order to allow a perspective of this signal in the light of positive and negative controls, whether from a phylogenetic, functional or metabolic perspective.

The development of a bioinformatics pipeline allows the identification of LGTs between different target taxonomic groups and consists of three major steps. Based on a selection of quality genomes and proteomes, representative of the diversity of living organisms, the orthologous groups (OGs or clusters) integrating the protocol are i) first filtered on criteria of presence of the target organisms, before being ii) enriched with the totality of the selected and filtered genomic and proteomic data, then iii) the phylogenetic reconstruction of each group will lead to the selection of the trees on topology criteria. This step identifies each tree with a phylogenetic branch (clan) between the target species (here, between Chlamydia and Archaeplastida). Manual analysis of the trees selected by the pipeline confirms or invalidates the endosymbiotic character of the LGT. In parallel, the inventory of genes listed as chlamydial in Archaeplastida in the literature, as well as the comparison of this inventory with the results obtained and the manual analysis of the corresponding trees, allows the validation of the protocol. The set of LGTs identified by the pipeline and then validated by the manual analysis thus constitutes the global phylogenetic signal. From the manual analysis of the trees, as much from the selection of the pipeline as from the inventory of the literature, the bioinformatic protocol is then adjusted to take into account the observations made and thus automate the pipeline. This automatic conceptualization of the protocol mimics the results of the manual analyses and allows the comparison of the chlamydial signal in Archaeplastida to others. Two controls are then designed: the control of LGT "donor" organisms (here Chlamydia), through the study of the signal of different bacterial groups in these photosynthetic eukaryotes, and the control of LGT "acceptor" organisms (here Archaeplastida) through the study of the chlamydial signal in other eukaryotic groups. The specificity of the chlamydial role during primary plastid endosymbiosis can then be assessed by characterizing and comparing each LGT selection. This involves first quantifying the number of gene transfers, then assessing their signal congruence, and then analyzing the

diversity of organisms displaying these transfers and finally putting them into a metabolic and functional context.

b. LGT Chlamydia - Archaeplastida screen

Our study is based on the identification of gene transfers between Chlamydia and Archaeplastida which, taken together and as a whole, reconstruct in part the common evolutionary history of these organisms, which we define here as a phylogenetic signal. Thus, with the idea of having the most accurate identification of these LGTs, limiting as much as possible the proportion of false-positives due to contamination or taxonomic over-representation, we opted for the use of a quality database representative of the diversity. The quality of each proteome or genome entering the pipeline was therefore evaluated. Among the set of species selected for this part of the project are 72 eukaryotes, including 54 photosynthetic eukaryotes, 33 Chlamydia, 48 cyanobacteria, and two sets of 49 and 92 bacteria, without distinction of taxonomy, from which we have removed Chlamydia and cyanobacteria (annexe 9).

The bioinformatics protocol developed for this study has three main steps (Figure 7). A pre-selection of the generated trees, including only the 57 eukaryotes, Chlamydia and cyanobacteria, allows to put in direct competition the three supposed partners of the Ménage à Trois Hypothesis, and thus to identify the gene transfers for which the chlamydial signal clearly outweighs the cyanobacterial one. Enriching this pre-selection with the rest of the genomes and proteomes then allows us to test the robustness of the Chlamydian signal, however, only manual tree analysis validates these transfers and allows us to determine their potential endosymbiotic context.

210451 orthologous groups were created by OrthoFinder from the proteomes of 57 eukaryotes (including 412 Rhodophyta and 8 Viridiplantae). Among them, only 8583 contained at least 2 Viridiplantae and/or Rhodophyta. After enrichment of each orthologous group with Chlamydia and cyanobacteria proteomes, we retained for the rest of the protocol only those with at least one chlamydial sequence. Thus, 987 orthologous groups underwent sequence filtering and phylogenetic reconstruction. Automatic bipartition analysis revealed 201 trees with a phylogenetic clan with at least one Chlamydia and at least one Archaeplastida (Figure 7).

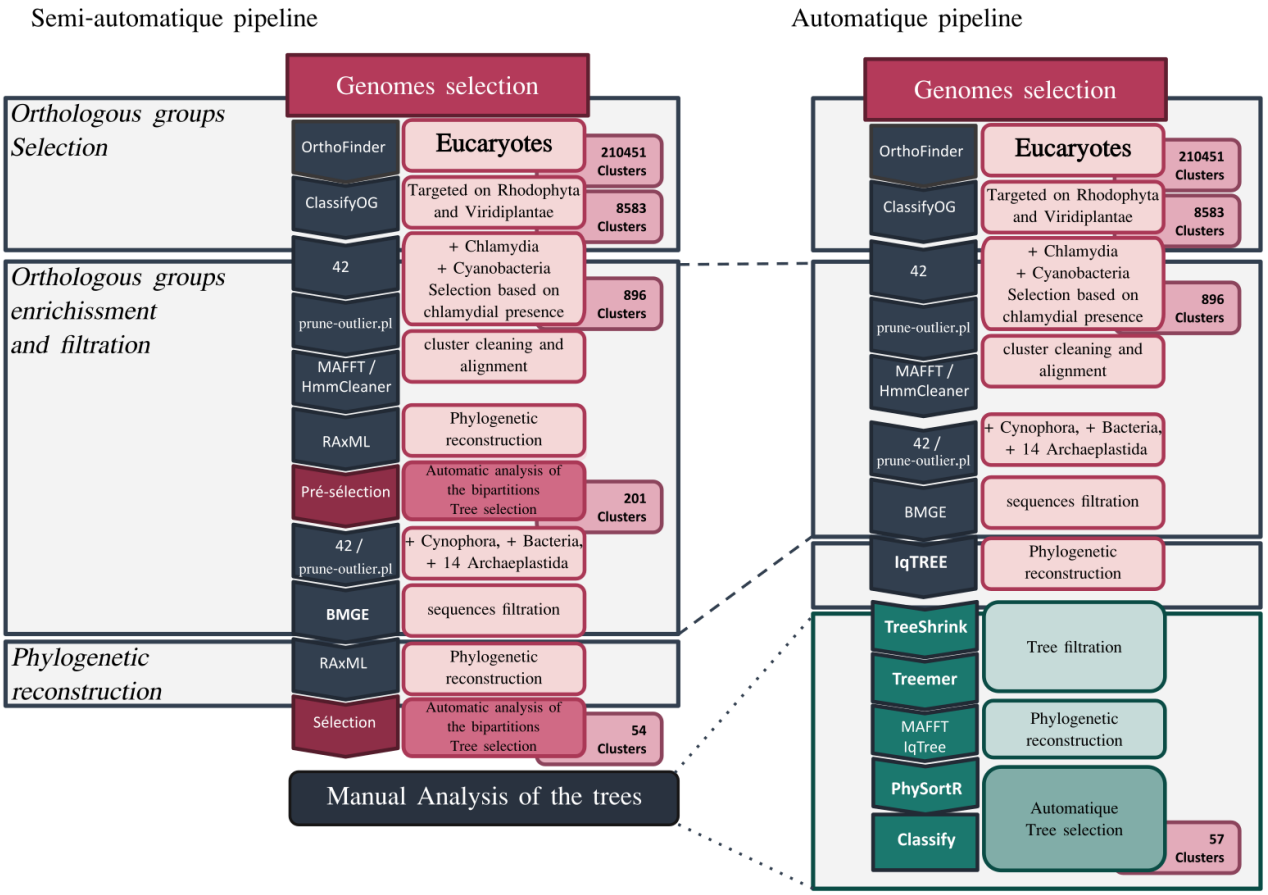


Figure 7: Methodological flowchart of the Chlamydia-Archaeplastida LGT screen. On the left, flowchart of the semi-automatic pipeline, based on the selection of the genomes and proteomes used, then composed of the three main steps: i) creation and selection of orthologous groups, ii) enrichment and filtering of these groups and iii) phylogenetic reconstruction of the pre-selected alignments. The sequence of tools used appears in the center of the flowchart. This pipeline is semi-automatic since a manual analysis of the generated trees validates the selection. The number of orthologous groups remaining at each step of the protocol is indicated on the right side of the flowchart. The automatic pipeline, on the right, is an adaptation of the semi-automatic pipeline calibrated by the manual analysis of the trees, so that the two versions of the pipeline are equivalent. The main steps remain the same, with the exception of removing the intermediate pre-selection of orthologous groups. The replacement of the manual tree analysis reduces the bias of the protocol and allows it to be applied to other bacterial groups.

The 201 corresponding alignments were then enriched with the rest of the proteomes and genomes, notably from other bacteria, but also from glaucophytes. Indeed, aware of the lower quality of the available data for Glaucocystophyceae, these were not integrated in the initial steps of the pipeline. As previously described, a cluster filtering step is necessary before performing the phylogenetic reconstruction. However, for this second step of the

protocol, several tools were first tested in order to optimize the results, both in quality and in computation time (Figure 8).

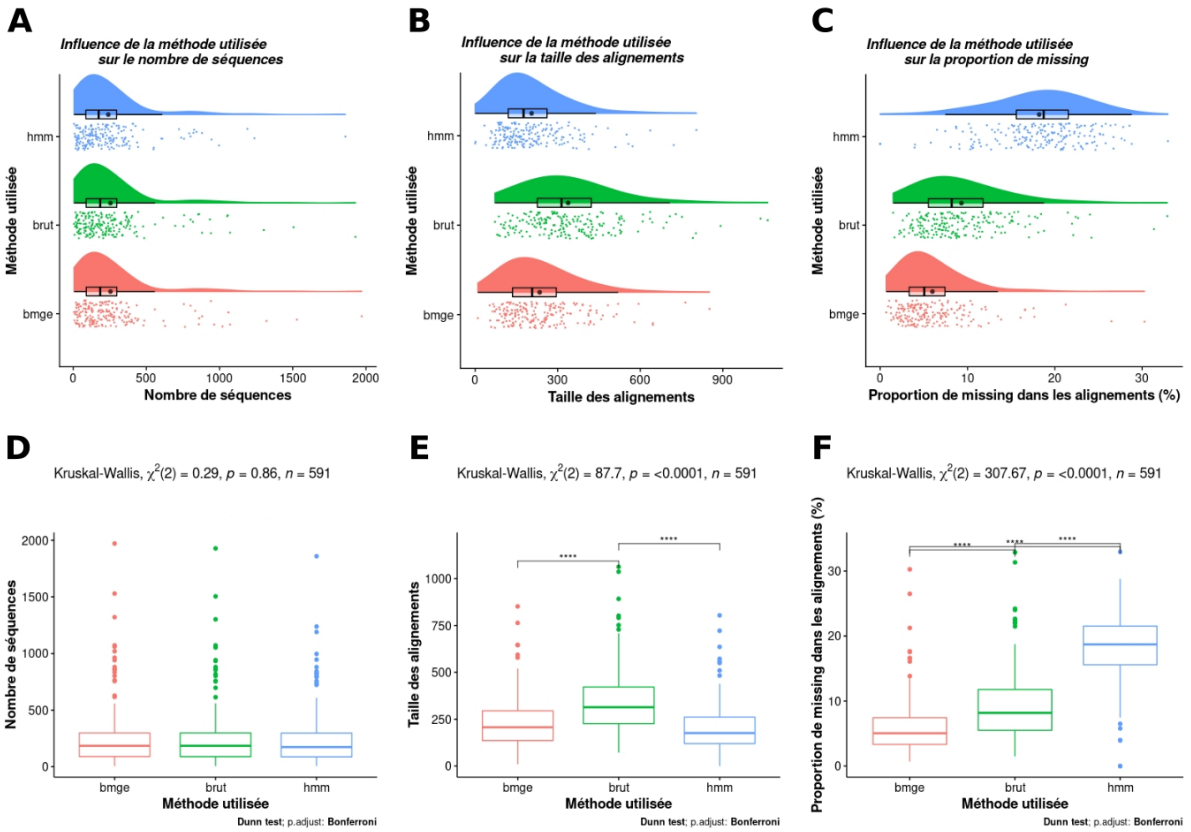


Figure 8: Tests of the influence of three sequence filtration methods on alignments. The same sample of orthologous groups selected following the semi-automatic pipeline pre-selection step was cleaned by HmmCleaner (blue), BMGE (and more precisely by ali2phylip.pl mask BMGE, red) and only by ali2phylip.pl("raw" or “brut”, green). For each of the methods used, the total number of sequences in the alignments (A, D), the length of the alignments (B, E), as well as the proportion of "missing" (C, F) were compared. The statistical evaluation, performed by a Kruskal-Wallis test, followed by a Dunn's test (p-adjust Bonferroni), shows a significant difference between the methods regarding the length of the alignments (E) and the proportion of "missing" (F).

Thus, we compared three methods of filtering orthologous groups: 1) HmmCleaner combined with ali2phylip.pl; 2) ali2phylip.pl combined with BMGE; and 3) ali2phylip.pl alone (named HmmCleaner, BMGE, and raw or “brut”, respectively). For each condition, phylogenetic reconstruction was performed with RAxML, under an LG4X and ultrafast-bootstrap model, on the same sample from the pipeline pre-selection. We were thus able to compare the selections obtained at the output of the pipeline, but also the influence of these methods on the alignments themselves, in order to keep the optimal version. The filtering methods mentioned impact the alignments as a whole, and influence the phylogenetic reconstruction afterwards. The main effects of these methods are on the length of the alignments and their proportion of missing sites. Thus, the analysis of the distributions

of the total number of sequences in the trees shows a stability, a similarity of the profiles for the three conditions tested (Figure 8A), with an average of about 250 sequences per tree. None of these three methods therefore reduces the sample too drastically and allows the taxonomic diversity of each tree linked to the enrichment steps to be maintained. However, the length of the alignments and the proportion of missing sites show significant differences between the three conditions. Concerning the size of the alignments first, HmmCleaner and BMGE seem to have the same distribution profile (Figure 8B), with an average length located around 220 amino acids, compared to the "raw" condition. This difference is confirmed by a Kruskal-Wallis statistical test (Figure 8E). This is expected since HmmCleaner and BMGE are methods for filtering alignments and therefore necessarily reduce their size. The distinctions become more pronounced when analyzing the impact of each condition on the proportion of missing sites in the alignments. Indeed, the three methods tested show different distribution profiles (Figure 8C), with a higher proportion of "missing" sites for HmmCleaner, located around 20%, compared to the "raw" method (~10%) and BMGE (~6%). These differences are significant according to a Kruskal-Wallis test (Figure 8F). Also, by relating the last two parameters analyzed, we find for HmmCleaner an identical alignment size to BMGE for a higher proportion of missing sites. As said before, these characteristics influence the phylogenetic reconstruction, and thus, in our case, also influence the number of trees selected by our protocol. At the end of the pipeline, 67 trees are selected for HmmCleaner, against 54 for BMGE and 54 for "raw", of which 37 are common to all three approaches. At first glance, the HmmCleaner condition seems to be the preferred one, since it allows the identification of a larger number of potential LGTs. However, a quick manual analysis invalidates a large majority of the trees selected only by HmmCleaner (only two trees out of the 27 selected only by HmmCleaner are indeed confirmed in manual analysis). Thus, we have chosen to use the combination of ali2phylip.pl and BMGE for sequence filtering. The automatic analysis of the bipartitions by clans-label.pl of the trees generated by RAxML then identifies 54 trees showing a phylogenetic association between at least one Chlamydia and one Archaeplastida (Figure 7). However, this selection remains imprecise. A manual analysis of the trees is necessary not only to validate the protocol and the methodology in general, but also to validate the identified gene transfers and to verify their endosymbiotic character.

c. Manual tree analysis

The manual analysis of the trees generated by the pipeline was carried out in partnership with Dr Ingrid Lafontaine and Dr Clotilde Garrido. The goals of this manual analysis are multiple. First, it allows us to validate the methodology used to identify gene transfers between target organisms. To this end, it is used throughout the development of the protocol. Secondly, it is necessary to validate the transfers, i.e. to distinguish the trees

correctly selected by the methodology from false positives. Finally, only the manual analysis of the trees allows the identified gene transfer to be placed in an endosymbiotic context. Thus, this part of the study is an important step in establishing the nature of the chlamydial signal in Archaeplastida. However, the analysis of single gene trees remains difficult and their interpretation varies depending on the observers and the evaluation criteria used. Moreover, depending on the reconstruction methods applied, the topology of the trees themselves can be modified. This particularity is one of the criticisms made against the Ménage à Trois Hypothesis. Therefore, in order to try to limit the interpretation bias of the trees, and to objectify the results obtained as much as possible, three independent observers were in charge of manually analyzing each of the 54 trees that came out of the pipeline described above, according to criteria defined beforehand.

The first criterion taken into account in the analysis of the trees is the quality and diversity of the species present in the Chlamydia - Archaeplastida clan. Indeed, a higher diversity, both on the donor side (Chlamydia) and on the acceptor side (Archaeplastida and more generally photosynthetic eukaryotes) of the LGTs, indicates an older transfer, having involved the common ancestor of a larger number of organisms. This is especially true when the transfer considered involves more than one lineage of Archaeplastida. The same logic can be applied when the most basal species of the donors are present in the clan of interest. Indeed, the chlamydial representation in the clan, and more particularly the presence of basal species, would confirm the origin of the gene transfer from this group. These two characteristics, both the presence of basal species and the transfer of genes in multiple lineages of Archaeplastida, allow us to estimate the direction and temporality of these LGTs, pointing then to the common ancestor of these photosynthetic eukaryotes as the recipient. The second criterion taken into account in this analysis of phylogenetic trees is the presence of intruders in the subtree of interest, which may indeed reflect more disseminated transfers in the evolution and the living world depending on the diversity of the species present. In other words, the potential presence of these non-target species may also reflect multiple independent transfer events, in an endosymbiotic or non-endosymbiotic context, as well as LGT unrelated to primary plastid endosymbiosis. Analysis of the diversity of organisms in the clan generally allows differentiation between the different cases. Finally, the general topology of the tree is studied, in order to identify possible paralogues in case of multigenic family, but also, in case the trees show a clan between Chlamydia and one or two lineages of Archaeplastida, to visualize the phylogenetic position of the other lineages of Archaeplastida. Thus, the presence in the same tree of a Chlamydia-Archaeplastida clan and another Cyanobacteria-Archaeplastida clan, attests to the MATH temporality of LGT. Using these criteria, the phylogenetic trees analyzed are classified into three different categories: 1) in favor of the Ménage à Trois Hypothesis, i.e., showing a signal that can be placed in an endosymbiotic context, 2) not in favor of MATH, and 3) uncertain, impossible to conclude.

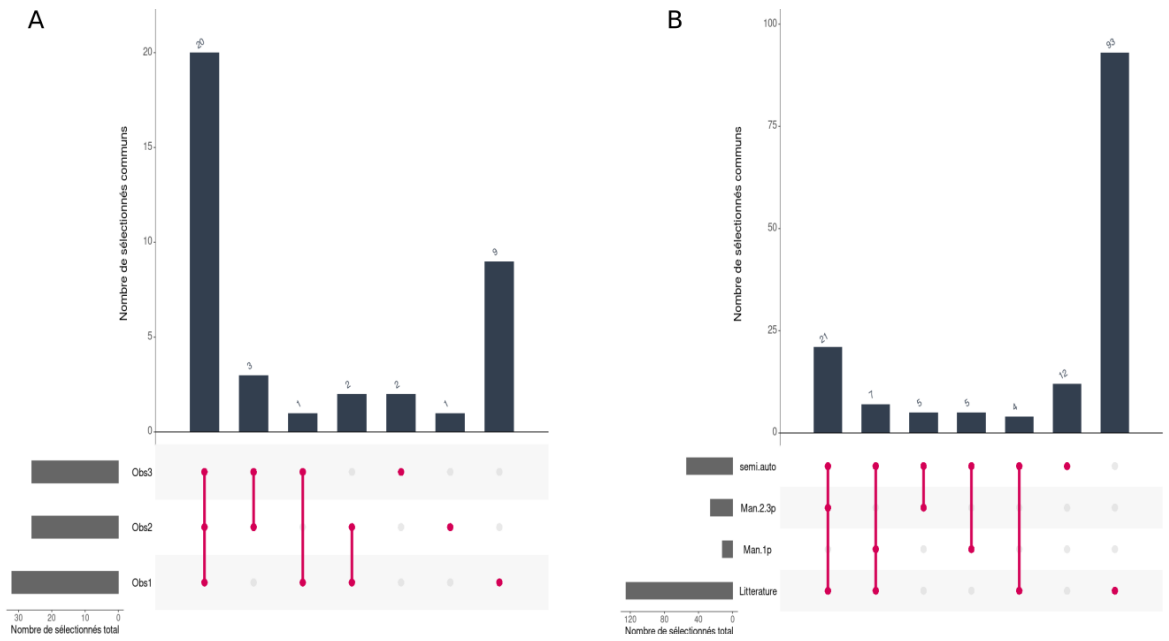


Figure 9: Summary diagrams of the semi-automatic pipeline selection and manual analysis of the generated trees. The semi-automatic pipeline selected 54 trees with a clustering clan between Chlamydia and Archaeplastida, which are then manually analyzed by three observers. 37 of these trees are confirmed to have primary plastid endosymbiosis-related gene transfer by at least one of the three observers. The diagrams represent the intersections between the different sets considered. Each pink dot visualizes the groups considered. Panel A shows the results of the manual analysis of the trees according to the observers (only the trees confirming a gene transfer). The details of the whole semi-automatic pipeline selection are presented in panel B, compared to the manual analysis and the literature data. Obs1: observer 1; Obs2: observer 2; Obs3: observer 3, semi.auto: semi-automatic pipeline; Man.2.3p: trees confirmed by 2 or 3 observers during manual analysis; Man.1p: trees confirmed by only 1 observer during manual analysis; Literature: genes reported as chlamydial in Archaeplastida in (Ball et al., 2013; Becker et al., 2008; Cenci et al., 2018, 2017; Huang and Gogarten, 2007; Moustafa et al., 2008).

Of 54 trees analyzed, 37 are selected by at least one observer as being in favor of the Ménage à Trois Hypothesis, i.e., showing gene transfer that can be linked to an endosymbiotic context, of which 20 are unanimous and 6 others are selected by two out of three observers (Figure 9A). By further analyzing the differences in selection among the three observers, we note that the 12 trees selected by only one are mostly categorized as uncertain or inconclusive for the other two. In general, we therefore decide to validate a tree as compatible with the Ménage à Trois Hypothesis if at least two out of three observers classified it as such. As a result, 26 trees emerging from the pipeline are validated by manual analysis and support a chlamydial involvement in the evolution of Archaeplastida.

d. Signal congruence

The manual analysis thus allows us to validate the gene transfers visible on the single gene phylogenetic trees, and to potentially confirm their occurrence in an endosymbiotic context. A supermatrix, created by SCaFoS (Roure et al., 2007) from the only Chlamydia and Archaeplastida sequences present in the clans of the 26 selected trees and for which the phylogenetic reconstruction is performed by IQ-TREE (Nguyen et al., 2015), then informs us on the congruence of the global Chlamydian signal in Archaeplastida. In other words, manual tree analysis identifies and validates MATH genes, while concatenation determines whether their signal is congruent. Regardless of the model used for the phylogenetic reconstruction (LG4X, C20 or C60), when the trees are rooted on the similichlamydia (basal Chlamydia), we observe the clustering of the three Archaeplastida lineages, branched directly with the environmental Chlamydia (Figure 10 + appendix 1). Moreover, within the Archaeplastida clan, the topology obtained respects the known evolutionary relationships in algae and plants. This grouping of Archaeplastida supports signal congruence, and thus suggests a similar evolutionary history of the selected genes, at least for the part of their post-endosymbiotic evolution. Indeed, otherwise, the Archaeplastida would have been dispersed within the Chlamydia.

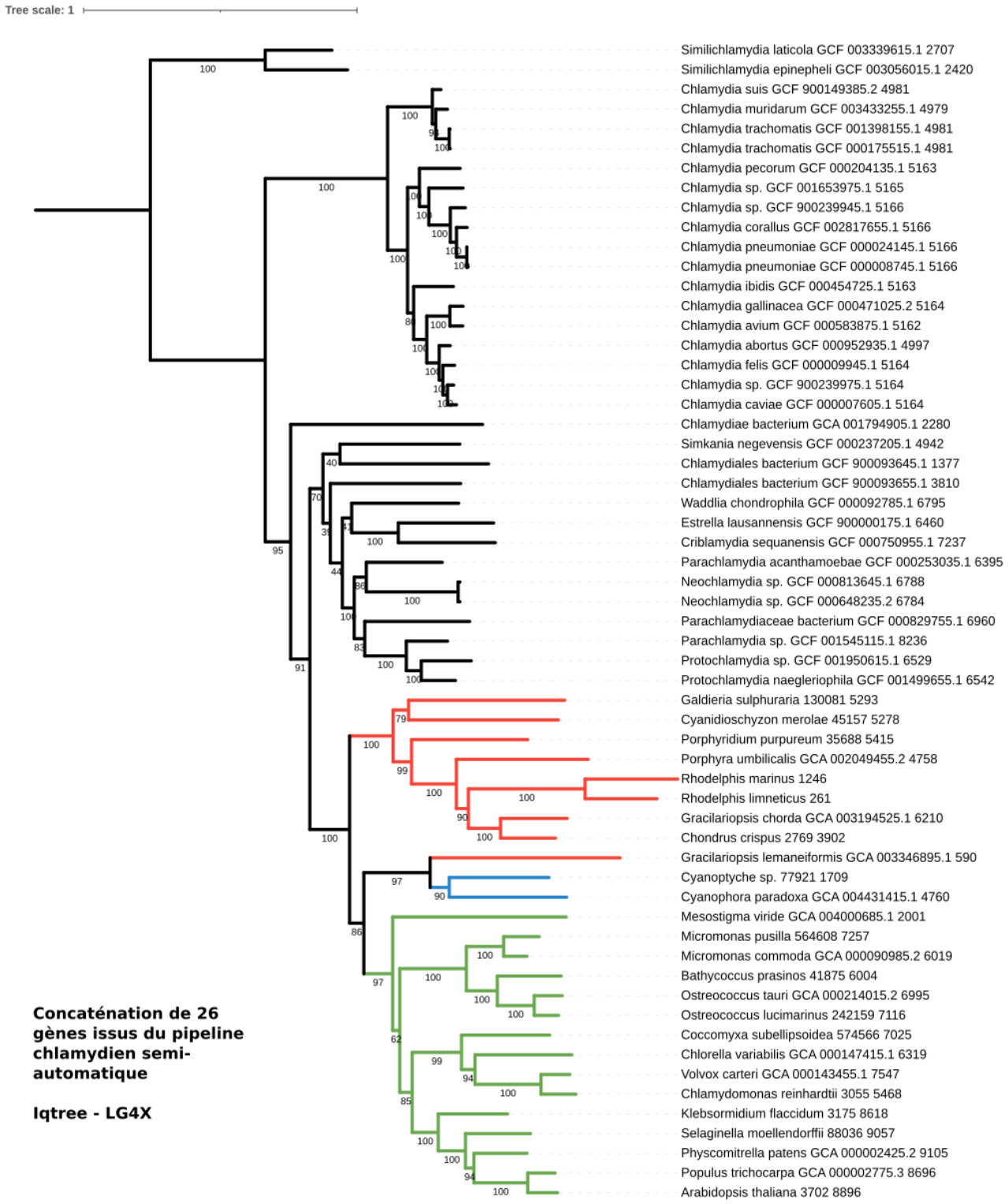


Figure 10: Phylogenetic tree from the concatenation of the 26 genes retained by manual analysis. The tree was obtained by IQ-TREE, under an LG4X model, after concatenation of the chlamydial and archaeplastid sequences of the 26 genes confirmed in manual analysis. Bootstrap values are shown on the branches. Rooting is manual on Similichlamydia. The scale counts in number of amino acid substitutions per site. In red: Rhodophyta and Rhodelphea, in green: Viridiplantae and in blue: Glaucophyta.

e. Inventory of chlamydial genes in the literature and correlation with our analysis

Several studies in the past have attempted to clarify the impact of Chlamydia in the evolution of Archaeplastida, including a catalog of genes transferred from the former to the latter. Based on these studies, we have taken up all Archaeplastida sequences reported to be of Chlamydian origin in Ball et al. 2013; Becker et al. 2008; Cenci et al. 2018, 2017; Huang and Gogarten 2007; Moustafa et al. 2008. This literature inventory includes 150 sequences that, when dereplicated to avoid duplication due to synonyms, are distributed in 128 of our starting orthologous groups. At the end of the pipeline, 32 OGs selected among the 54 are also found in the literature and, among the 26 validated in manual analysis, 21 are also identified in this inventory (Figure 9B). All the trees corresponding to the inventory (especially those not selected) were then manually analyzed to validate and understand the pipeline selection. It was during this manual analysis that we noticed a limitation of our pipeline: the majority of the trees in the inventory are considered false positives according to our methods. However, on reflection, some should have been selected. When looking further into these trees, it appears that the trees in question have clans for which Chlamydia and Archaeplastida branch together, but also with other species. Thus, the tool in charge of the automatic analysis of bipartitions could not identify the subtree of interest since it does not take into account clans interrupted by "intruders", i.e. non-target species. In this case it is all bacteria and non-photosynthetic eukaryotes.

The manual analysis of the trees, both from the pipeline selection and from the literature inventory, allows us not only to validate and adjust our methods, but also, and more importantly, to serve as a basis for fully automating the pipeline so that it mimics the results obtained by hand. Indeed, according to the observations made manually, we have parameterized our methods in order to correct selection biases and automatically transcribe our analysis criteria.

2. Significance of the chlamydial signal in Archaeplastida

Analysis of the Ménage à Trois hypothesis requires first identifying the Chlamydial signal in Archaeplastida, and only then determining whether this signal is different from others, so as to assess the specific impact of Chlamydia on the origin of photosynthetic eukaryotes, not only quantitatively but also functionally and metabolically. So far, we have focused on the identification of this signal and its phylogenetic nature, showing the presence of gene transfers between Chlamydia and Archaeplastida whose origin goes back to primary plastid endosymbiosis.

It is now necessary to place this identified signal within the overall evolutionary history of photosynthetic bacteria and eukaryotes, in order to determine whether Chlamydia played a particular role during primary plastid endosymbiosis. This involves comparing the chlamydial signal in Archaeplastida to other bacterial signals, but also assessing the specificity of the Archaeplastida linkage compared to other eukaryotic groups. This control of the chlamydial signal in Archaeplastida is also highlighted in relation to the cyanobacterial signal, used here as a reference.

a. Pipeline automation

The limiting step of the protocol described in the first chapter remains the manual analysis of the trees. Indeed, this step is certainly necessary to validate the gene transfers and especially to identify the so-called "MATH" genes, but it is time consuming and open to criticism. Since the goal is now to compare different phylogenetic signals, both on the acceptor and donor sides of the LGTs, and not to establish an exhaustive listing of MATH genes, we calibrate the pipeline on the results of the manual analysis to develop a fully automated protocol. As a result, based on the observations made in manual analysis, the results obtained and the analysis criteria set up previously, we adjusted the bioinformatics pipeline (Figure 7).

The tree pre-selection step, still only dealing with target species found in a MATH context (eukaryotes, Chlamydia and cyanobacteria), was first removed, so as to save computational time. Then, still with the idea of saving computational time, we replaced RAxML (Stamatakis, 2014) with IQ-TREE (Nguyen et al., 2015). Indeed, for phylogenetic reconstruction of the same alignments under the same models, RAxML and IQ-TREE give very similar results. The difference in maximum likelihood scores for each pair of trees is negligible, although consistently in favor of RAxML (average score of 27500, with an average difference between RAxML and IQ-TREE of 32). Therefore, the final selection by the previously described pipeline identifies the majority of the same trees as having a phylogenetic interaction between at least one Chlamydia and one Archaeplastida. However, IQ-TREE is (much) faster. To fully validate the change in method, the 8 trees selected by RAxML and not by IQ-TREE were manually analyzed.

The automation of the pipeline is based entirely on the manual analysis done previously, as much on the results obtained as on the criteria set up for the study of the trees. Indeed, each step of the manual analysis has been translated into an automated bioinformatics tool. The quality of donor and acceptor inference is first improved by Treeshrink (Mai and Mirarab, 2018) and Treemer (Menardo et al., 2018), which, respectively, remove long branches and dereplicate the trees, without reducing the number of species present. PhySortR (Stephens et

al., 2016) and classify-ali.pl (https://metacpan.org/dist/Bio-MUST-Core) combined then handle the identification of the gene transfers themselves, while taking care of the other criteria established during the manual analysis. PhySortR is an R package that identifies phylogenetic trees with a clan composed of predetermined species, while taking into account the possibility of non-target species intersecting the clans of interest (intruders), as well as the proportion of target species in the selected clan relative to the entire tree. Several tests of these parameters were performed to best adjust the selection performance of the package. Thus, the minimum proportion of target species in the sub-tree corresponding to the branch of interest compared to the whole tree is set to 30% and we allow the presence of 10% of intruders. In other words, each tree for which at least 30% of the total target species (Chlamydia and eukaryotes from the primary plastid endosymbiosis with respect to the chlamydial pipeline) of the total tree are present in the subtree of interest, and for which less than 10% of the species in this subtree are other than these target species, is selected. While PhySortR handles the criteria of overall relative tree diversity and the presence of intruders that may cut across clans, classify-ali.pl refines the selection made by filtering the identified subtrees on an absolute diversity criterion. Here, we set the minimum number of species present in the clan of interest to 5, including at least 2 donors and 3 acceptors, for a tree to be selected. The set of pipeline modifications are materialized in Figure 7 .

Concretely, if we take up the LGT screen between Chlamydia and Archaeplastida, the new version of the pipeline selects 57 trees with a phylogenetic interaction between at least 2 Chlamydia and 3 Archaeplastida. In comparison, the previous semi-automated pipeline output 54 trees, 26 of which were validated in manual analysis (2 or 3 opinions). Of these 57 trees, 26 are common to the previous selection, and all are confirmed by manual analysis by at least one observer (21 trees are selected by 2 or 3 observers, 5 are selected by one observer) (Figure 11A). The automatic analysis thus misses 28 trees (54-26), though mostly doubtful or invalid (23/28), while bringing in an additional 31 trees (57-26). Looking in more detail at these additional trees, especially by comparing them to their correspondent from the semi-automatic pipeline, we find, as in the analysis of the literature inventory, that it is mostly trees with clans interrupted by intruders that were not identified by the first pipeline. The adjustment of the protocol, in particular by the choice of PhySortR, thus allows to recover these false negatives and to assemble a more complete catalog of MATH genes. On the other hand, 5 trees confirmed in manual analysis by at least two of the three observers are not selected by this new version of the pipeline. The objective being to compare the phylogenetic signals of the different bacterial groups, we nevertheless accept these new false negatives. Indeed, it is important to keep in mind that the goal of the automatic protocol presented here is not to establish an exhaustive catalog of MATH genes. Therefore, the parameters have been chosen to optimize this comparison, with full awareness of a sensitivity that is not necessarily maximal, leading to the possible rejection of some MATH genes.

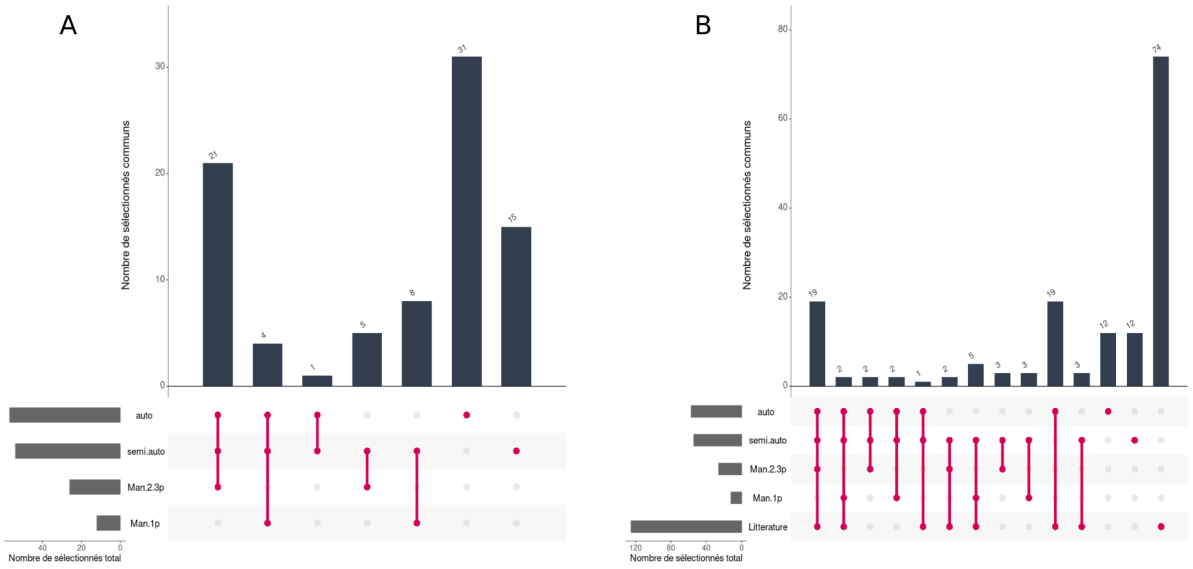


Figure 11: Summary diagrams of selections based on different methods. The diagrams represent the intersections between the different sets considered. Each pink dot visualizes the groups considered. The semi-automatic pipeline (semi.auto) selected 54 trees with a branching between Chlamydia and Archaeplastida, which are then manually analyzed by three observers. 26 of these trees are confirmed as having a gene transfer related to primary plastid endosymbiosis by at least two of the three observers (Man.2.3p), 15 others are validated by only one of the observers (Man.1p). The automatic (auto) pipeline, meanwhile, identified 57 trees with a phylogenetic relationship between Chlamydia and Archaeplastida. Panel A shows the common selections between the different methods. Panel B also includes the chlamydial genes identified in Archaeplastida in the literature. semi.auto: semi-automatic pipeline; Man.2.3p: trees confirmed by 2 or 3 observers during manual analysis; Man.1p: trees confirmed by only 1 observer during manual analysis; Literature: genes reported as chlamydian in Archaeplastida in (Ball et al., 2013; Becker et al., 2008; Cenci et al., 2018, 2017; Huang and Gogarten, 2007; Moustafa et al., 2008).

b. Pipeline controls

Two types of controls are considered to evaluate the overall chlamydial signal in Archaeplastida: a control of the "donors" at the origin of the gene transfers and a control of the "acceptors". The first allows us to evaluate the chlamydial signal in Archaeplastida compared to other bacterial groups, the second to determine the impact of Chlamydia specifically in Archaeplastida compared to other eukaryotes. In all cases, a reorientation of the pipeline on the identification of LGTs of interest is necessary.

On the "acceptor" side of these controls, we have chosen to redirect the pipeline to the identification of gene transfers between Chlamydia and Amoebozoa and then between Chlamydia and Fungi. In view of their antiquity, and their predilection for gene transfer, amoebae play the role of positive control and are likely to have integrated a significant

number of Chlamydian genes into their genome. Conversely, Fungi, known to be more reactive to gene transfer, are the negative control. Since we are dealing here with the identification of chlamydial gene transfer to different eukaryotic groups, the reorientation of the pipeline concerns the very first steps of the protocol, and in particular the selection filter applied to the orthologous groups.
Regarding fungi, the dataset of 57 eukaryotes used to create the orthologous groups by OrthoFinder contains 11 species. We consider this number sufficient to compare their signal to that of Archaeplastida, since we count 12 species of the latter. From the initial 210451 orthologous groups, classify-mcl.pl retains 5808 with at least 2 Fungi species. After enrichment with cyanobacteria and Chlamydia, only 625 orthologous clusters are selected on the criterion of the presence of at least one Chlamydia. These clusters then continue in the protocol as previously described (Figure 7).
For Amoebozoa, on the other hand, only 3 species are present in the initial dataset used for the creation of orthologous groups. Before filtering the clusters on the presence of these target species, we enrich them with 10 additional genomes - proteomes of amoebae. 5734 clusters are then selected on the presence of at least two Amoebozoa and 1558 after enrichment in Chlamydia and cyanobacteria to continue the protocol. The selection of Amoebozoa genomes and proteomes, as well as the corresponding pipeline, was performed by Clotilde Garrido. Thus, only 1 tree was identified by the pipeline as having a phylogenetic interaction between at least 2 Chlamydia and 3 Fungi, and 16 were selected for the pipeline re-oriented towards Amoebozoa (Figure 12).

Regarding the "donor" side of these controls, we relied on the results of Dagan et al., 2013, to choose the bacterial groups to redirect the pipeline to. Indeed, these authors estimated the gene transfer contributions of each bacterial group in Archaeplastida. According to their results, the chlamydian contribution would only come in 6th position, after cyanobacteria, Proteobacteria, Firmicutes, Actinobacteria and Bacteroidetes. Therefore, in addition to cyanobacteria, we chose to include all these groups in our analyses, in order to compare them to the chlamydial signal in Archaeplastida. The cyanobacterial signal, which is the majority in the evolution of Archaeplastida, will be used as a reference for comparison and to evaluate the sensitivity of our pipeline.
The reorientation of the pipeline is therefore carried out on each group chosen and requires first of all adjusting the selection of species entering the bioinformatics protocol (Figure 7). For each bacterial group identified, we used TQMD (ToRQuEMaDA: tool for retrieving queried Eubacteria, metadata and dereplicating assemblies Léonard et al., 2021) to produce a list of target species, representative of the diversity of the group. The main criterion for these TQMD selections was to produce lists of species similar in number to Chlamydia (the phylogenetic structure of each selection being much more difficult to control). Indeed, in order to compare the different results obtained, it is important to homogenize the data and

methods. Four lists of organisms were therefore produced, containing respectively 36 Proteobacteria, 37 Bacteroidetes, 20 Actinobacteria and 22 Firmicutes. The entry of the genomes and proteomes associated with these lists in the protocol will therefore depend on the bacterial signal that we want to quantify (Figure 12, appendix 9). In addition, the list of general bacteria, without distinction of taxa, is also affected by the orientation of the pipeline. Indeed, this list of 49+92 organisms, entering the second phase of enrichment, represents the complete diversity of bacteria, and therefore contains different target organisms depending on the orientation. Therefore, depending on the signal studied, the target species present are removed.

The set of orthologous starting groups remains the same as before, as well as the first filter on the presence of at least 2 Viridiplantae and/or Rhodophyta. The reorientation of the pipeline takes place afterwards. Instead of enriching with Chlamydia and then selecting orthologous clusters based on the presence of these pathogens, each pipeline is enriched with the proteomes of the target organisms for which we want to evaluate the signal, and then only orthologous clusters with at least 1 target species are selected. Thus, 1143 clusters continue the protocol for Bacteroidetes, 1147 for Proteobacteria, 1002 for Firmicutes and 1024 for Actinobacteria and follow the same steps of the automatic pipeline developed for Chlamydia.

For cyanobacteria, since the 48 proteome dataset is already present in all control pipelines, it was sufficient to reorient the initial pipeline at the time of cluster selection. Indeed, using the protocol described for the Chlamydian pipeline, after enrichment in cyanobacteria and Chlamydia, to keep the target bacterial signal competitive, the selection filter for orthologous clusters is done on the presence of at least one cyanobacteria and not on the presence of Chlamydia. 1572 OGs then continue the pipeline.

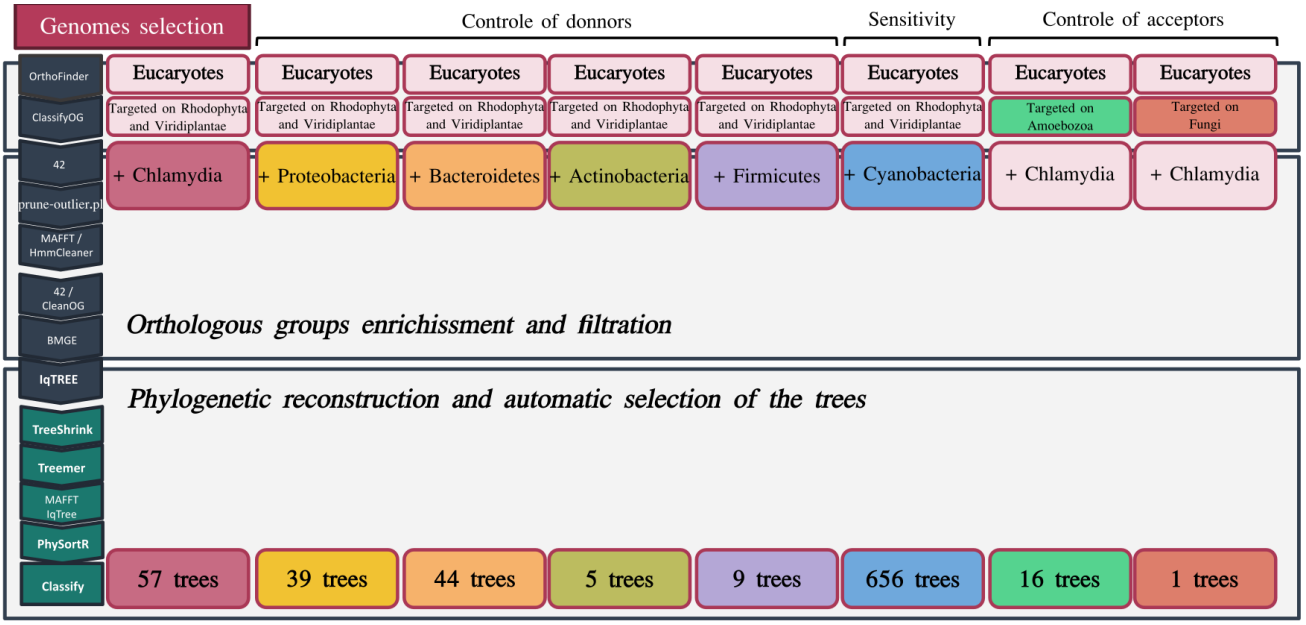


Figure 12: Flowchart of the pipeline reorientation on the identification of lateral gene transfers involving control groups. The methodological protocol of the automatic pipeline, detailed in Figure X, visible on the left of the diagram, is unchanged. The modifications made appear in bright colors. Two types of controls are performed: the control of donors at the origin of LGTs and the control of acceptors. For the donor controls, the pipeline is redirected to the first step of orthologous group enrichment, in order to identify LGTs between Archaeplastida and different bacterial groups. In pink: Chlamydia, in yellow: Proteobacteria, in orange: Bacteroidetes, in green: Actinobacteria and in purple: Firmicutes. The identification of LGT between Cyanobacteria and Archaeplastida (in blue) serves as a reference for comparison and evaluation of the sensitivity of the protocol. On the side of the acceptor controls, the pipeline reorientation allows the identification of LGTs between Chlamydia and two target groups, and was performed on the selection of orthologous groups integrating the protocol. For the bacterial controls, in fact, it is targeted on the presence of at least 2 Vridiplantae and/or Rhodophyta, while for the acceptors, the sorting is targeted on the presence of at least 2 Amoebozoa (in bright green) or 2 Fungi (in red). In grey is represented the separate pipeline (see below) specifically targeting LGT between Chlamydia and Glaucophyta

The rest of the protocol remains unchanged here as well. For all pipelines tested, trees are selected if a minimum of 2 donors and 3 acceptors are present together in the same clan. Therefore, re-pipelining on each bacterial group selects 44 trees with a phylogenetic relationship between 2 Bacteroidetes and 3 Archaeplastida, 39 for Proteobacteria, 9 for Firmicutes, 5 for Actinobacteria, and 656 for Cyanobacteria (Figure 12). This last proportion is expected, since the ancestor of the plastid was a cyanobacterium. The estimate of cyanobacterial genes in Archaeplastida varies between 600 and 5000 depending on the study. Our methods are therefore, as expected, at the low end of the sensitivity range described in previous publications. However, since the goal is to compare different signals with identical stringency, we accept this loss of signal identified by our methods. In general, we can see a higher number of gene transfers for Chlamydia, with 57 trees selected, compared to other bacterial groups. Compared to Bacteroidetes, the second group with the highest number of gene transfers to Archaeplastida, Chlamydia contributed about 23% more to this genetic mix.

According to our methods, Chlamydia thus represent the first contributor of gene transfers, after cyanobacteria, to Archaeplastida. This first result confirms the undeniable relative impact of Chlamydia on the evolution of photosynthetic lineages, but it is not sufficient to validate or refute the Ménage à Trois Hypothesis as such.

Concatenation of the genes selected by the control pipelines, both by creating a supermatrix and a supertree, reveals a similar congruence profile for each (Figures 12-16). Since the Fungi control pipeline only led to the selection of one tree, concatenation was not performed for it. Focusing on the specific study of the chlamydial signal in eukaryotes, concatenation of the amoeba selection shows a clear separation between eukaryotes and bacteria, but appears to be less well supported than for Archaeplastida (Figures 12 and 13). Indeed, although grouping Amoebozoa together shows a monophyletic signal, the bootstrap values support the phylogenetic relationship less well, both between the two taxonomic groups and within the chlamydial diversity itself, than when analyzing the chlamydial signal in Archaeplastida. This, combined with the difference in the number of transferred genes, is therefore rather evidence for a privileged relationship of Chlamydia with Archaeplastida.
On the side of the bacterial controls, taking only the target organisms present in the clans identified for each pipeline, the phylogenetic reconstruction of each bacterial selection, whether in LG4X, C20 or C60, shows a clustering of Archaeplastida, respecting moreover the known topology of the evolution of these organisms. This monophyletic set is often connected with the most basal representatives of the evaluated bacterial groups. For each control pipeline, the different concatenations tested (figures 15 to 16, appendix 2, 3, 4) present similar profiles. The few notable differences lie in the internal topology of the Archaeplastida, for which it is the Rhodophyta or the Glaucophyta that diverge first, depending on the trees considered. However, this instability is already known, the order of diversification of three primary lineages still being debated. Within the different selections, the signal is therefore congruent and does not allow to differentiate a privileged contribution of Chlamydia in eukaryotes compared to the other bacterial groups tested.

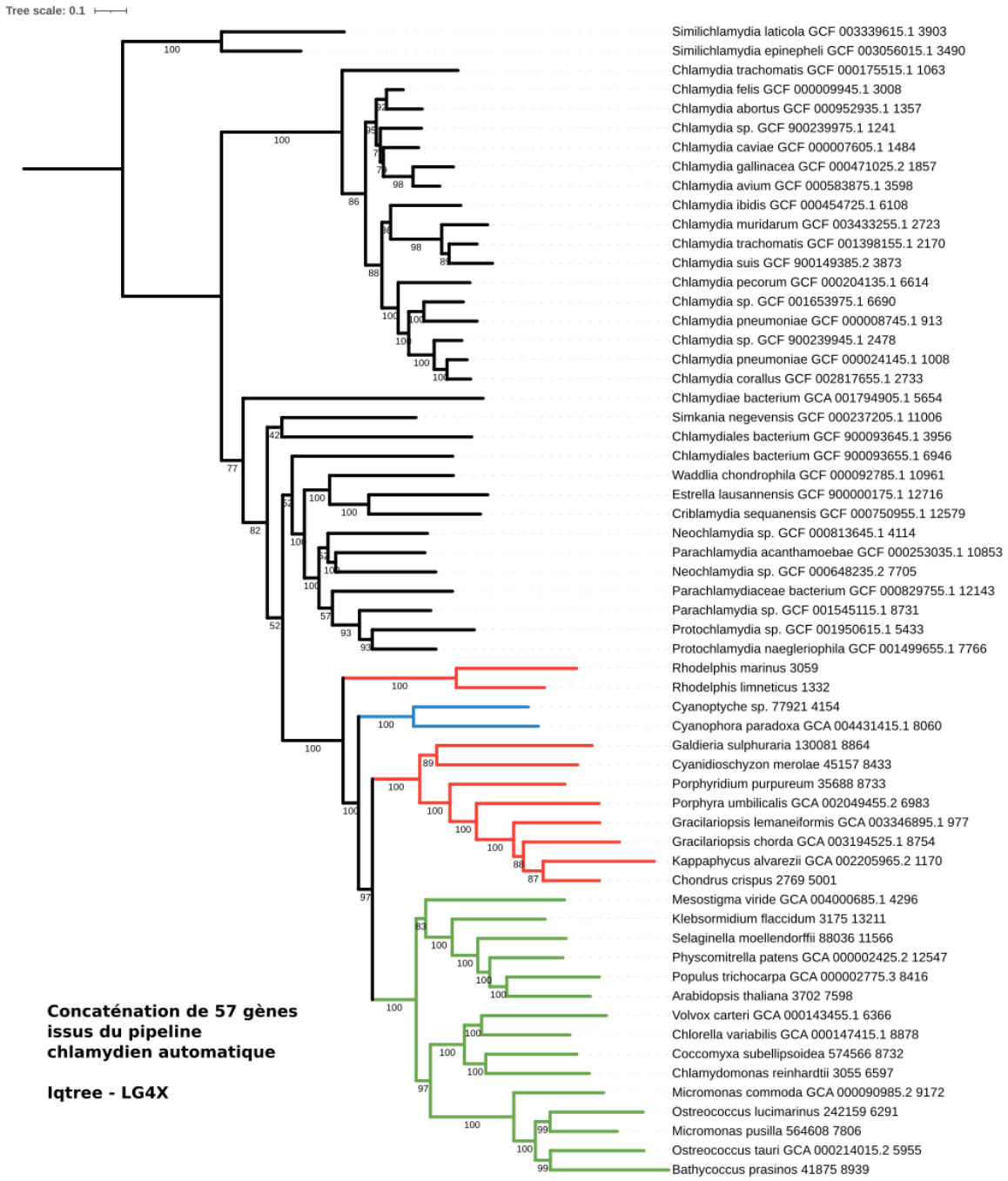


Figure 13: Phylogenetic tree from the concatenation of the 57 genes selected by the automatic chlamydial pipeline. The tree was obtained by IQ-TREE, under an LG4X model, after concatenation of the chlamydial and archaeplastida sequences of the genes selected by the pipeline. Bootstrap values are shown on the branches. Rooting is manual on Similichlamydia. The scale counts in number of amino acid substitutions per site. In red: Rhodophyta and Rhodelphea, in green: Viridiplantae and in blue: Glaucophyta.

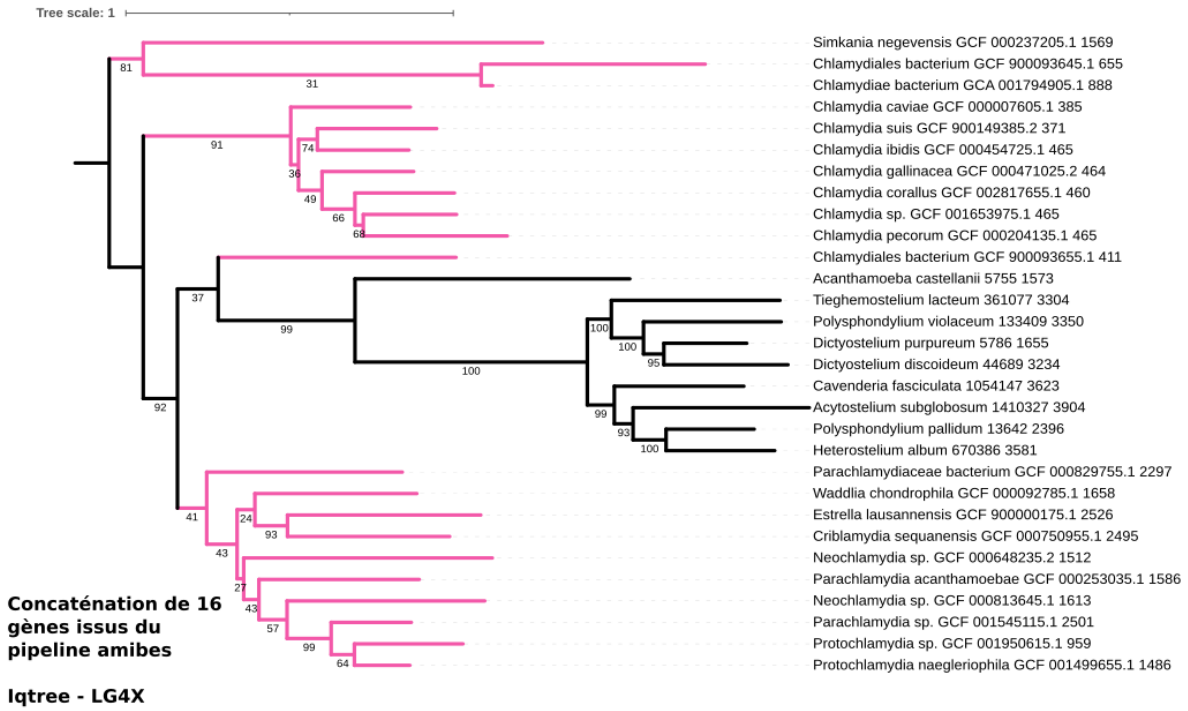


Figure 14: Phylogenetic tree from the concatenation of the 16 genes selected by the automatic amoebozoa pipeline. The tree was obtained by IQ-TREE, under an LG4X model, after concatenation of Chlamydia and Amoebozoa sequences of the genes selected by the pipeline. Bootstrap values are shown on the branches. Rooting is manual on basal Chlamydia. The scale counts in number of amino acid substitutions per site. In pink: Chlamydia, in black: amoboezoa.

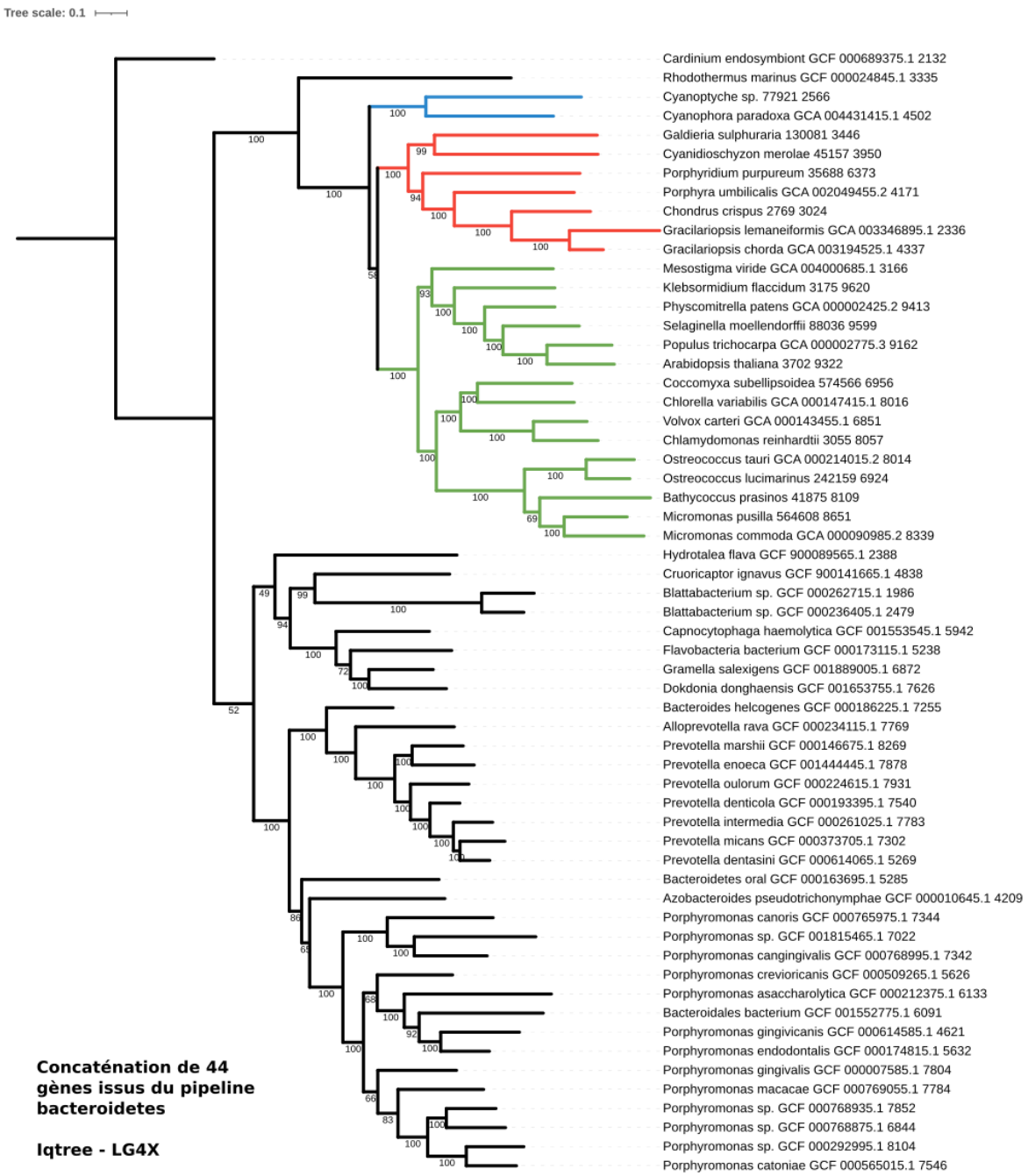


Figure 15: Phylogenetic tree from the concatenation of the 44 genes selected by the Bacteroidetes automatic pipeline. The tree was obtained by IQ-TREE, under an LG4X model, after concatenation of the Bacteroidetes and Archaeplastida sequences of the genes selected by the pipeline. Bootstrap values are shown on the branches. Rooting is manual on basal Bacteroidetes. The scale counts in number of amino acid substitutions per site. In red: Rhodophyta and Rhodelphea, in green: Viridiplantae and in blue: Glaucophyta.

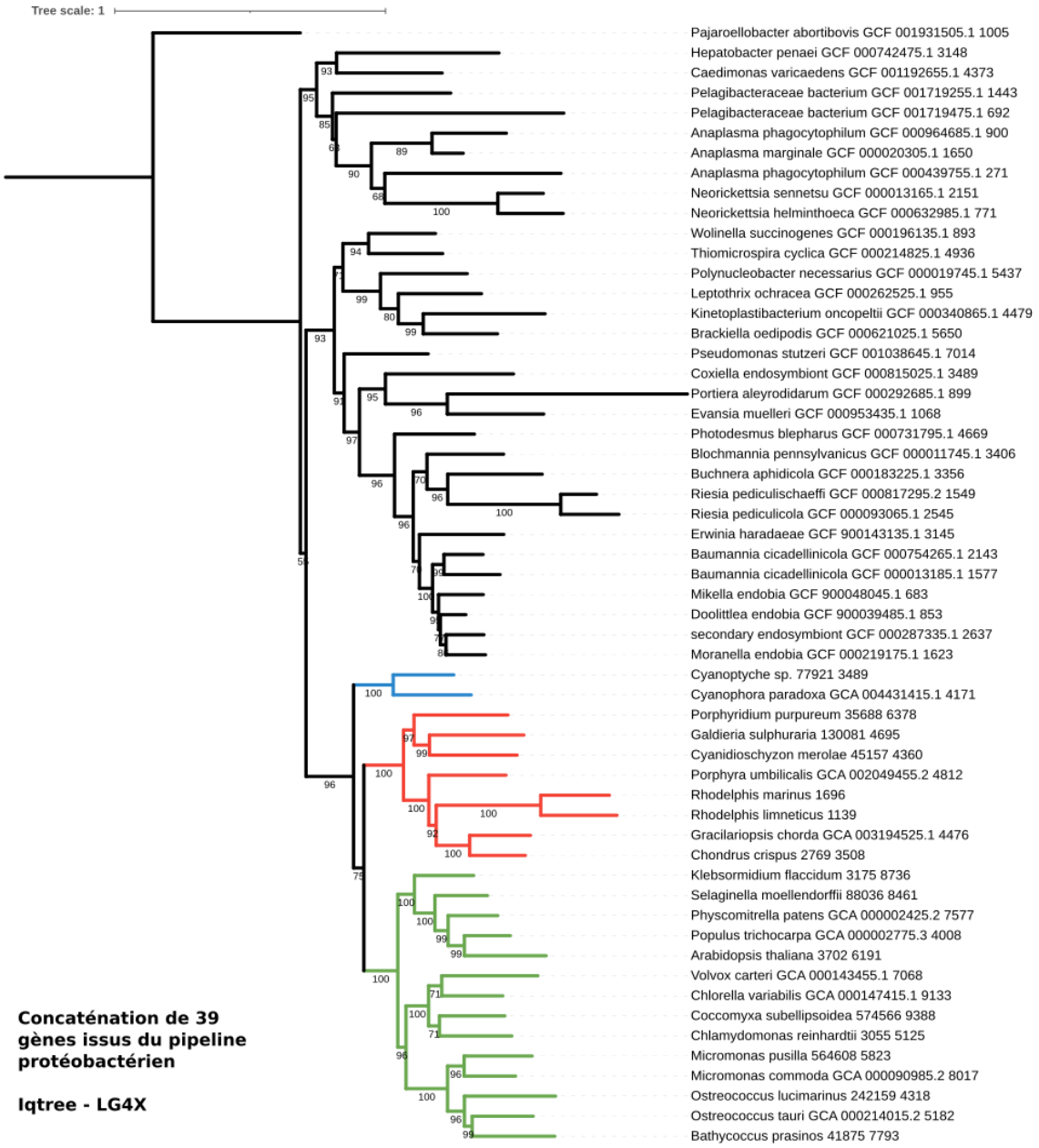


Figure 16: Phylogenetic tree from the concatenation of the 39 genes selected by the Proteobacteria automatic pipeline. The tree was obtained by IQ-TREE, under an LG4X model, after concatenation of Proteobacteria and Archaeplastida sequences of the genes selected by the pipeline. Bootstrap values are shown on the branches. Rooting is manual on basal Proteobacteria. The scale counts in number of amino acid substitutions per site. In red: Rhodophyta and Rhodelphea, in green: Viridiplantae and in blue: Glaucophyta.

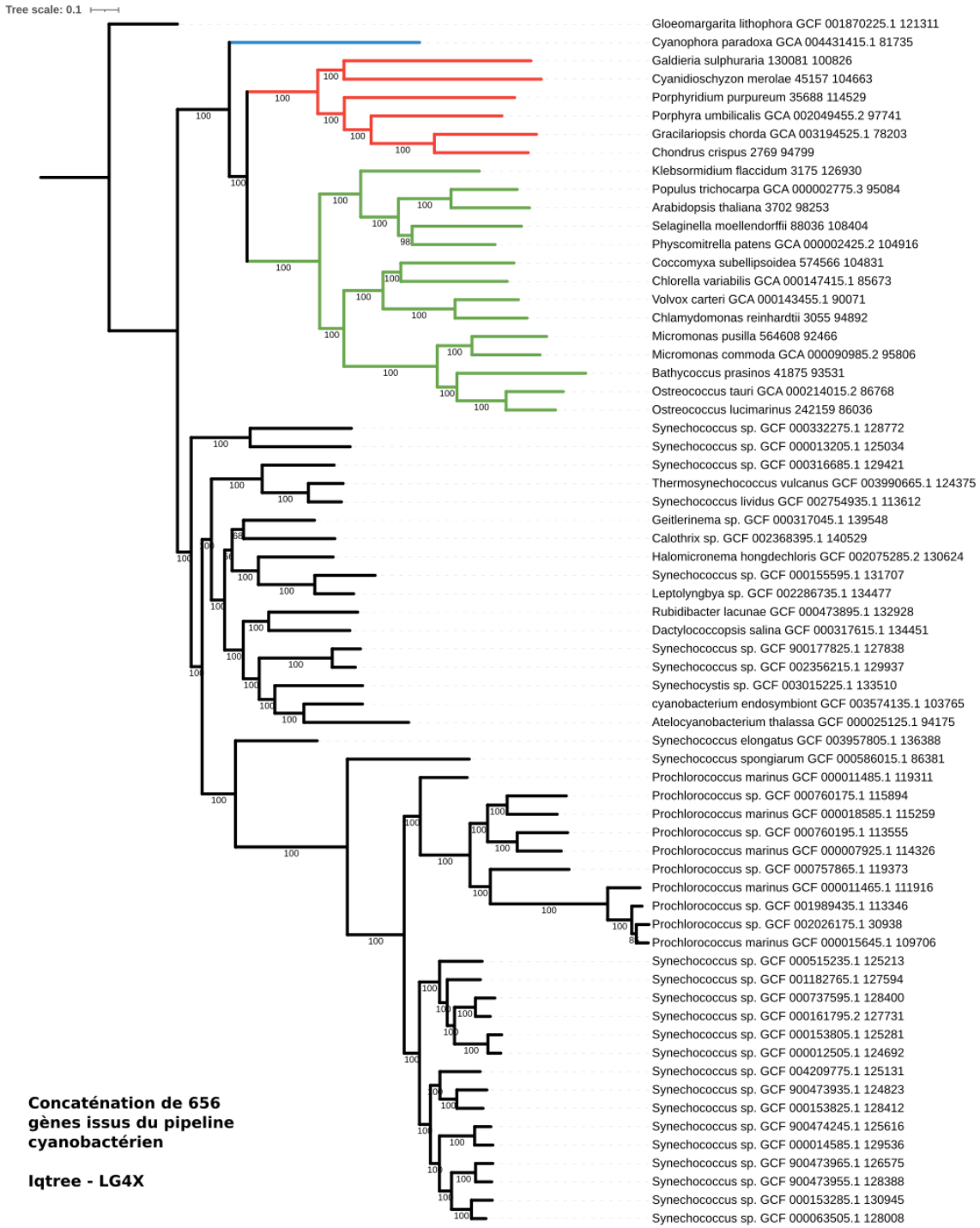


Figure 17: Phylogenetic tree from the concatenation of the 656 genes selected by the Cyanobacteria automatic pipeline. The tree was obtained by IQ-TREE, under an LG4X model, after concatenation of Cyanobacteria and Archaeplastida sequences of the genes selected by the pipeline. Bootstrap values are shown on the branches. Rooting is manual on basal cyanobacteria. The scale counts in number of amino acid substitutions per site. In red: Rhodophyta and Rhodelphea, in green: Viridiplantae and in blue: Glaucophyta.

c. pattern and diversity

So far we have identified the phylogenetic nature of several bacterial signals in Archaeplastida. For each, there is indeed a bacterial contribution in these photosynthetic eukaryotes, for which the signal appears congruent. The phylogenies identified by the pipelines, although selected to mimic a manual tree analysis, may reflect disparate transfers during the evolution of Archaeplastida and not a common evolutionary origin, as predicted by MATH. To go further in the analysis, and to define the signal in its subtleties, we studied the set of identified gene transfers by taking more specifically into account the diversity of Archaeplastida impacted.

Assuming that gene transfers from primary plastid endosymbiosis would be more generally distributed within photosynthetic eukaryotes, thus impacting several Archaeplastida lineages, the specific contribution of a bacterial group should then be felt in the diversity of the selected clans. Indeed, gene transfer in the common ancestor of Archaeplastida implies a subsequent co-evolution of donor and acceptor organisms and a wider distribution of the transferred gene in the diversity. On the contrary, a gene transfer later in the evolution will be characterized by a less important distribution among donors and acceptors, in number of species but also in number of different lineages impacted. The study of diversity within the identified clans thus allows us to partly distinguish LGTs established in the common ancestor of Archaeplastida, and thus potentially linked to primary plastid endosymbiosis, from later LGTs.

We listed the different distribution patterns of the transferred genes for each pipeline within Archaeplastida according to two types of analyses: i) analysis of the topological pattern of the selected trees, i.e. the number of LGT-receiving Archaeplastida lineages (Figure 18A and B), and ii) a more detailed analysis of the lineages involved in the transfers (Figure 18 C and D). Of the 57 trees selected by the chlamydial pipeline, 16 show branching with a single Archaeplastida lineage (28%), 17 with 2 lineages (30%), and 24 with 3 (42%). The majority of chlamydial gene transfers (72%) therefore impact at least 2 Archaeplastida lineages. Without taking into account the actinobacterial signal or the Firmicutes signal, since only 5 and 9 LGTs were identified in total, this proportion of multi-lineage transfers is higher than that observed for the other bacterial control groups, and rather similar to that of cyanobacteria. Indeed, the Bacteroidetes profile is the reverse of the Chlamydia profile, with the presence of 18 LGT (41%) in only 1 Archaeplastida line, 16 (36%) in 2 lines and 10 (23%) in all 3 lines (Figure 18A-B). This prevalence for transfers impacting only one Archaeplastida lineage is also visible for Proteobacteria since 17 unique LGTs (44%) are identified, against 6 LGTs (15%) in 2 lineages, and 16 (41%) in 3 lineages. The predominance of transfers to a single Archaeplastida lineage, for Bacteroidetes and Proteobacteria in particular, may reflect a phenomenon of lateral gene transfer that is not

concerted and/or occurs later than the emergence of these organisms. Therefore, we cannot assert their potential endosymbiotic origin in these cases. However, the proportions of LGT identified in 2 or 3 Archaeplastida lineages, respectively 72%, 59% and 56% for Chlamydia, Bacteroidetes and Proteobacteria, are not to be neglected since they highlight an antiquity of transfer, having taken place in the common ancestor of these organisms.

To go into more detail in the analysis of the diversity impacted by the identified LGTs, we listed the selected trees for each pipeline according to the Archaeplastida lineages receiving the transfer (Figure 18C-D). Thus, we can note different distribution patterns in Archaeplastida diversity depending on the bacterial group that initiated the transfer. First, we can see that, again, the Chlamydia and Cyanobacteria profiles are quite similar, with a majority proportion of transfers to the three Archaeplastida lineages (42 and 29% respectively), followed by joint transfers to the Rhodophyta and Viridiplantae (16 and 28%), and then only to the Viridiplantae (19 and 16%) (Figure 18D). In contrast, Bacteroidetes and Proteobacteria show a proportion of gene transfers only to Viridiplantae almost double that of the previous two groups, with 36 and 38% of LGTs identified, respectively. Bacteroidetes and Proteobacteria thus seem to have influenced the evolution of Viridiplantae more than Chlamydia and cyanobacteria. The proportion is reversed with respect to the impact of the different bacteria on Rhodophyta (Figure 18D), since Chlamydia and Cyanobacteria present this time a higher proportion of LGTs specifically to Rhodophyta compared to Bacteroidetes and Proteobacteria (respectively 9, 8, 5 and 5% of LGTs identified)

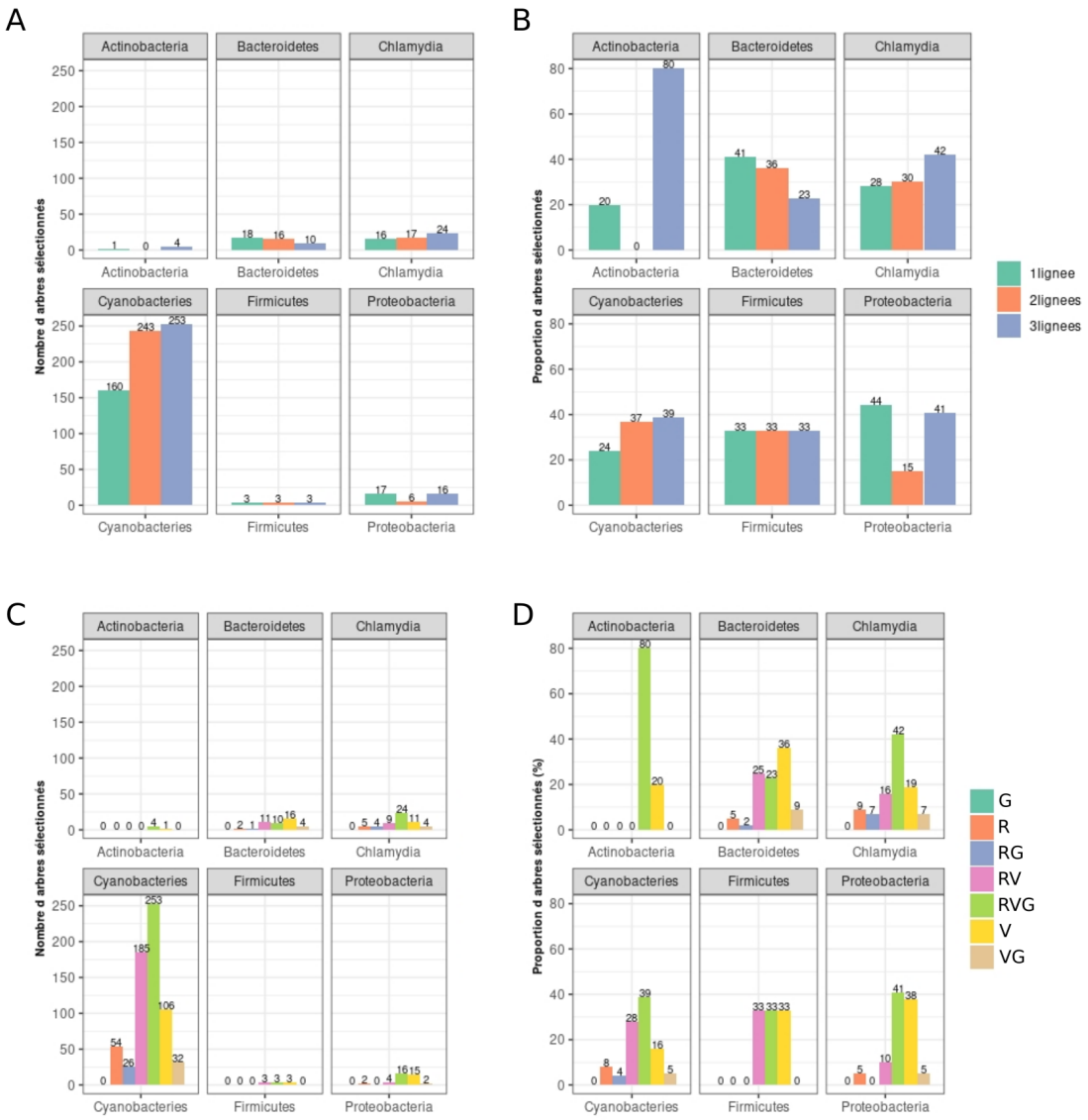


Figure 18: Tropism analysis of each selection. Panels A and B count the amount of trees branching 1, 2, and 3 Archaeplastida lineages for each pipeline in number (A) and proportion (B). Panels C and D further detail this diversity by counting trees based on the Archaeplastida lineages involved, in number (C) and proportion (D). G = Glaucophyta; R = Rhodophyta, V = Viridiplantae.

So far, we have evaluated the different phylogenetic pattern of the identified LGTs in a binary way (presence, absence of an Archaeplastida lineage), without really taking into account the total diversity within clans. Recall that the selection criteria of the automatic pipeline identify each tree for which there is a phylogenetic interaction between at least two donors and three acceptors. However, the diversity of organisms plays a role in the interpretation of LGTs and especially in their temporality.

We therefore analyzed the proportions of donors and acceptors within the selected clans for the different pipelines. Apart from Cyanobacteria, no significant difference is observable between the different pipelines, either at the level of donors (Figure 19B and E) or acceptors

(Figure 19C and F), but also in the total number of species included in the clans (Figure 19A and D). The distributions are similar, so the orthologous groups generally branch the same number of donors and acceptors. Indirectly, these results also show the robustness of the pipeline, since for each control condition tested, the diversity of clans is equivalent. Taken together, the acceptors show a similar distribution pattern for the different pipelines. However, separate evaluation of the Archaeplastida lineages reveals some differences (Figure 19 G to K). This feature is verified by analyzing the distribution of the number of Viridiplantae within the selected clans. Indeed, a higher proportion of Viridiplantae is observed within the Bacteroidetes selection, all pattern combined, compared to the other signals studied (Figure 19G). According to a Kruskal-Wallis statistical test, this signal is significantly larger in Bacteroidetes compared to the other bacterial groups (Figure 19G). This peculiarity of the joint evolution of Bacteroidetes and Viridiplantae may indicate more continuous gene transfers during the evolution of these organisms. Similarly, we also analyzed the presence of intruders in the selected clans. Here, we call "intruders" non-target species, whether eukaryotes or bacteria, that are part of the subtree identified by the pipeline, thus interrupting the monophyly of target species in the clan. Photosynthetic eukaryotes present from secondary endosymbiosis of a red or green alga are not considered intruders here. These intruders may reveal an ancient gene sharing, going back to the common ancestor of all these organisms and followed by multiple losses, and thus in this case potentially invalidate the hypothesis of endosymbiotic transfer, a later transfer or contamination, but also a phylogenetic artifact. Thus, again, the diversity of organisms in clans, and particularly here that of intruders, can help differentiate between signals related to primary plastid endosymbiosis, for which low numbers of intruders are expected, and those that are not. Analysis of the number of non-target species in clans, based on each pipeline studied, highlights a higher proportion of intruders in clans selected for the Bacteroidetes pipeline (Figure 19J), significantly different from the Chlamydia, Cyanobacteria and Proteobacteria pipelines. This, along with the prevalence of Viridiplantae as acceptors, suggests a different profile for Bacteroidetes.

When evaluating the distributions of donors and acceptors separately, no differences emerge, except for Viridiplantae and intruders. However, when the acceptors and donors of these gene transfers are examined jointly within clans, different patterns emerge depending on the pipelines. Again, if we remove the results obtained for the Firmicutes and Actinobacteria pipelines, since the number of identified gene transfers is much lower than for the other bacterial groups, we can distinguish two different profiles, grouping Proteobacteria and Bacteroidetes on one side, and Chlamydia and Cyanobacteria on the other. Indeed, Figure 20 visualizes in part the relationship between the different target species of each clan selected by the pipelines. In general, we can observe a correlation between the numbers of donors and acceptors of gene transfers. For the Chamydian and cyanobacterial pipelines, the number of Archaeplastida species seems to increase along with the number of target bacteria in the clan.

Conversely, for Bacteroidetes and Proteobacteria, it appears that the increase in the number of donors in the clan is related to the decrease in the number of photosynthetic target eukaryotes. To summarize, a gene transfer in Proteobacteria and Bacteroidetes affects either a higher diversity of bacteria or a higher diversity of Archaeplastida, whereas for Chlamydia and cyanobacteria, these gene transfers affect more intensely both bacterial and Archaeplastida diversity. This particularity could reflect a different evolutionary history depending on the bacterial groups studied, rather based on lateral transfers of genes dispersed during evolution on the one hand, and from endosymbiotic events on the other. Moreover, a greater diversity of target species, notably Archaeplastida, within the identified clans necessarily implies a transfer of the gene in the common ancestor of these organisms, potentially going back to the primary endosymbiosis of the plastid. Depending on the tropism of these transfers and the diversity of the impacted organisms, we can imagine four different scenarios, combining two dimensions, depth and density. i) In the case of a deep and dense scenario, the gene is transferred to the last common ancestor (LCA for Last Common Ancestor) and conserved since, it will thus be visible in more than one lineage of Archaeplastida and a significant diversity of acceptors. ii) For the second scenario, deep and rare, either the gene was transferred to the LCA but lost or replaced since then, or the transfer occurred several times immediately following the emergence of the LCA, in which case, although the presence of the gene is visible in more than one lineage of Archaeplastida, the diversity represented is less important. iii) the superficial and dense scenario represents a gene transferred later but completely conserved in the acceptor subgroup, and finally iv) the last scenario, superficial and rare, corresponding to multiple independent transfers, will be visualized by a low diversity, especially of acceptors, within the clans, which mostly concern only one lineage of Archaeplastida

The proportion of gene transfers to a single lineage of Archaeplastida, their privileged impact on Viridiplantae, as well as the donor profiles depending on the different acceptors, may testify to lateral gene transfers dispersed within the evolution of its organisms, without direct linkage to primary plastid endosymbiosis, particularly for Bacteroidetes. However, a majority of LGTs, including those from Bacteroidetes, are observable in two or three lineages of Archaeplastida, implying an earlier timing. Thus, although there appear to be different diversity patterns between Chlamydia, cyanobacteria and other bacterial control groups, this analysis is not yet sufficient to confirm or refute MATH.

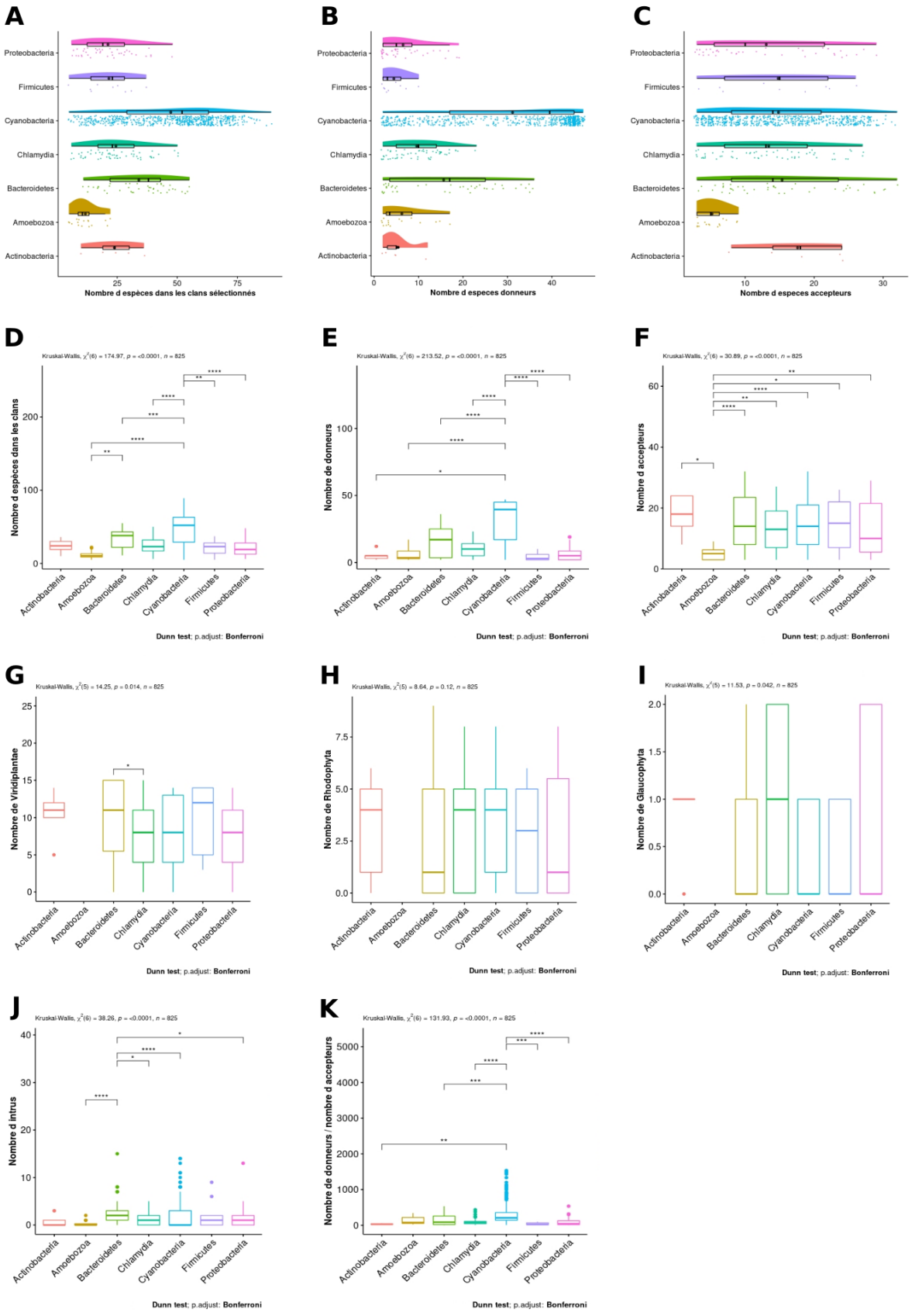


Figure 19: Diversity analysis of selected clans for each pipeline. For each pipeline studied, the distribution of the number of species in clans (A) was analyzed and then detailed into donors (B) and acceptors (C). A Kruskal Wallis statistical test assesses differences in diversity across pipelines. Several parameters are reported to evaluate the diversity of the selected clans, namely the number of species (D), the number of donors (E), and

acceptors (F), but also the diversity in Viridiplantae (G), Rhodophyta (H), and Glaucophytes (I), but also the proportion of intruders (J). The last parameter analyzed is the ratio between donors and acceptors in the clans (K). The specific parameters for the Kruskal Willis test for each condition evaluated is indicated on the graphs. Significant differences between pipelines are represented by stars (*).

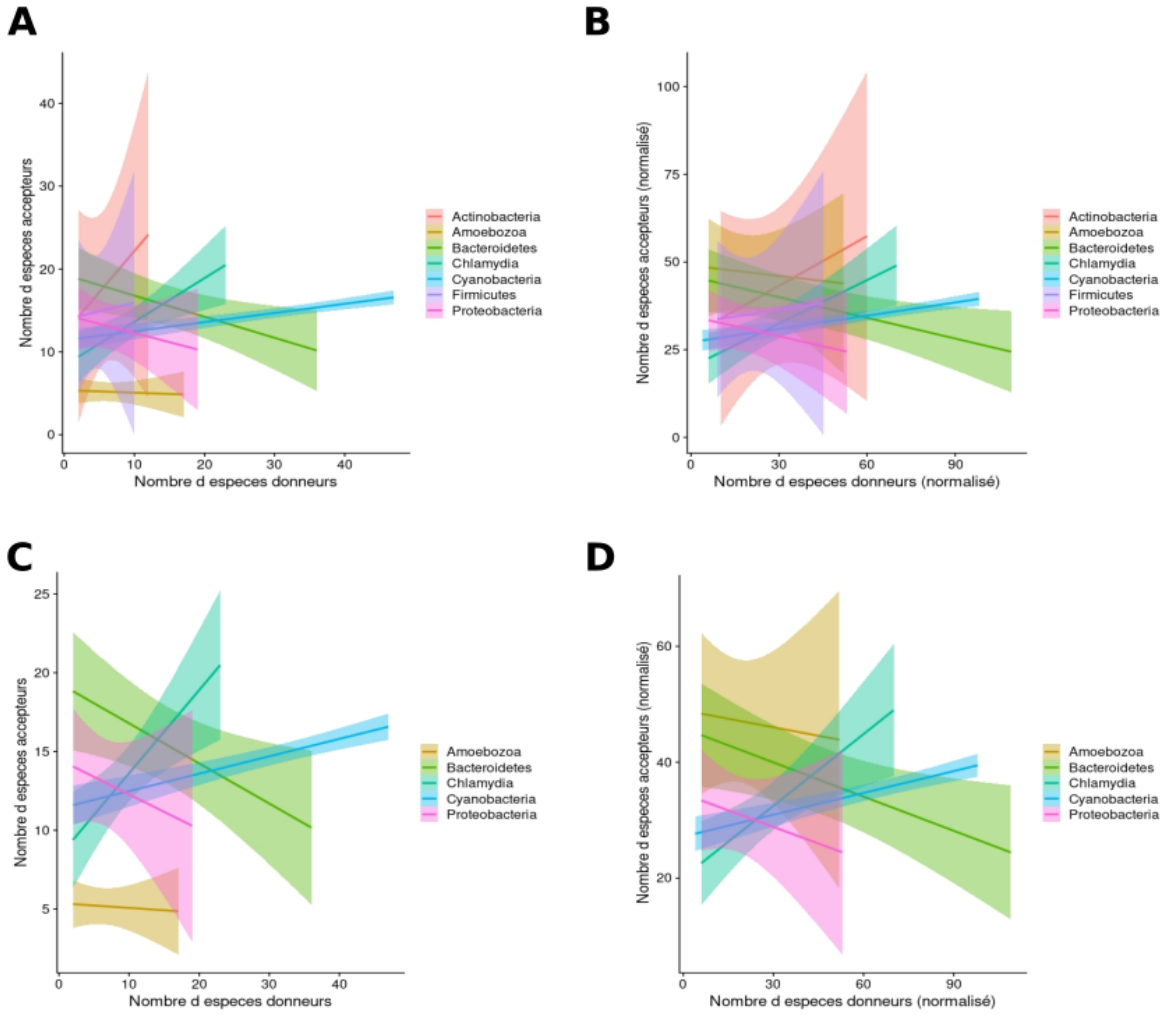


Figure 20: Analysis of the number of acceptors in the selected clans as a function of the number of donors. For each pipeline, the diversity of acceptors in the clans as a function of the number of donors is reported. In (A) are represented the raw data, in absolute number of species present in the selected clans for donors and for acceptors. In (B), these data are normalized with the number of donor and acceptor species entering the pipeline. For graphs (C) and (D), the Actinobacteria and Firmicutes pipelines have been removed, for clarity (contain only 5 and 9 selected clans respectively).

d. Functional annotations

Beyond the phylogenetic analysis, which allows us to discern the nature of the signals studied and to clarify the specific impact of one bacterial group compared to another, a functional study of the identified gene transfers can provide further information to assess the plausibility of the Ménage à Trois hypothesis, especially regarding the establishment of biochemical flows at the origin of the metabolic integration of cyanobacteria. For each

selection in each pipeline, we therefore annotated the sequences present in the branches of interest using EggNog Mapper and BlastKoala. We generalized to the entire clan when at least 50% of the sequences present were annotated in the same way (Table 1, appendix 8).

Table 1: Summary table of the functional annotation of the 57 trees selected by the Chlamydian automatic pipeline. The annotation of each selected clan, performed by eggNOG mapper and BlastKOALA, was generalized if at least 50% of the sequences were annotated in the same way. The tropism of each clan is indicated in the second column. R: Rhodophyta, V: Viridiplantae, G: Glaucophyta.

Chlamydia Pipeline				
OG	**Tropism**	**Protein identification**	**No. Kegg**	**Function**
OG0000013	RVG	NLRC3, NOD3; NLR family CARD domain-containing protein 3	K22614	Signaling and cellular processes
OG0000134	R	aqpZ; aquaporin Z	K06188	**Transporters**
OG0000479	RVG	aroDE, DHQ-SDH; 3-dehydroquinate dehydratase / shikimate dehydrogenase [EC:4.2.1.10 1.1.1.25]	K13832	Amino acid metabolism
OG0000649	VG	ksgA; 16S rRNA (adenine1518-N6/adenine1519-N6)-dimethyltransferase [EC:2.1.1.182]	K02528	Ribosome biogenesis
OG0000674	V	TC.PIT; inorganic phosphate transporter, PiT family	K03306	**Transporters**
OG0000812	VG	K07146; UPF0176 protein	K07146	
OG0000904	V	rluB; 23S rRNA pseudouridine2605 synthase [EC:5.4.99.22]	K06178	Ribosome biogenesis
OG0000913	RVG	fabF, OXSM, CEM1; 3-oxoacyl-[acyl-carrier-protein] synthase II [EC:2.3.1.179]	K09458	Lipid metabolism - Metabolism of cofactors and vitamins
OG0001000	V	NSF, SEC18; vesicle-fusing ATPase [EC:3.6.4.6]	K06027	Membrane trafficking
OG0001059	V	E1.1.1.82; malate dehydrogenase (NADP+) [EC:1.1.1.82].	K00024	Carbohydrate metabolism - Energy metabolism - Amino acid metabolism
OG0001078	RV	ribBA; 3,4-dihydroxy 2-butanone 4-phosphate synthase / GTP cyclohydrolase II [EC:4.1.99.12 3.5.4.25]	K14652	Metabolism of cofactors and vitamins
OG0001293	VG	glgA; starch synthase [EC:2.4.1.21].	K00703	Carbohydrate metabolism
OG0001410	RVG	RP-L24, MRPL24, rplX; large subunit ribosomal protein L24	K02895	Ribosome
OG0001468	R	trpC; indole-3-glycerol phosphate synthase [EC:4.1.1.48].	K01609	Amino acid metabolism
OG0001851	RV	mraW, rsmH; 16S rRNA (cytosine1402-N4)-methyltransferase [EC:2.1.1.199]	K03438	Ribosome biogenesis
OG0001950	RV	trxB, TRR; thioredoxin reductase (NADPH) [EC:1.8.1.9].	K00384	Metabolism of other amino acids
OG0002167	RVG	PHYH; phytanoyl-CoA hydroxylase [EC:1.14.11.18].	K00477	Peroxisome
OG0002222	RVG	SAL; 3'(2'), 5'-bisphosphate nucleotidase / inositol polyphosphate 1-phosphatase [EC:3.1.3.7 3.1.3.57]	K15422	Carbohydrate metabolism - Energy metabolism
OG0002300	R	PTH1, pth, spoVC; peptidyl-tRNA hydrolase, PTH1 family [EC:3.1.1.29]	K01056	Translation factors
OG0002395	RVG	truA, PUS1; tRNA pseudouridine38-40 synthase [EC:5.4.99.12]	K06173	Transfer RNA biogenesis
OG0002498	RV	YARS, tyrS; tyrosyl-tRNA synthetase [EC:6.1.1.1]	K01866	Transfer RNA biogenesis
OG0002584	RV	pnp, PNPT1; polyribonucleotide nucleotidyltransferase [EC:2.7.7.8]	K00962	Messenger RNA biogenesis
OG0002591	RVG			
OG0003272	RVG	trpD; anthranilate phosphoribosyltransferase [EC:2.4.2.18]	K00766	Amino acid metabolism
OG0003309	RG	rnhB; ribonuclease HII [EC:3.1.26.4]	K03470	DNA replication proteins
OG0003312	RVG	tyrP; tyrosine-specific transport protein	K03834	**Transporters**
OG0003383	RV		K03319	**Transporters**
OG0003449	RVG	TC.AAA; ATP:ADP antiporter, AAA family	K03301	**Transporters**
OG0003873	RVG	ISA, treX; isoamylase [EC:3.2.1.68]	K01214	Carbohydrate metabolism

OG0003961	V	fabI; enoyl-[acyl-carrier protein] reductase I [EC:1.3.1.9 1.3.1.10]	K00208	Lipid metabolism - Metabolism of cofactors and vitamins
OG0004281	RV	K09858; SEC-C motif domain protein	K09858	
OG0004382	RVG			
OG0004493	RVG	Na H antiporter		**Transporters**
OG0004746	RVG	uhpC; MFS transporter, OPA family, sugar phosphate sensor protein UhpC	K07783	**Transporters**
OG0004766	RVG	ispE; 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase [EC:2.7.1.148]	K00919	Metabolism of terpenoids and polyketides
OG0004954	RVG			
OG0005053	V	gcpE, ispG; (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase [EC:1.17.7.1 1.17.7.3]	K03526	Metabolism of terpenoids and polyketides
OG0005097	RVG			
OG0005231	RVG	E2.6.1.83; LL-diaminopimelate aminotransferase [EC:2.6.1.83].	K10206	Amino acid metabolism
OG0005232	RVG	ispD; 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase [EC:2.7.7.60]	K00991	Metabolism of terpenoids and polyketides
OG0005255	RVG			
OG0005308	RV	CHS; chalcone synthase [EC:2.3.1.74]	K00660	Biosynthesis of other secondary metabolites
OG0005374	R	rlmH; 23S rRNA (pseudouridine1915-N3)-methyltransferase [EC:2.1.1.177]	K00783	Ribosome biogenesis
OG0005382	RVG		K03215	Ribosome biogenesis
OG0005581	RVG	ATS1; glycerol-3-phosphate O-acyltransferase [EC:2.3.1.15].	K00630	Lipid metabolism
OG0006000	V	kdsB; 3-deoxy-manno-octulosonate cytidylyltransferase (CMP-KDO synthetase) [EC:2.7.7.38]	K00979	Glycan biosynthesis and metabolism
OG0007168	RVG			
OG0008425	VG	UGP3; UTP---glucose-1-phosphate uridylyltransferase [EC:2.7.7.9]	K22920	Lipid metabolism
OG0008763	V			Ribosome biogenesis
OG0008957	V			
OG0008974	RV	ddl; D-alanine-D-alanine ligase [EC:6.3.2.4]	K01921	Metabolism of other amino acids - Glycan biosynthesis and metabolism
OG0009869	RG	apbE; FAD:protein FMN transferase [EC:2.7.1.180]	K03734	
OG0014617	V	dnaQ; DNA polymerase III subunit epsilon [EC:2.7.7.7].	K02342	DNA replication proteins
OG0017499	RG	wbpA; UDP-N-acetyl-D-glucosamine dehydrogenase [EC:1.1.1.136]	K13015	Carbohydrate metabolism - Glycan biosynthesis and metabolism
OG0017872	R		K01046	Glycerolipid metabolism
OG0024221	V	murB; UDP-N-acetylmuramate dehydrogenase [EC:1.3.1.98].	K00075	Carbohydrate metabolism - Glycan biosynthesis and metabolism
OG0028045	RG	queD, ptpS, PTS; 6-pyruvoyltetrahydropterin/6-carboxytetrahydropterin synthase [EC:4.2.3.12 4.1.2.50]	K01737	Metabolism of cofactors and vitamins

At first sight, the identified gene transfers impact multiple metabolic pathways and cellular processes. However, we can observe a predominance of transporter transfers among the genes from Chlamydia. In an endosymbiotic context, the presence of transporters is necessary for the establishment of biochemical flows and thus lead to metabolic integration. Indeed, out of the 57 orthologous groups selected by the automatic pipeline, 7 are annotated as transporters, among which we can find the key transporters of MATH (namely UhpC, an

ATP transporter and TyrP). For Bacteroidetes, no such protein was identified. On the other hand, Proteobacteria seem to have transmitted two to Archaeplastida. By further analyzing these proteins in particular, we can notice that these two transporters identified for Proteobacteria are also identified in the chlamydial pipeline, which appears incompatible. However, manual analysis of the corresponding trees shows the branching of two Proteobacteria with Archaeplastida, then directly with Chlamydia for the tree corresponding to TyrP (Figure 21) and the ADP-ATP transporter (Figure 23). Analysis of the corresponding trees selected by the Chlamydian pipeline supports the identification of a paralogous clan composed of both Proteobacteria and Stramenopiles of the TyrP tree (Figure 22). With respect to the ADP-ATP transporter, the tree selected by the Chlamydial pipeline has no Proteobacteria, despite enrichment with a bacterial dataset consisting of 27 of these organisms (Figure 24, appendix 9). It is highly likely in these cases that the identified genes are chlamydial, transferred in Archaeplastida during primary plastid endosymbiosis, and in the two Proteobacteria in question during lateral gene transfer events. In an endosymbiotic context, the presence of transporters is necessary for the establishment of biochemical flows and thus lead to metabolic integration. This proportion of chlamydial transporters found in Archaeplastida and absent in the other bacterial groups studied, supports the hypothesis of a particular involvement of these pathogens during primary plastid endosymbiosis.

The visualization of selected gene transfers on metabolic maps allows us to realize the impact of certain groups on the metabolism of Archaeplastida. Indeed, Chlamydia seem to play a more pronounced role in certain metabolic pathways, compared to other bacterial groups, in particular due to the transfer of several genes impacting the same pathways. For example, the tryptophan biosynthesis pathway, already studied by Cenci et al, for which 4 of the 9 steps are catalyzed by enzymes of Chlamydian origin, is also identified in our results. This particularity is also found for the synthesis of fatty acids and isoprenoids and seems to be specific to Chlamydia-affiliated genes. For the other bacterial groups studied, gene transfers are dispersed within the metabolism and not grouped into metabolic pathways. An exception exists, however, since the Bacteroidetes impact the fatty acid synthesis pathway on two enzymes: FabZ and FATA. Thus, a brief analysis of the functions of the gene transfers selected by the pipelines reveals a Chlamydian profile that is different from other bacterial groups, impacting more specifically on transporters and revealing multiple transfers in certain metabolic pathways.

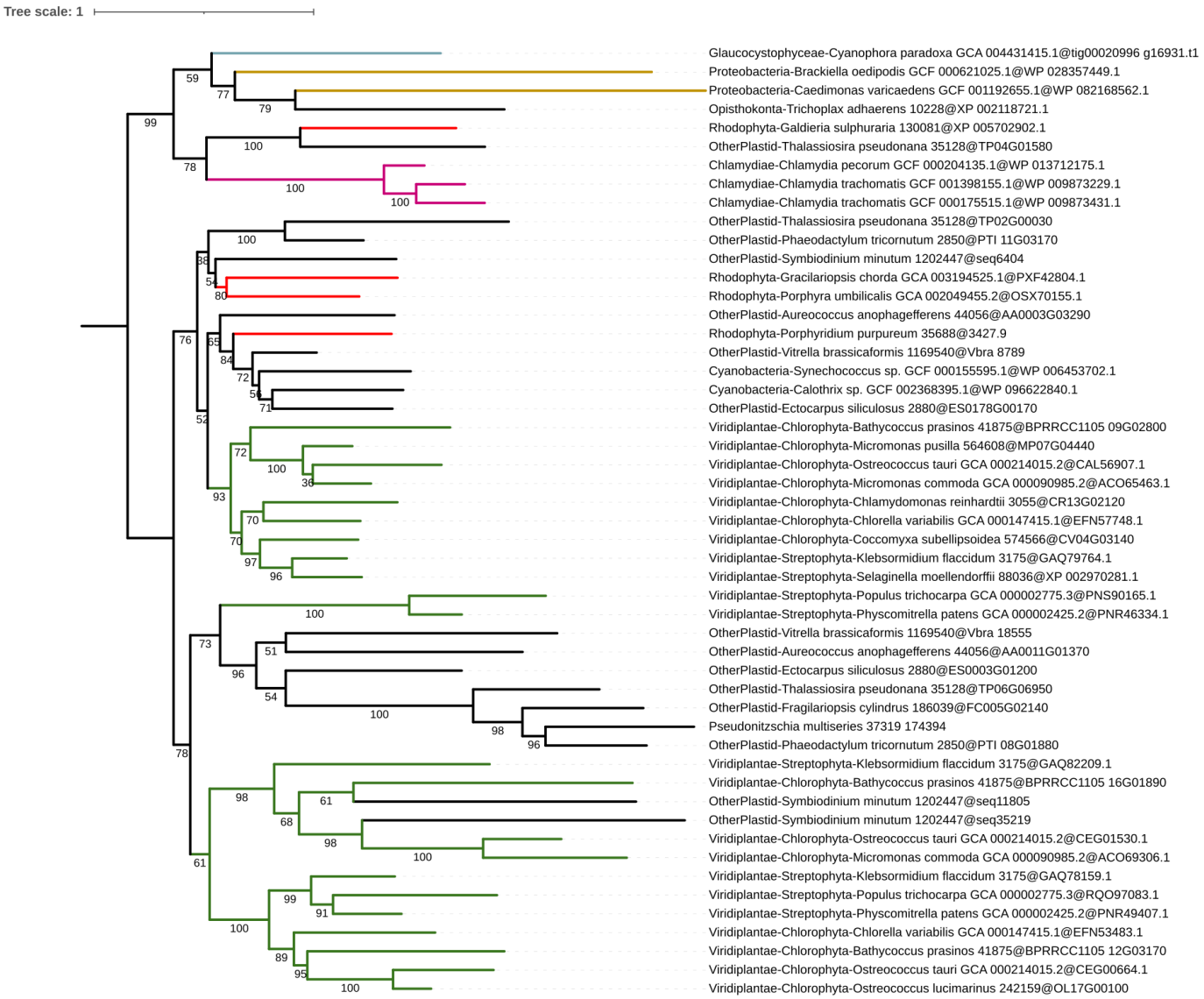


Figure 21: Single gene phylogenetic tree of the TyrP transporter identified by the Proteobacteria pipeline. This tree corresponds to the phylogenetic reconstruction of the orthologous group OG0003312, annotated as ATP:ADP antiporter, selected by the Proteobacteria pipeline. The tree was obtained by IQ-TREE, under an LG4X ultrafast-bootstrap model. Species identified as "other-plastid" are not considered as intruders, and are therefore part of the clan selection criteria. Bootstrap values are indicated on the branches. The rooting is in mid-point. The scale counts in number of amino acid substitutions per site. In red: Rhodophyta and Rhodelphea, in green: Viridiplantae and in blue: Glaucophyta. In pink: Chlamydia, in yellow: Proteobacteria


ee scale: 1
Glaucocystophyceae-Cyanophora paradoxa GCA 004431415.1@tig00020996 g16931.t1
Viridiplantae-Streptophyta-Physcomitrella patens GCA 000002425.2@PNR46334.1
Viridiplantae-Streptophyta-Arabidopsis thaliana 3702@AT2G33260
Viridiplantae-Streptophyta-Klebsormidium flaccidum 3175@GAQ82209.1
Viridiplantae-Chlorophyta-Bathycoccus prasinos 41875@BPRRCC1105 16G01890
OtherPlastid-Symbiodinium minutum 1202447@seq11805
OtherPlastid-Symbiodinium minutum 1202447@seq35219
Viridiplantae-Chlorophyta-Ostreococcus tauri GCA 000214015.2@CEG01530.1
Viridiplantae-Chlorophyta-Micromonas commoda GCA 000090985.2@ACO69306.1
Viridiplantae-Chlorophyta-Chlorella variabilis GCA 000147415.1@EFN53483.1
Viridiplantae-Chlorophyta-Bathycoccus prasinos 41875@BPRRCC1105 12G03170
Viridiplantae-Chlorophyta-Ostreococcus tauri GCA 000214015.2@CEG00664.1
Viridiplantae-Chlorophyta-Ostreococcus lucimarinus 242159@OL17G00100
Viridiplantae-Streptophyta-Mesostigma viride GCA 004000685.1@BIFU01004623.1
Viridiplantae-Streptophyta-Klebsormidium flaccidum 3175@GAQ78159.1
Viridiplantae-Streptophyta-Populus trichocarpa GCA 000002775.3@RQO97083.1
Viridiplantae-Streptophyta-Physcomitrella patens GCA 000002425.2@PNR49407.1
Viridiplantae-Chlorophyta-Bathycoccus prasinos 41875@BPRRCC1105 09G02800
Viridiplantae-Chlorophyta-Ostreococcus tauri GCA 000214015.2@CAL56907.1
Viridiplantae-Chlorophyta-Micromonas commoda GCA 000090985.2@ACO65463.1
Viridiplantae-Chlorophyta-Micromonas pusilla 564608@MP07G04440
Viridiplantae-Chlorophyta-Volvox carteri GCA 000143455.1@EFJ49671.1
Viridiplantae-Chlorophyta-Chlorella variabilis GCA 000147415.1@EFN57748.1
Viridiplantae-Chlorophyta-Coccomyxa subellipsoidea 574566@CV04G03140
Viridiplantae-Streptophyta-Selaginella moellendorffii 88036@XP 002978446.1
Viridiplantae-Streptophyta-Klebsormidium flaccidum 3175@GAQ79764.1
Viridiplantae-Streptophyta-Populus trichocarpa GCA 000002775.3@RQO85218.1
Viridiplantae-Streptophyta-Physcomitrella patens GCA 000002425.2@PNR45804.1
OtherPlastid-Symbiodinium minutum 1202447@seq6404
OtherPlastid-Aureococcus anophagefferens 44056@AA0003G03290
OtherPlastid-Thalassiosira pseudonana 35128@TP02G00030
OtherPlastid-Pseudonitzschia multiseries 37319@188974
OtherPlastid-Phaeodactylum tricornutum 2850@PTI 11G03170
OtherPlastid-Ectocarpus siliculosus 2880@ES0178G00170
OtherPlastid-Vitrella brassicaformis 1169540@Vbra 8789
Cyanobacteria-Synechococcus sp. GCF 000155595.1@WP 006453702.1
Cyanobacteria-Calothrix sp. GCF 002368395.1@WP 096622840.1
Rhodophyta-Porphyridium purpureum 35688@3427.9
Rhodophyta-Porphyra umbilicalis GCA 002049455.2@OSX70155.1
Rhodophyta-Gracilariopsis lemaneiformis GCA 003346895.1@NFUL01006118.1
Rhodophyta-Chondrus crispus 2769@XP 005713394.1
OtherPlastid-Vitrella brassicaformis 1169540@Vbra 18555
OtherPlastid-Ectocarpus siliculosus 2880@ES0003G01200
OtherPlastid-Aureococcus anophagefferens 44056@AA0011G01370
OtherPlastid-Thalassiosira pseudonana 35128@TP06G06950
OtherPlastid-Fragilariopsis cylindrus 186039@FC003G06610
OtherPlastid-Fragilariopsis cylindrus 186039@FC005G02140
OtherPlastid-Phaeodactylum tricornutum 2850@PTI 08G01880
Rhodophyta-Galdieria sulphuraria 130081@XP 005702902.1
OtherPlastid-Thalassiosira pseudonana 35128@TP04G01580
Chlamydiae-Simkania negevensis GCF 000237205.1@WP 013944641.1
Chlamydiae-Chlamydiae bacterium GCA 001794905.1@OGN62444.1
Chlamydiae-Parachlamydia acanthamoebae GCF 000253035.1@WP 013924112.1
Chlamydiae-Protochlamydia sp. GCF 001950615.1@WP 075882953.1
Chlamydiae-Waddlia chondrophila GCF 000092785.1@WP 013181292.1
Chlamydiae-Neochlamydia sp. GCF 000648235.2@WP 042240804.1
Chlamydiae-Estrella lausannensis GCF 900000175.1@WP 098038658.1
Chlamydiae-Criblamydia sequanensis GCF 000750955.1@WP 041017870.1
Chlamydiae-Parachlamydiaceae bacterium GCF 000829755.1@WP 042105790.1
Chlamydiae-Parachlamydia sp. GCF 001545115.1@WP 079989498.1
Chlamydiae-Parachlamydia acanthamoebae GCF 000253035.1@WP 006339747.1
Chlamydiae-Simkania negevensis GCF 000237205.1@WP 013943875.1
Chlamydiae-Criblamydia sequanensis GCF 000750955.1@WP 079977941.1
Chlamydiae-Estrella lausannensis GCF 900000175.1@WP 098037790.1
Chlamydiae-Protochlamydia naegleriophila GCF 001499655.1@WP 079992843.1
Proteobacteria-Escherichia coli GCF 000026545.1@WP 000797555.1
Proteobacteria-Neisseria meningitidis GCF 000083565.1@WP 012779567.1
Proteobacteria-Escherichia coli GCF 000026545.1@WP 000233849.1
Proteobacteria-Escherichia coli GCF 000026545.1@WP 000224347.1
Chlamydiae-Chlamydiae bacterium GCA 001794905.1@OGN62445.1
Chlamydiae-Simkania negevensis GCF 000237205.1@WP 013944638.1
Opisthokonta-Trichoplax adhaerens 10228@XP 002118721.1
Chlamydiae-Parachlamydiaceae bacterium GCF 000829755.1@WP 042105788.1
Chlamydiae-Neochlamydia sp. GCF 000648235.2@WP 042240807.1
Chlamydiae-Waddlia chondrophila GCF 000092785.1@WP 013181293.1
Chlamydiae-Parachlamydia sp. GCF 001545115.1@WP 068470789.1
Chlamydiae-Parachlamydia acanthamoebae GCF 000253035.1@WP 006341673.1
Chlamydiae-Criblamydia sequanensis GCF 000750955.1@WP 041017871.1
Chlamydiae-Estrella lausannensis GCF 900000175.1@WP 098039227.1
Chlamydiae-Waddlia chondrophila GCF 000092785.1@WP 013181175.1
Chlamydiae-Simkania negevensis GCF 000237205.1@WP 013943921.1
Chlamydiae-Neochlamydia sp. GCF 000813645.1@WP 039386625.1
Chlamydiae-Chlamydiae bacterium GCA 001794905.1@OGN61662.1
Chlamydiae-Parachlamydiaceae bacterium GCF 000829755.1@WP 042102609.1
Chlamydiae-Protochlamydia naegleriophila GCF 001499655.1@WP 051981685.1
Chlamydiae-Neochlamydia sp. GCF 000813645.1@WP 052243958.1
Chlamydiae-Chlamydiales bacterium GCF 900093655.1@WP 088206379.1
Chlamydiae-Waddlia chondrophila GCF 000092785.1@WP 013181675.1
Chlamydiae-Parachlamydia acanthamoebae GCF 000253035.1@WP 013925138.1
Chlamydiae-Estrella lausannensis GCF 900000175.1@WP 098038183.1
Chlamydiae-Chlamydiae bacterium GCA 001794905.1@OGN62198.1
Chlamydiae-Chlamydia muridarum GCF 003433255.1@WP 100065006.1
Chlamydiae-Chlamydia suis GCF 900149385.2@WP 080124562.1
Chlamydiae-Chlamydia ibidis GCF 000454725.1@WP 020370494.1
Chlamydiae-Chlamydia sp. GCF 001653975.1@WP 066482341.1
Chlamydiae-Simkania negevensis GCF 000237205.1@WP 013944187.1
Chlamydiae-Chlamydiales bacterium GCF 900093655.1@WP 088206541.1
Chlamydiae-Waddlia chondrophila GCF 000092785.1@WP 041941473.1
Chlamydiae-Parachlamydia sp. GCF 001545115.1@WP 068470028.1
Chlamydiae-Parachlamydiaceae bacterium GCF 000829755.1@WP 042104409.1
Chlamydiae-Neochlamydia sp. GCF 000648235.2@WP 042243128.1
Chlamydiae-Estrella lausannensis GCF 900000175.1@WP 098039013.1
Chlamydiae-Criblamydia sequanensis GCF 000750955.1@WP 041017697.1

Figure 22: Single gene phylogenetic tree of the TyrP transporter identified by the Chlamydia pipeline. (previous page) This tree corresponds to the phylogenetic reconstruction of the orthologous group OG0003312, annotated as TyrP, selected by the Chlamydia pipeline. The tree was obtained by IQ-TREE, under an LG4X ultrafast-bootstrap model. Species identified as "other-plastid" are not considered as intruders, and are therefore part of the clan selection criteria. Bootstrap values are indicated on the branches. The rooting is in mid-point. The scale counts in number of amino acid substitutions per site. In red: Rhodophyta and Rhodelphea, in green: Viridiplantae and in blue: Glaucophyta. In pink: Chlamydia, in yellow: Proteobacteria

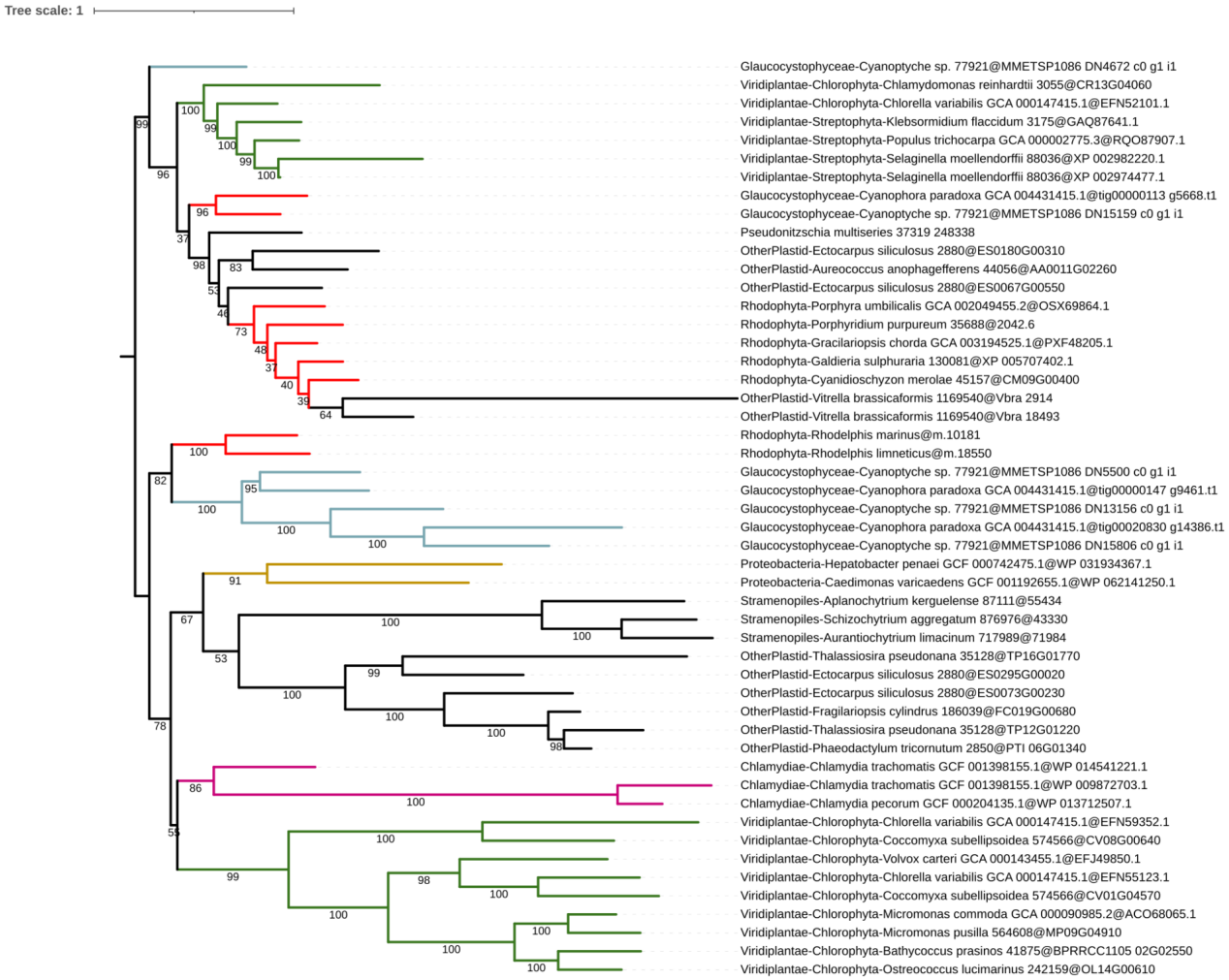


Figure 23: Single gene phylogenetic tree of the ATP:ADP antiporter identified by the Proteobacteria pipeline. This tree corresponds to the phylogenetic reconstruction of the orthologous group OG0003449, annotated as ATP:ADP antiporter, selected by the Proteobacteria pipeline. The tree was obtained by IQ-TREE, under an LG4X ultrafast-bootstrap model. Species identified as "other-plastid" are not considered as intruders, and are therefore part of the clan selection criteria. Bootstrap values are indicated on the branches. The rooting is in mid-point. The scale counts in number of amino acid substitutions per site. In red: Rhodophyta and Rhodelphea, in green: Viridiplantae and in blue: Glaucophyta. In pink: Chlamydia, in yellow: Proteobacteria

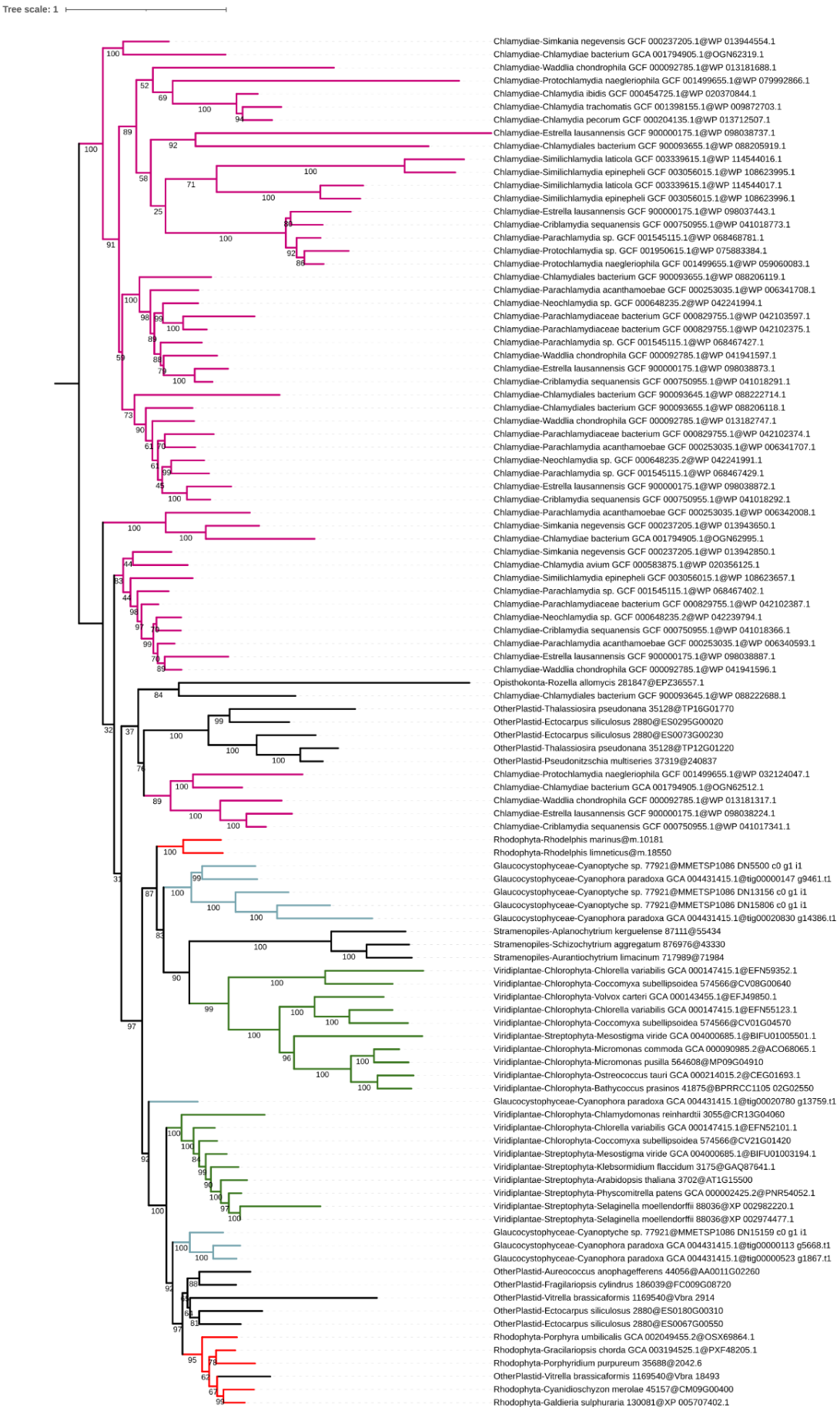

Tree scale: 1
Chlamydiae-Simkania negevensis GCF 000237205.1@WP 013944554.1
Chlamydiae-Chlamydiae bacterium GCA 001794905.1@OGN62319.1
Chlamydiae-Waddlia chondrophila GCF 000092785.1@WP 013181688.1
Chlamydiae-Protochlamydia naegleriophila GCF 001499655.1@WP 079992866.1
Chlamydiae-Chlamydia ibidis GCF 000454725.1@WP 020370844.1
Chlamydiae-Chlamydia trachomatis GCF 001398155.1@WP 009872703.1
Chlamydiae-Chlamydia pecorum GCF 000204135.1@WP 013712507.1
Chlamydiae-Estrella lausannensis GCF 900000175.1@WP 098038737.1
Chlamydiae-Chlamydiales bacterium GCF 900093655.1@WP 088205919.1
Chlamydiae-Similichlamydia laticola GCF 003339615.1@WP 114544016.1
Chlamydiae-Similichlamydia epinepheli GCF 003056015.1@WP 108623995.1
Chlamydiae-Similichlamydia laticola GCF 003339615.1@WP 114544017.1
Chlamydiae-Similichlamydia epinepheli GCF 003056015.1@WP 108623996.1
Chlamydiae-Estrella lausannensis GCF 900000175.1@WP 098037443.1
Chlamydiae-Criblamydia sequanensis GCF 000750955.1@WP 041018773.1
Chlamydiae-Parachlamydia sp. GCF 001545115.1@WP 068468781.1
Chlamydiae-Protochlamydia sp. GCF 001950615.1@WP 075883384.1
Chlamydiae-Protochlamydia naegleriophila GCF 001499655.1@WP 059060083.1
Chlamydiae-Chlamydiales bacterium GCF 900093655.1@WP 088206119.1
Chlamydiae-Parachlamydia acanthamoebae GCF 000253035.1@WP 006341708.1
Chlamydiae-Neochlamydia sp. GCF 000648235.2@WP 042241994.1
Chlamydiae-Parachlamydiaceae bacterium GCF 000829755.1@WP 042103597.1
Chlamydiae-Parachlamydiaceae bacterium GCF 000829755.1@WP 042102375.1
Chlamydiae-Parachlamydia sp. GCF 001545115.1@WP 068467427.1
Chlamydiae-Waddlia chondrophila GCF 000092785.1@WP 041941597.1
Chlamydiae-Estrella lausannensis GCF 900000175.1@WP 098038873.1
Chlamydiae-Criblamydia sequanensis GCF 000750955.1@WP 041018291.1
Chlamydiae-Chlamydiales bacterium GCF 900093645.1@WP 088222714.1
Chlamydiae-Chlamydiales bacterium GCF 900093655.1@WP 088206118.1
Chlamydiae-Waddlia chondrophila GCF 000092785.1@WP 013182747.1
Chlamydiae-Parachlamydiaceae bacterium GCF 000829755.1@WP 042102374.1
Chlamydiae-Parachlamydia acanthamoebae GCF 000253035.1@WP 006341707.1
Chlamydiae-Neochlamydia sp. GCF 000648235.2@WP 042241991.1
Chlamydiae-Parachlamydia sp. GCF 001545115.1@WP 068467429.1
Chlamydiae-Estrella lausannensis GCF 900000175.1@WP 098038872.1
Chlamydiae-Criblamydia sequanensis GCF 000750955.1@WP 041018292.1
Chlamydiae-Parachlamydia acanthamoebae GCF 000253035.1@WP 006342008.1
Chlamydiae-Simkania negevensis GCF 000237205.1@WP 013943650.1
Chlamydiae-Chlamydiae bacterium GCA 001794905.1@OGN62995.1
Chlamydiae-Simkania negevensis GCF 000237205.1@WP 013942850.1
Chlamydiae-Chlamydia avium GCF 000583875.1@WP 020356125.1
Chlamydiae-Similichlamydia epinepheli GCF 003056015.1@WP 108623657.1
Chlamydiae-Parachlamydia sp. GCF 001545115.1@WP 068467402.1
Chlamydiae-Parachlamydiaceae bacterium GCF 000829755.1@WP 042102387.1
Chlamydiae-Neochlamydia sp. GCF 000648235.2@WP 042239794.1
Chlamydiae-Criblamydia sequanensis GCF 000750955.1@WP 041018366.1
Chlamydiae-Parachlamydia acanthamoebae GCF 000253035.1@WP 006340593.1
Chlamydiae-Estrella lausannensis GCF 900000175.1@WP 098038887.1
Chlamydiae-Waddlia chondrophila GCF 000092785.1@WP 041941596.1
Opisthokonta-Rozella allomycis 281847@EPZ36557.1
Chlamydiae-Chlamydiales bacterium GCF 900093645.1@WP 088222688.1
OtherPlastid-Thalassiosira pseudonana 35128@TP16G01770
OtherPlastid-Ectocarpus siliculosus 2880@ES0295G00020
OtherPlastid-Ectocarpus siliculosus 2880@ES0073G00230
OtherPlastid-Thalassiosira pseudonana 35128@TP12G01220
OtherPlastid-Pseudonitzschia multiseries 37319@240837
Chlamydiae-Protochlamydia naegleriophila GCF 001499655.1@WP 032124047.1
Chlamydiae-Chlamydiae bacterium GCA 001794905.1@OGN62512.1
Chlamydiae-Waddlia chondrophila GCF 000092785.1@WP 013181317.1
Chlamydiae-Estrella lausannensis GCF 900000175.1@WP 098038224.1
Chlamydiae-Criblamydia sequanensis GCF 000750955.1@WP 041017341.1
Rhodophyta-Rhodelphis marinus@m.10181
Rhodophyta-Rhodelphis limneticus@m.18550
Glaucocystophyceae-Cyanoptyche sp. 77921@MMETSP1086 DN5500 c0 g1 i1
Glaucocystophyceae-Cyanophora paradoxa GCA 004431415.1@tig00000147 g9461.t1
Glaucocystophyceae-Cyanoptyche sp. 77921@MMETSP1086 DN13156 c0 g1 i1
Glaucocystophyceae-Cyanoptyche sp. 77921@MMETSP1086 DN15806 c0 g1 i1
Glaucocystophyceae-Cyanophora paradoxa GCA 004431415.1@tig00020830 g14386.t1
Stramenopiles-Aplanochytrium kerguelense 87111@55434
Stramenopiles-Schizochytrium aggregatum 876976@43330
Stramenopiles-Aurantiochytrium limacinum 717989@71984
Viridiplantae-Chlorophyta-Chlorella variabilis GCA 000147415.1@EFN59352.1
Viridiplantae-Chlorophyta-Coccomyxa subellipsoidea 574566@CV08G00640
Viridiplantae-Chlorophyta-Volvox carteri GCA 000143455.1@EFJ49850.1
Viridiplantae-Chlorophyta-Chlorella variabilis GCA 000147415.1@EFN55123.1
Viridiplantae-Chlorophyta-Coccomyxa subellipsoidea 574566@CV01G04570
Viridiplantae-Streptophyta-Mesostigma viride GCA 004000685.1@BIFU01005501.1
Viridiplantae-Chlorophyta-Micromonas commoda GCA 000090985.2@ACO68065.1
Viridiplantae-Chlorophyta-Micromonas pusilla 564608@MP09G04910
Viridiplantae-Chlorophyta-Ostreococcus tauri GCA 000214015.2@CEG01693.1
Viridiplantae-Chlorophyta-Bathycoccus prasinos 41875@BPRRCC1105 02G02550
Glaucocystophyceae-Cyanophora paradoxa GCA 004431415.1@tig00020780 g13759.t1
Viridiplantae-Chlorophyta-Chlamydomonas reinhardtii 3055@CR13G04060
Viridiplantae-Chlorophyta-Chlorella variabilis GCA 000147415.1@EFN52101.1
Viridiplantae-Chlorophyta-Coccomyxa subellipsoidea 574566@CV21G01420
Viridiplantae-Streptophyta-Mesostigma viride GCA 004000685.1@BIFU01003194.1
Viridiplantae-Streptophyta-Klebsormidium flaccidum 3175@GAQ87641.1
Viridiplantae-Streptophyta-Arabidopsis thaliana 3702@AT1G15500
Viridiplantae-Streptophyta-Physcomitrella patens GCA 000002425.2@PNR54052.1
Viridiplantae-Streptophyta-Selaginella moellendorffii 88036@XP 002982220.1
Viridiplantae-Streptophyta-Selaginella moellendorffii 88036@XP 002974477.1
Glaucocystophyceae-Cyanoptyche sp. 77921@MMETSP1086 DN15159 c0 g1 i1
Glaucocystophyceae-Cyanophora paradoxa GCA 004431415.1@tig00000113 g5668.t1
Glaucocystophyceae-Cyanophora paradoxa GCA 004431415.1@tig00000523 g1867.t1
OtherPlastid-Aureococcus anophagefferens 44056@AA0011G02260
OtherPlastid-Fragilariopsis cylindrus 186039@FC009G08720
OtherPlastid-Vitrella brassicaformis 1169540@Vbra 2914
OtherPlastid-Ectocarpus siliculosus 2880@ES0180G00310
OtherPlastid-Ectocarpus siliculosus 2880@ES0067G00550
Rhodophyta-Porphyra umbilicalis GCA 002049455.2@OSX69864.1
Rhodophyta-Gracilariopsis chorda GCA 003194525.1@PXF48205.1
Rhodophyta-Porphyridium purpureum 35688@2042.6
OtherPlastid-Vitrella brassicaformis 1169540@Vbra 18493
Rhodophyta-Cyanidioschyzon merolae 45157@CM09G00400
Rhodophyta-Galdieria sulphuraria 130081@XP 005707402.1

Figure 24: Single gene phylogenetic tree of the ATP:ADP antiporter identified by the Chlamydia pipeline. (previous page) This tree corresponds to the phylogenetic reconstruction of the orthologous group OG0003449, annotated as ATP:ADP antiporter, selected by the Chlamydia pipeline. The tree was obtained by IQ-TREE, under an LG4X ultrafast-bootstrap model. Species identified as "other-plastid" are not considered as intruders, and are therefore an integral part of the clan selection criteria. Bootstrap values are indicated on the branches. The rooting is in mid-point. The scale counts in number of amino acid substitutions per site. In red: Rhodophyta and Rhodelphea, in green: Viridiplantae and in blue: Glaucophyta. In pink: Chlamydia, in yellow: Proteobacteria

e. Inventory of the literature and correlation with our results

Estimates of the number of chlamydial genes found in Archaeplastida vary between 30 and 100 depending on the studies and the stringency of the protocols used. By pooling publications from (Ball et al., 2013; Becker et al., 2008; Cenci et al., 2018, 2017; Huang and Gogarten, 2007; Moustafa et al., 2008) we have created a literature inventory that distributes across 125 of our initial orthologous groups. Our methods, on the other hand, identify 26 chlamydial genes in Archaeplastida during manual tree analysis and 57 through the automatic pipeline. In total, 73 different genes were selected in one way or another by our methods. Among them, 51 were also identified in these previous studies (Figure 9B). Depending on the different analyses conducted, some trees are more or less convincing to support the Ménage à Trois hypothesis. We ranked the totality of chlamydial genes recovered from Archaeplastida according to signal robustness, i.e., according to the number of methods for which they were recovered (appendix 7). Thus, 19 genes selected by the 2 pipelines implemented in this study, validated in manual analysis by 2 or 3 individuals were also found in the literature. These therefore constitute a robust core of genes with a common evolutionary history that may support the involvement of Chlamydia in primary plastid endosymbiosis.

f. Identification of specific LGT between Chlamydia and Glaucophytes

As said before, 51 genes from the literature inventory are also found by our methods (Figure 9B). For the remaining 74, a manual analysis of the corresponding trees was performed, allowing the validation of our methods, but also a conservative update of the MATH gene inventory. Thus, this analysis confirms the invalidation of a number of genes previously identified in the literature, indicating that their rejection by our pipeline was justified. Their initial selection was likely due to the unavailability of some genomic and proteomic data at the time of these previous studies, although their disqualification may also be due to the new analytical tools used here.

Manual analysis of these trees also reveals the existence of genes with a phylogenetic interaction between Chlamydia and Glaucocystophyceae only. These trees are indeed not recognized by our methods, since Glaucocystophyceae are absent from the initial dataset used to create the orthologous groups. Therefore, to specifically identify gene transfers between

these organisms, a new clustering was performed. From the sequences of the 33 Chlamydia and *Cyanophora paradoxa*, 27087 orthologous clusters were created by OrthoFinder (named OFCh). As some clusters are also present in OF57, the set of orthologous groups of the 57 eukaryotes used for previous pipelines, we removed them from the OFCh dataset. Only those with at least 2 Chlamydia and 1 *C. paradoxa* (795 clusters) were then enriched with cyanobacteria, then eukaryotes and the rest of the bacteria, in the same way as described above (Figure 7). At the output of the automatic pipeline, 60 trees with a phylogenetic relationship between at least 2 Chlamydia and 1 Glaucocystophyceae are identified, of which 7 are also reported in the literature. We had to change the selection criteria here since only 2 Glaucocystophyceae are present in the dataset. However, given the quality of the *Cyanophora paradoxa* data, as well as the lower proportion of glaucophytes available and integrated in our protocol, the choice of selection criteria should be optimized according to the identification of these particular gene transfers. Indeed, a quick and succinct manual analysis, performed here by one person and not three as previously, estimates a validation of barely 30% of the selected trees. The 60 identified trees therefore require further analysis to confirm the corresponding gene transfers.

Conclusions and Discussion

The different issues raised by this thesis project aimed at evaluating the chlamydial signal in Archaeplastida, not only from the point of view of the existence of a signal linked to the primary endosymbiosis of the plastid, but also from the point of view of the particular contribution of the pathogen, different from the other bacterial actors. The bioinformatics protocol thus allowed us to answer the different questions, and thus test the Ménage à Trois hypothesis. The manual analysis of the trees generated in the first place places the gene transfers identified in an endosymbiotic context, however, it is the complete automation of the pipeline that allows us to put the chlamydian signal into perspective within the global evolutionary history, both bacterial and eukaryotic. It is important to keep in mind that the goal of this project was not to establish an exhaustive list of MATH genes, but to test the phylogenetic predictions of this hypothesis. In doing so, it is likely, and even certain, that some endosymbiotic gene transfers may not have been identified by our methods, but also that others are actually false positives.

The different pipelines implemented allow to compare the signal from a donor point of view but also from an acceptor point of view of these gene transfers linked to primary plastid endosymbiosis. Concerning eukaryotic hosts, Chlamydia have played a particular role in the evolution of Archaeplastida compared to Fungi and Amoebozoa. Indeed, we count 1 LGT identified between Chlamydia and Fungii, and 16 between Chlamydia and Amoebozoa. For the latter, the window of time during which these eukaryotes were potentially in contact with Chlamydia can be considered larger than for Archaeplastida. Combined with the possible access of membranes to infection, this suggests a theoretically larger contribution of pathogens. However, the number of LGTs identified is significantly lower than those found in Archaeplastida, and the signal congruence is not as clear. In fact, since amoebae are supposed to be a positive witness of the chlamydial impact on eukaryotic evolution, we can infer that these pathogens had a particular role in Archaeplastida compared to other eukaryotes.

On the bacterial side, 97 LGTs are identified for the 5 main bacterial groups. The study confirms 656 cyanobacterial LGTs in the Archaeplastida genome, 57 chlamydial LGTs, 44 LGTs of Bacteroidetes, 39 LGTs for all Proteobacteria, 9 LGTs of Firmicutes and 5 LGTs of Actinobacteria. These results show the main contribution of Chlamydia in the diversity of bacterial LGTs. Although encouraging, they do not validate the MATH model since the

typology of transfers is similar in all bacteria, showing in all cases the existence of very ancient events contemporary to endosymbiosis and not allowing to distinguish Chlamydias from other bacterial groups on this level.

A more detailed analysis of these LGTs, both from a metabolic and functional point of view and from the point of view of the diversity of organisms affected by this transfer, however, suggests differences between bacterial donors. Indeed, as expected, no bacterial control group shows, despite the analysis of 97 distinct LGTs, the existence of multiple lateral transfers affecting the same metabolic pathway. On the other hand, Chlamydia show this characteristic for 4 different metabolisms; those of starch/glycogen (2 LGTs), isoprenoids (3/4 LGTs), menaquinone (2 LGTs) and tryptophan (3/4 LGTs). With the exception of the menaquinone synthesis pathway, which will be discussed later, each of the chlamydial genes impacting these metabolic pathways, and already reported in the literature, are also identified by our methods. Moreover, two of these pathways are associated with three of the seven plastid transporters found in chlamydial LGTs, whereas no plastid transporter is found in all 97 non-chlamydial bacterial LGTs. Although two transporters are identified for the proteobacterial pipeline, a thorough analysis of the corresponding trees redefines these transfers as chlamydial and not proteobacterial. These phylogenies indeed show a branching of only two Proteobacteria with the Archaeplastida (despite the enrichment with the 36 selected species) and the three Chlamydia then present in the dataset, which are thus considered as intruders but in sufficiently low proportion to validate the selection criteria of PhySortR. Logically, these two transporters identified with the proteobacterial pipeline are also identified with the chlamydial pipeline, showing then a more important diversity of pathogens connected with the Archaeplastida. According to the parameters retained in our study (calibrated on the basis of manual analysis), a tree is identified by the pipeline if 2 donors branch with 3 acceptors minimum. By adjusting these parameters to identify trees with at least 3 donors and 3 acceptors, we remove 6 chlamydian LGTs, 5 Bacteroidetes, 1 actinobacterial, 4 Amoebae, 3 Firmicutes, 11 proteobacterials and 16 cyanobacterials. While this does not induce a major change in the conclusions reached previously, it does remove the two proteobacterial transporters initially identified, thus supporting the absence of bacterial transporters other than cyanobacterial and chlamydian in Archaeplastida. The early connectivity of prokaryotic origin (the bulk being clearly of eukaryotic origin) of the plastid appears to be under the exclusive control of these two groups.

As said before, the bioinformatics pipeline was designed to test the Ménage à Trois hypothesis and not to build an exhaustive list of MATH genes. In fact, we have accepted some limitations in our protocol. One of them is the grouping into gene families of the different eukaryotic proteomes. Indeed, following the manual analysis of the trees selected by the semi-automatic pipeline and the inventory of the literature, it became obvious that some distinct orthologous groups should have been grouped together when others, largely multigenic, would have required a stricter cut. This is the case for the genes of the menaquinone synthesis pathway. In green plants, PHYLLO encodes four of the enzymes involved in this metabolic pathway. In bacteria, on the other hand, the corresponding genes are separated from each other. Since the grouping into gene families is based on eukaryotes, a single orthologous group is identified for a majority of the Men pathway, which will then be enriched with the bacterial sequences in the rest of the protocol. The multigenic character of the generated tree thus necessarily leads to a non-selection of the gene, although its chlamydial origin was demonstrated in Cenci et al.

Still concerning menaquinone, we have set up an additional analysis in order to identify a potential alternative synthesis pathway of this vitamin K in Gloeobacterales. After having determined the distribution of the menaquinone synthesis pathways, this analysis was based on the grouping of the 48 cyanobacteria into gene families, and then on the identification of the genes common to the only holders of this potential alternative pathway. However, the bioinformatics approach of this side project proved inconclusive, since very few candidate genes were identified, and would at least need to be rethought and deepened. From a functional point of view, it was planned to follow this bioinformatics analysis of the menaquinone synthesis pathway with the creation and characterization of a mutant library of *Gloeobacter violaceus*. However, due to the health crisis of 2020-2021, this project has been postponed.

All of the results presented above are based on the selection of organisms entering the protocol and on the grouping of eukaryotic proteomes into gene families. However, in this initial dataset, no glaucophyte was present. The reorientation of the pipeline on the identification of LGTs between Chlamydia and glaucophytes therefore makes up for this absence. However, the 60 LGTs identified should be taken with caution. Indeed, only two glaucophytes have data of good enough quality to integrate the pipeline, the selection criteria of the trees were thus adapted in this direction, inducing the selection of each tree having at

least one glaucophyte connected with at least two Chlamydia. After a very brief manual analysis of these trees, it appears that the majority would be reconsidered.

As a result of this thesis project, we can identify three criteria for which the chlamydial signal in Archaeplastida can be considered different from other bacterial signals: i) the slightly higher number of LGTs identified, ii) the multiple impact of these LGTs on metabolic pathways and iii) the presence of transporters. These three features, combined with the manual tree analysis that places LGTs in an endosymbiotic context, support the hypothesis of a Ménage à Trois Hypothesis during primary plastid endosymbiosis, which would be useful to compare to the total bacterial contribution.

The automation of the pipeline set up in the second part of the project reduces the difficulties of interpretation of the phylogenetic trees. It has been optimized, certainly, on the manual analysis of chlamydial trees, but especially to establish a comparison of the different bacterial signals. In fact, some parameters would have to be revised to be able to differentiate EGTs from ERGTs in particular, but also to differentiate the different possible scenarios. We have indeed imagined several phylogenetic scenarios according to the "depth" and "rarity" of the signal found in the archaeplastid diversity. It is important to remember that the aim is to describe different possible scenarios in order to test the Ménage à Trois hypothesis and not to list the MATH genes in an exhaustive way. Thus, two same topologies, according to the established criteria, can be categorized in the same scenario without being linked to the primary plastid endosymbiosis. Indeed, if one can easily associate the deep and dense scenario with a potential MATH gene (EGT or ERGT) and, conversely, associate the superficial and rare scenario with a late gene transfer (LGT), the interpretation of the other two intermediate scenarios (deep and rare or superficial and dense) is more open to question. The diversity of species impacted by gene transfer as well as phylogenetic tropisms will shed light on the possibility of multiple transfers on the one hand, versus a transfer in the common ancestor but lost in the descendants on the other. This ultimately reflects the difficulties of manual tree analysis, in which case the deep dense scenario would correspond to trees validated by observers, and the shallow rare scenario to trees rejected by the pipeline. Trees meeting the other two scenarios, meanwhile, would be categorized in manual analysis as uncertain.

In this Ph.D I focus on how cyanobiont could have become a plastid using a functional point of view, because I believe that the first requirement of such symbiosis should be metabolic. This will therefore create a link between organisms and the selective pressure

to keep a symbiotic relationship. I, thus, agree that a more complete view of the organellogenesis should incorporate a more genomic view of the cyanobiont integration. Indeed, the idea of genomic prerequisite, as defended in “the shopping bag model” (Howe, et al., 2008), could help define how links could have been created. In particular, this hypothesis is a good complement, which does not necessarily reject the MAT hypothesis. In addition, our view of lateral gene transfer (i.e. endosymbiotic gene transfer and endosymbiotic related gene transfer) that happened during the process of endosymbiosis are particularly in line with “the shopping bag model”. The point of divergence being that our view is mainly functional while their view is centered around the preadaptation. A second hypothesis that should be taken into account to understand lateral gene transfer is the ability to incorporate genes directly from the symbiont. Indeed it has been proposed that the need of more than one endosymbiont will facilitate their acquisition once endosymbiont dies and DNA is released. This hypothesis is called the “The limited transfer window hypothesis” (Barbrook et al., 2006) could be explored more deeply. However, this hypothesis is hardly testable here, and can only be discussed, theoretically. Indeed, if the MAT hypothesis is real, the number of chlamydia cells in its inclusion vesicle should be high and that nothing prevents, due to the non-coupling of cell division between host, chlamydia and the cyanobiont, the presence of several cyanobacteria inside the inclusion vesicle.

Futures Recherches et Perspectives

Following this project, I have been granted an additional 6 months to finalize the analysis of the chlamydial signal in Archaeplastida. These 6 months will allow us to analyze the total bacterial contribution in the evolution of Archaeplastida and to deepen the functional aspect of the identified LGT. Indeed, for the moment, only 4 bacterial signals have been studied, and although important differences in phylogenetic, topological and diversity profiles can already be observed, the analysis of the total bacterial contribution will put chlamydia in a more complete evolutionary context. Combined with a functional analysis of the identified LGTs, as well as an analysis of the diversity impacted by these gene transfers (both on the LGT donor and acceptor sides), the results of the project will put chlamydia back into the evolutionary context in a general way, and more specifically in the context of primary plastid endosymbiosis. From a theoretical and phylogenetic point of view, the further and more detailed conceptualization of EGTs and ERGTs, their topological and diversity profile, would allow to distinguish the different gene transfers and to rule on the chlamydial role. From a functional point of view, a significant proportion of the LGTs identified have not been identified in previous publications on the subject, and in fact suggest the chlamydial impact on new metabolic pathways.

However, this project was carried out by taking advantage of all the new resources available, both at the genomic and proteomic level, and at the methodological level. Thus, once finalized, and in the current state of resources, this study marks the end point of testing the Ménage à Trois Hypothesis in a bioinformatic way. Perhaps it will be possible to identify environmental evidence for the Ménage à Trois?

Materials and Methods

All of the computational work presented in this study was performed on an IBM/Lenovo Flex compute cluster running CentOS 6.6. The system consists of one large compute node (x440) and 11 smaller ones (x240), for a total of 228 physical cores, 3TB of RAM and 160TB of shared storage.

1. Selection of genomic and proteomic data

a. Eukaryotic and bacterial data

The analysis presented here is based on a selection of genomic and proteomic data optimized to maximize the representation of diversity while minimizing contamination. We took the option of using datasets already available in the laboratory, for which the quality of the data has been verified in previous projects, but also the use of automatic tools for the creation and evaluation of our selection, respectively TQMD (Léonard et al., 2021) and 42 (Van Vlierberghe et al., 2021). TQMD (ToRQuEMaDA: tool for retrieving queried Eubacteria, metadata and dereplicating assemblies) allows, from public databases in particular, to produce lists of dereplicated prokaryotic genomes representative of diversity. The bacterial datasets necessary for this study (detailed below) were thus created from TQMD, and are published in Léonard et al., 2021 (publication for which I am co-author). 42 as for it allows not only to assess the quality of genomic and proteomic data (whether by estimating contaminations or to a lesser extent by the completeness of the data), but also to enrich orthologous groups with sequences of additional organisms. This second function of the tool will be used in the second part of the protocol.

Thus, at the start of our study we selected 57 eukaryotic proteomes (http://hdl.handle.net/2268/254874; https://doi.org/10.6084/m9.figshare.13550267.v2), for which clustering into orthologous groups was already available in the lab (created by OrthoFinder (Emms and Kelly, 2015) under an inflation parameter of 1.5), as well as the various datasets generated by TQMD and published in Léonard et al., 2021. To this listing, we manually selected and added 15 Archaeplastida genomes, including 2 Glaucocystophyceae and 2 Rhodelphea (Gawryluk et al., 2019), 10 Amoebozoa, as well as 2 cyanobacteria considered basal *Gloeobacter violaceus* and *Pseudanabaena biceps*. All selected genomes and proteomes were evaluated by 42, whose configuration file is available in the appendix 5 (for more details see: (Cornet and Baurain, 2022; and Van Vlierberghe et al., 2021), to ensure data quality.

In total, we selected 72 eukaryotes (including 54 photosynthetic eukaryotes), 10 Amoebozoa, 33 Chlamydia, 48 Cyanobacteria, 37 Bacteroidetes, 36 Proteobacteria, 22 Firmicutes and 20 Actinobacteria. Depending on the orientation of the pipeline described below, and thus on the signal studied, a subset of this selection will enter the bioinformatics protocol (Figure 12). Therefore, to ensure a global bacterial representation, we combined two other taxa-neutral bacterial datasets also from TQMD: the first containing 49 species, published in Léonard et al., 2021, and the second of 92 species already available in the laboratory.

b. Distribution of the selected genomes and proteomes in the different analyses

The Ménage à Trois Hypothesis test has a part of identifying the chlamydial signal in Archaeplastida, but also a second part of comparing this signal with other groups. For each control, the pipeline is oriented on the screen of LGTs specific to the studied groups. This reorientation is implemented in particular by sampling the genomes and proteomes integrating the pipeline. Thus, each condition is assigned a partially specific dataset. The 72 eukaryotes selected previously, the 48 cyanobacteria and a majority of the two lists of 49+92 bacteria are common to all controls. However, the specific bacterial datasets vary from pipeline to pipeline. The LGT screen between Chlamydia and Archaeplastida requires the introduction of all 33 Chlamydia into the protocol, and thus the adaptation of the general bacterial list by removing this taxonomic group. The same applies to the other control groups: the datasets of the target organisms tested are added to the pipeline, while ensuring that they are removed from the general bacterial selection (of 49+92 species). All the selected genomes and proteomes, as well as their entry in the pipeline according to the studied signal, are listed in appendix 9.

2. Identification of the chlamydial signal in Archaeplastida

a. Semi-automatic pipeline: general methodological approach

The semi-automated pipeline described here is divided into three steps (Figure 7): 1) partitioning and selection of the 57 eukaryotic proteomes into orthologous groups; 2) enrichment and filtration of the selected orthologous groups; and 3) selection of LGTs after phylogenetic reconstruction. After separation of the genomic and proteomic data into gene families or orthologous groups (OG or clusters), the chlamydial signal is directly put in competition with the cyanobacterial signal. Each selected cluster is then aligned with MAFFT (Katoh and Standley, 2013) and enriched with the previously selected eukaryotic and bacterial data Phylogenetic reconstruction is then used to verify whether the chlamydial signal still holds. Finally, a manual analysis of the trees selected by the pipeline is used to confirm or deny gene transfer and then, if transfer is validated, to determine if the context is endosymbiotic. The entire pipeline is shown in Figure 7.

b. Sorting of orthologous groups

From the 57 eukaryotes, a clustering into gene families was performed by OrthoFinder (inflation parameter set to 1.5, (Emms and Kelly, 2015)) (called OF57). From the orthologous groups (OG) containing less than 3000 sequences, classify-mcl.pl (https://metacpan.org/dist/Bio-MUST-Tools-Mcl) retained only those containing at least 2 Viridiplantae and/or Rhodophyta. A second step of orthologous group reduction occurs then, after enrichment with chlamydian sequences. Classify-ali.pl then retains only the OG with at least one chlamydia.

c. Enrichments and filtration of orthologous groups

Each orthologous group selected in the previous step is enriched with the proteomes of 33 Chlamydia and 48 Cyanobacteria. This step is performed by 42, whose configuration file is available in the Appendix 6 (for details see: (Cornet and Baurain, 2022; and Van Vlierberghe et al., 2021). Classify-ali.pl (https://metacpan.org/dist/Bio-MUST-Core) then allowed us to filter a second time for orthologous groups on the presence of at least one Chlamydia. These OGs selected by classify-ali.pl were then cleaned by prune-outliers.pl (https://metacpan.org/dist/Bio-MUST-Apps-FortyTwo, -min-threshold=0, max-threshold=0.9, evalue=1e-02) and HmmCleaner (Di Franco et al., 2019) and aligned with MAFFT v7.453 (Katoh and Standley, 2013). ali2phylip.pl (min=0.3, max=05) then reduced the proportion of missing sites in the alignments. Phylogenetic reconstruction of the trees corresponding to each OG was performed by RAxML (Stamatakis, 2014), under a PROTGAMMALG4X model, and in ultra-fast bootstrap. The automatic analysis of the generated trees, performed by clans-label.pl, then locates trees with a phylogenetic interaction between at least one Chlamydia and one Archaeplastida. This tool analyzes each bipartition of each tree and allows the identification of clans composed of at least one Chlamydia and one Archaeplastida, without interruption by other species.

A second phase of enrichment then occurs with 42, on the tree alignments of interest, with the previously selected proteomes and genomes not yet included in the alignments. At this point, all orthologous groups have received, or had the opportunity to receive, homologous sequences from the 286 organisms selected for the first stage of this study. As described previously, we eliminated sequences with low similarity and reduced the proportion of missing sites with prune-outliers.pl (treshold selected at 0.2, evalue=1e-02) and ali2phylip.pl (min=0.3, max=0.5, mask=BMGE) (https://metacpan.org/dist/Bio-MUST-Core).

Several tools and parameters were tested in the development of the pipeline. Regarding alignment filtering, we compared the results obtained with HmmCleaner combined with ali2phylip.pl (min=0.3, max=0.5), ali2phylip.pl applying block shrinkage by BMGE (Criscuolo and Gribaldo, 2010) (min=0.3, max0.5, bmge-mask=loose), and ali2phylip.pl reducing the proportions of missing sites only (min=0.3, max=0.5). Similarly, phylogenetic

tree reconstruction was performed by RAxML (PROTGAMMALG4X, ultrafast-bootstrap) compared to IQ-TREE (Nguyen et al., 2015)(LG4X, ultrafast-bootstrap).

d. Phylogenetic reconstruction and LGT selection

After the successive steps of enrichment and filtration of orthologous groups, phylogenetic reconstruction was performed by RAxML, with the PROTGAMMALG4X model and ultrafast bootstrap (Stamatakis, 2014). clans-label.pl then identified trees with a phylogenetic interaction between at least one Chlamydia and at least one Archaeplastida.

e. Manual tree analysis

Selected trees were automatically colored by format-tree.pl (found in Bio::MUST::Core) and visualized by iTOL (https://itol.embl.de) (Letunic and Bork, 2019). Manual analysis of the trees was then performed by three independent individuals. Several evaluation criteria were initially defined to allow classification of the trees into three categories: 1) in favor of MATH, i.e., showing a phylogenetic interaction that indicates an endosymbiotic context; 2) not in favor of MATH; and 3) uncertain, impossible to conclude. These criteria include i) the quality and diversity of the donor (Chlamydia) in the subtrees of interest, i.e., the abundance of organisms in the clan, but also the presence of multiple sequence ii) the quality and diversity of the acceptors (Archaeplastida and other photosynthetic organisms) in the subtrees of interest, iii) the presence of intruders in the subtrees of interest and finally iv) the general topology of the tree, in order to identify potential paralogues or multigene families, but also to evaluate the total diversity of donors and acceptors.

f. Inventory of chlamydial genes in the literature

To create an inventory of Chlamydia genes in Archaeplastida and compare it to our own data, we retrieved sequences published in Ball et al. 2013; Becker et al. 2008; Cenci et al. 2018, 2017; Huang and Gogarten 2007; Moustafa et al. 2008. We dereplicated these sequences with a 95% cd-hit (Li and Godzik, 2006) and identified orthologous groups in our pipeline corresponding to this inventory using BLAST (Altschul et al., 1997).

3. Specificity of the chlamydial signal in Archaeplastida

a. Pipeline automation

While manual analysis is necessary to evaluate and place the gene transfer identified by the semi-automated pipeline in a MATH context, comparison of the chlamydial signal in Archaeplastida to the signal in other bacteria requires full automation of the protocol. Based on the manual analysis, we calibrated this automatic pipeline so that it best mimics the manual application of the previously mentioned tree evaluation criteria (Figure 7).

After an initial filtering of orthologous clusters on the presence of at least 2 Viridiplantae and/or Rhodophyta, all OGs are enriched in Chlamydia and cyanobacteria with 42. The clusters selected by classify-ali.pl were then cleaned and aligned by cleanOG.pl (-min-threshold=0, max-threshold=0.9, evalue=1e-02) and MAFFT v7.453 (Katoh and Standley, 2013). We used 42 to enrich all orthologous clusters with the rest of the eukaryotic and bacterial genomes and proteomes selected for this pipeline. After a second filtration of the clusters by cleanOG.pl and ali2phylip.pl (min=0.3, max=0.5, mask-bmge=loose), IQ-TREE (Nguyen et al., 2015, LG4X, ultrafast-bootstrap) handled the phylogenetic reconstruction of the trees. We then used Treeshrink (Mai and Mirarab, 2018) to remove long branches and Treemer (Menardo et al., 2018) to phylogenetically dereplicate the trees by retaining at least one sequence per species. Each sequence identified by Treeshrink and Treemer was then removed from the alignment files and IQ-TREE performed the final phylogenetic reconstruction with an LG4X model and in ultrafast-bootstrap. PhySortR (Stephens et al., 2016) then automatically selected each tree with a clan of interest, allowing 10% intruders and with a minimum of 30% of all target species in the tree being in the subtree of interest. Finally, classif-ali.pl identified clans with a minimum diversity of 2 donors and 3 acceptors https://metacpan.org/dist/Bio-MUST-Core).

b. Chlamydial signal controls in Archaeplastida

To assess the chlamydial signal in Archaeplastida, 2 types of controls are proposed: first, the control of the chlamydial signal in Archaeplastida relative to other bacterial groups and second, the control of the chlamydial signal specifically in Archaeplastida relative to other eukaryotic groups, called the "donor" and "acceptor" controls respectively.

For the donor side of this control, we compared the chlamydial signal in Archaeplastida with Proteobacteria, Bacteroidetes, Firmicutes and Actinobacteria. In the same spirit, we also quantified the cyanobacterial signal in Archaeplastida, in order to evaluate the sensitivity of our methods. For each case, we redirected the automatic pipeline to identify gene transfers between each bacterial group and Archaeplastida. The protocol remains exactly the same as previously described, but the combination of genomes and proteomes entering the pipeline differs for each condition (appendix 9). Thus, instead of adding the 33 Chlamydia proteomes and filtering the clusters on the presence of these organisms, we enriched the orthologous groups with the dataset corresponding to the control pipeline studied and filtered the clusters on the presence of these target species. We also adjusted the bacterial dataset by removing the target species from the combination of the two lists of 49 and 92 species.

For the acceptor side of the analysis, we compared the chlamydial signal in Archaeplastida to the chlamydial signal in amoebae and fungi. The pipeline was also redirected to the identification of corresponding gene transfers by modifying the clustering filter. After clustering by OrthoFinder (Emms and Kelly, 2015), we filtered orthologous groups on the

presence of at least 2 Fungi or 2 Amoebozoa. For the amoebozoa pipeline, this step was preceded by enriching all clusters with amoebozoa genomes and proteomes. The rest of the protocol remains exactly the same as described for the chlamydial signal in Archaeplastida.

c. Concatenation and signal congruence

For each pipeline redirection condition, we assessed the congruence of the identified signal. Using the sequences of the target species present in the subtrees of interest, the set of gene transfers identified by the different pipelines underwent two types of analysis: a super-matrix concatenation analysis of signal congruence, and a supertree concatenation analysis.

After filtration and sequence alignments by ali2phylip.pl implemented of the BMGE filter (min=0.3, max=0.5, bmge-mask=loose, https://metacpan.org/dist/Bio-MUST-Core, Criscuolo and Gribaldo, 2010,) and MAFFT v7.453 (Katoh and Standley, 2013), the supermatrices were created by SCaFoS v1.30k (Roure et al., 2007), using the minimum evolutionary distance as a selection criterion among paralogous sequences of the same OTU (25% complete elimination threshold), the maximum percentage of missing sites for a complete sequence set to 10, and the maximum number of missing OTUs set to 25 (except for Firmicutes, set to 22, and for Actinobacteria, set to 20). In addition, species with sequence frequencies below 10% were removed from the alignment. IQ-TREE then allowed phylogenetic reconstruction of the supermatrices with LG4X, C20, and C60 models and in ultrafast-bootstrap (Nguyen et al., 2015).

Using split-matrix.pl (https://metacpan.org/dist/Bio-MUST-Core) to redivide the supermatrices into individual genes, we resampled each set of orthologous groups to a size corresponding to one-third of the supermatrices. For each of the 100 replicates created by jack-ali-dir.pl (https://metacpan.org/dist/Bio-MUST-Core) for each pipeline (except for cyanobacteria for which we made 2 sets of replicates, one corresponding to a third of the original supermatrix and the other with a length set at 4500 AA), SCaFoS v1.30k recreated the supermatrices using the same parameters as described above and IQ-TREE enabled phylogenetic reconstruction of the trees, with an LG4X model and in ultrafast-bootstrap.

In parallel, we also performed a signal congruence analysis based on supertree generation. ASTRAL-III (v5.7.5) (Mirarab et al., 2014) produced a supertree for each pipeline from single-gene phylogenetic trees generated by IQ-TREE (LG4X, ultra-fast bootstrap) picking up only the target species of interest for each condition.

d. Rooting of trees from concatenations

The trees resulting from the concatenations are only composed of target species, and do not have an outgroup. The rooting of these trees is therefore manual, on the most basal

donor species present. We therefore first reconstructed the species phylogeny for each bacterial dataset.

For each TQMD selection (Léonard et al., 2021), we retrieved the corresponding proteomes and used 42 to recover their ribosomal proteins (Cornet and Baurain, 2022; Van Vlierberghe et al., 2021). The taxonomy of these proteins were labeled by calculating the last common ancestor (best hit BLAST) in the corresponding alignments (excluding self-matches), provided they had a bit-score ≥80 and were within 99% of the bit-score of the first hit (MEGAN-like algorithm (Cornet et al., 2018)). The supermatrices corresponding to each TQMD selection were assembled from the ribosomal proteins. Briefly, the sequences were aligned with MAFFT v7.453 (Katoh and Standley, 2013), and then the alignments were filtered using ali2phylip.pl (https://metacpan.org/dist/Bio-MUST-Core), implemented with the BMGE filter (Criscuolo & Gribaldo, 2010) (min=0.3, max=0.5, bmge-mask=loose). This step reduced the proportion of missing sites in the alignments. Next, we used Scafos v1.30k (Roure et al., 2007) to create the six different supermatrices, using the minimum evolutionary distance as a sequence selection criterion (threshold set to 25%), the maximum percentage of missing sites for a "complete sequence" set to 10 and the maximum number of missing OTUs set to 25, except for Firmicutes (22) and Actinobacteria (20). Finally, IQ-TREE (Hoang et al., 2018; Nguyen et al., 2015) was used to reconstruct the phylogenomic tree associated with each supermatrix, using the LG4X model with ultra-fast bootstraps. Trees were automatically annotated and colored using format-tree.pl (also from Bio::MUST::Core), and then visualized with iTOL v4 (Letunic and Bork, 2019).

e. Comparisons of the different selections

Comparing the different selections to each other allows us to eventually identify divergences within alignments, or trees, and thus to refine the role of Chlamydia during primary plastid endosymbiosis. For each selected pipeline tree, we inventoried several parameters: the number of sequences and total species in the clans, then separated into donors - acceptors and intruders and the topology of the clans (i.e. the number of Archaeplastida lineages present). These data were then visualized with different R packages. The distribution of each parameter within the different selections was done with the Rainclouds plot package (Allen et al., 2021). Significance of differences between the different pipelines was evaluated by a Kruskal-Wallis test (specific parameters for each condition are listed on the corresponding figures) (Figures 7 and 18).

Data visualization generally required ggplot2 (Wickham, 2009), for the creation of diagrams and graphs. The intersection of the different chlamydial selections, on the other hand, was evaluated by the UpSetR package (Conway et al., 2017).

f. Functional annotations

For each selected clan in each pipeline, we retrieved the sequences of the target species. These different sequence lists were then annotated by EggNog mapper (Huerta-Cepas et al., 2017) and BlastKoala (Kanehisa et al., 2016). We assigned an annotation to the set of clans selected by the pipeline when at least 50% of the sequences were annotated in the same way.

We also inferred the localization of each protein identified by the pipeline. Based solely on the Archaeplastida sequences in each selected cluster, and the localization annotations present in (Van Vlierberghe et al., 2021), annotate.pl (https://doi.org/10.6084/m9.figshare.18544955.v2) retrieved, for each of our sequences, the corresponding annotation. Again, we assigned a localization to the selected clan if at least 50% of the sequences had the same.

4. Identification of specific LGT between Chlamydia and Glaucophyta

To address the lack of glaucophyte data in the initial clustering performed from the 57 eukaryotes, we refocused the pipeline on specifically identifying gene transfers between Chlamydia and these organisms. To do so, from the 33 selected Chlamydia proteomes, to which we added the available *Cyanophora paradoxa* transcriptome (Price et al., 2019), a second clustering into gene families was performed by OrthoFinder (inflation parameter = 1.5, (Emms and Kelly, 2015)). After an initial selection of orthologous clusters on the presence of at least 1 Chlamydia and 1 Glaucystophyceae, we removed from this dataset all OGs already identified in the main pipeline. The rest of the protocol remains the same as previously described. The clusters selected by classify-mcl.pl are then enriched with the cyanobacteria and the dataset of 57 eukaryotes before being cleaned up by prune-outlier.pl (threshold selected at 0.2 and evalue = 1e-05, https://metacpan.org/dist/Bio-MUST-Apps-FortyTwo) and aligned with MAFFT v7.453 (Katoh and Standley, 2013). The proteomes of the additional 15 Archaeplastida (except *Cyanophora paradoxa*, which is already part of the clustering), and the proteomes of the 49+92 bacterial species selected by TQMD (Leonard et al., 2021) (from which we removed Cyanobacteria and Chlamydia) are then added to each alignment by 42. The sequences are then cleaned a second time by prune-tree.pl and ali2phylipp.pl implemented with the BMGE filter (min=0.3, max=0.5, bmge-mask=loose, (Criscuolo and Gribaldo, 2010)), and then IQ-TREE (Nguyen et al., 2015) performs phylogenetic reconstruction of each orthologous group with an LG4X model and an ultrafast-bootstrap. Treeshrink (Mai and Mirarab, 2018)

and Treemer (Menardo et al., 2018) then perform long branch removal and sequence dereplication, taking care to leave at least one sequence per species present in the tree. PhySortR (Stephens et al., 2016) finally takes care of analyzing the trees generated by IQ-TREE (LG4X, ultrafast-bootstrap) and selecting those with a phylogenetic association of at least 1 Chlamydia with at least one Archaeplastida (Glaucophyta), allowing 10% intruders as well as the minimum presence of 30% of the total target species in the identified clan. Classify-ali.pl finally ends by distinguishing clans for which at least 3 Chlamydia branch with at least 1 Glaucophyta.

Bibliography

AbdelRahman, Y.M., Belland, R.J., 2005. The chlamydial developmental cycle. FEMS Microbiol. Rev. 29, 949–959. https://doi.org/10.1016/j.femsre.2005.03.002

Adl, S.M., Simpson, A.G.B., Farmer, M.A., Andersen, R.A., Anderson, O.R., Barta, J.R., Bowser, S.S., Brugerolle, G., Fensome, R.A., Fredericq, S., James, T.Y., Karpov, S., Kugrens, P., Krug, J., Lane, C.E., Lewis, L.A., Lodge, J., Lynn, D.H., Mann, D.G., McCourt, R.M., Mendoza, L., Moestrup, O., Mozley-Standridge, S.E., Nerad, T.A., Shearer, C.A., Smirnov, A.V., Spiegel, F.W., Taylor, M.F.J.R., 2005. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J. Eukaryot. Microbiol. 52, 399–451. https://doi.org/10.1111/j.1550-7408.2005.00053.x

Allen, M., Poggiali, D., Whitaker, K., Marshall, T.R., Langen, J. van, Kievit, R.A., 2021. Raincloud plots: a multi-platform tool for robust data visualization. https://doi.org/10.12688/wellcomeopenres.15191.2

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Archibald, J.M., 2015. Endosymbiosis and Eukaryotic Cell Evolution. Curr. Biol. CB 25, R911-921. https://doi.org/10.1016/j.cub.2015.07.055

Archibald, J.M., 2009. The puzzle of plastid evolution. Curr. Biol. CB 19, R81-88. https://doi.org/10.1016/j.cub.2008.11.067

Ball, S.G., Colleoni, C., Kadouche, D., Ducatez, M., Arias, M.-C., Tirtiaux, C., 2015. Toward an understanding of the function of Chlamydiales in plastid endosymbiosis. Biochim. Biophys. Acta 1847, 495–504. https://doi.org/10.1016/j.bbabio.2015.02.007

Ball, S.G., Subtil, A., Bhattacharya, D., Moustafa, A., Weber, A.P.M., Gehre, L., Colleoni, C., Arias, M.-C., Cenci, U., Dauvillée, D., 2013. Metabolic Effectors Secreted by Bacterial Pathogens: Essential Facilitators of Plastid Endosymbiosis?[W][OA]. Plant Cell 25, 7–21. https://doi.org/10.1105/tpc.112.101329

Baum, D., 2013. The Origin of Primary Plastids: A Pas de Deux or a Ménage à Trois? Plant Cell 25, 4–6. https://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 Evol. Biol. 8, 203. https://doi.org/10.1186/1471-2148-8-203

Bodył, A., 2018. Did some red alga-derived plastids evolve via kleptoplastidy? A hypothesis. Biol. Rev. Camb. Philos. Soc. 93, 201–222. https://doi.org/10.1111/brv.12340

Cavalier-Smith, T., 2003. Genomic reduction and evolution of novel genetic membranes and protein-targeting machinery in eukaryote-eukaryote chimaeras (meta-algae). Philos. Trans. R. Soc. Lond. B. Biol. Sci. 358, 109–133; discussion 133-134. https://doi.org/10.1098/rstb.2002.1194

Cavalier-Smith, T., 1998. A revised six-kingdom system of life. Biol. Rev. Camb. Philos. Soc. 73, 203–266. https://doi.org/10.1017/s0006323198005167

Cenci, U., Bhattacharya, D., Weber, A.P.M., Colleoni, C., Subtil, A., Ball, S.G., 2017. Biotic Host-Pathogen Interactions As Major Drivers of Plastid Endosymbiosis. Trends Plant Sci. 22, 316–328. https://doi.org/10.1016/j.tplants.2016.12.007

Cenci, U., Ducatez, M., Kadouche, D., Colleoni, C., Ball, S.G., 2016. Was the Chlamydial Adaptative Strategy to Tryptophan Starvation an Early Determinant of Plastid Endosymbiosis? Front. Cell. Infect. Microbiol. 6, 67. https://doi.org/10.3389/fcimb.2016.00067

Cenci, U., Moog, D., Archibald, J.M., 2015. Origin and Spread of Plastids by Endosymbiosis, in: Algal and Cyanobacteria Symbioses. WORLD SCIENTIFIC (EUROPE), pp. 43–81. https://doi.org/10.1142/9781786340580_0002

Cenci, U., Qiu, H., Pillonel, T., Cardol, P., Remacle, C., Colleoni, C., Kadouche, D., Chabi, M., Greub, G., Bhattacharya, D., Ball, S.G., 2018. Host-pathogen biotic interactions shaped vitamin K metabolism in Archaeplastida. Sci. Rep. 8. https://doi.org/10.1038/s41598-018-33663-w

Chan, C.X., Yang, E.C., Banerjee, T., Yoon, H.S., Martone, P.T., Estevez, J.M., Bhattacharya, D., 2011. Red and green algal monophyly and extensive gene sharing found in a rich repertoire of red algal genes. Curr. Biol. CB 21, 328–333. https://doi.org/10.1016/j.cub.2011.01.037

Colleoni, C., Linka, M., Deschamps, P., Handford, M.G., Dupree, P., Weber, A.P.M., Ball, S.G., 2010. Phylogenetic and biochemical evidence supports the recruitment of an ADP-glucose translocator for the export of photosynthate during plastid endosymbiosis. Mol. Biol. Evol. 27, 2691–2701. https://doi.org/10.1093/molbev/msq158

Collingro, A., Tischler, P., Weinmaier, T., Penz, T., Heinz, E., Brunham, R.C., Read, T.D., Bavoil, P.M., Sachse, K., Kahane, S., Friedman, M.G., Rattei, T., Myers, G.S.A., Horn, M., 2011. Unity in Variety—The Pan-Genome of the Chlamydiae. Mol. Biol. Evol. 28, 3253–3270. https://doi.org/10.1093/molbev/msr161

Collins, M.D., Jones, D., 1981. Distribution of isoprenoid quinone structural types in bacteria and their taxonomic implication. Microbiol. Rev. 45, 316–354.

Conway, J.R., Lex, A., Gehlenborg, N., 2017. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinforma. Oxf. Engl. 33, 2938–2940. https://doi.org/10.1093/bioinformatics/btx364

Cornet, L., Baurain, D., 2022. Contamination detection in genomic data: more is not enough. Genome Biol. 23, 60. https://doi.org/10.1186/s13059-022-02619-9

Cornet, L., Bertrand, A.R., Hanikenne, M., Javaux, E.J., Wilmotte, A., Baurain, D., 2018. Metagenomic assembly of new (sub)polar Cyanobacteria and their associated microbiome from non-axenic cultures. Microb. Genomics 4. https://doi.org/10.1099/mgen.0.000212

Criscuolo, A., Gribaldo, S., 2011. Large-scale phylogenomic analyses indicate a deep origin of primary plastids within cyanobacteria. Mol. Biol. Evol. 28, 3019–3032. https://doi.org/10.1093/molbev/msr108

Criscuolo, A., Gribaldo, S., 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol. Biol. 10, 210. https://doi.org/10.1186/1471-2148-10-210

Dagan, T., Roettger, M., Stucken, K., Landan, G., Koch, R., Major, P., Gould, S.B., Goremykin, V.V., Rippka, R., Tandeau de Marsac, N., Gugger, M., Lockhart, P.J., Allen, J.F., Brune, I., Maus, I., Pühler, A., Martin, W.F., 2013. Genomes of Stigonematalean Cyanobacteria (Subsection V) and the Evolution of Oxygenic Photosynthesis from Prokaryotes to Plastids. Genome Biol. Evol. 5, 31–44. https://doi.org/10.1093/gbe/evs117

Deschamps, P., 2014. Primary endosymbiosis: have cyanobacteria and Chlamydiae ever been roommates? Acta Soc. Bot. Pol. 83. https://doi.org/10.5586/asbp.2014.048

Deschamps, P., Colleoni, C., Nakamura, Y., Suzuki, E., Putaux, J.-L., Buléon, A., Haebel, S., Ritte, G., Steup, M., Falcón, L.I., Moreira, D., Löffelhardt, W., Raj, J.N., Plancke, C., d'Hulst, C., Dauvillée, D., Ball, S., 2008. Metabolic symbiosis and the birth of the plant kingdom. Mol. Biol. Evol. 25, 536–548. https://doi.org/10.1093/molbev/msm280

Di Franco, A., Poujol, R., Baurain, D., Philippe, H., 2019. Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. BMC Evol. Biol. 19, 21. https://doi.org/10.1186/s12862-019-1350-2

Domman, D., Horn, M., Embley, T.M., Williams, T.A., 2015. Plastid establishment did not require a chlamydial partner. Nat. Commun. 6, 6421. https://doi.org/10.1038/ncomms7421

Emms, D.M., Kelly, S., 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16, 157. https://doi.org/10.1186/s13059-015-0721-2

Eugeni Piller, L., Besagni, C., Ksas, B., Rumeau, D., Bréhélin, C., Glauser, G., Kessler, F., Havaux, M., 2011. Chloroplast lipid droplet type II NAD(P)H quinone oxidoreductase is essential for prenylquinone metabolism and vitamin K1 accumulation. Proc. Natl. Acad. Sci. U. S. A. 108, 14354–14359. https://doi.org/10.1073/pnas.1104790108

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 Plant Sci. 18, 673–679. https://doi.org/10.1016/j.tplants.2013.09.006

Gabr, A., Grossman, A.R., Bhattacharya, D., 2020. Paulinella, a model for understanding plastid primary endosymbiosis. J. Phycol. 56, 837–843. https://doi.org/10.1111/jpy.13003

Gawryluk, R.M.R., Tikhonenkov, D.V., Hehenberger, E., Husnik, F., Mylnikov, A.P., Keeling, P.J., 2019. Non-photosynthetic predators are sister to red algae. Nature. https://doi.org/10.1038/s41586-019-1398-6

Gehre, L., Gorgette, O., Perrinet, S., Prevost, M.-C., Ducatez, M., Giebel, A.M., Nelson, D.E., Ball, S.G., Subtil, A., n.d. Sequestration of host metabolism by an intracellular pathogen. eLife 5. https://doi.org/10.7554/eLife.12552

Gil, R., Latorre, A., 2019. Unity Makes Strength: A Review on Mutualistic Symbiosis in Representative Insect Clades. Life Basel Switz. 9, E21. https://doi.org/10.3390/life9010021

Green, B.R., 2011. After the primary endosymbiosis: an update on the chromalveolate hypothesis and the origins of algae with Chl c. Photosynth. Res. 107, 103–115. https://doi.org/10.1007/s11120-010-9584-2

Gross, J., Meurer, J., Bhattacharya, D., 2008. Evidence of a chimeric genome in the cyanobacterial ancestor of plastids. BMC Evol. Biol. 8, 117. https://doi.org/10.1186/1471-2148-8-117

Hackett, J.D., Yoon, H.S., Li, S., Reyes-Prieto, A., Rümmele, S.E., Bhattacharya, D., 2007. Phylogenomic analysis supports the monophyly of cryptophytes and haptophytes and the association of rhizaria with chromalveolates. Mol. Biol. Evol. 24, 1702–1713. https://doi.org/10.1093/molbev/msm089

Henrissat, B., Deleury, E., Coutinho, P.M., 2002. Glycogen metabolism loss: a common marker of parasitic behaviour in bacteria? Trends Genet. TIG 18, 437–440. https://doi.org/10.1016/s0168-9525(02)02734-8

Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., Vinh, L.S., 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol. Biol. Evol. 35, 518–522. https://doi.org/10.1093/molbev/msx281

Huang, J., Gogarten, J.P., 2007. Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? Genome Biol. 8, R99. https://doi.org/10.1186/gb-2007-8-6-r99

Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., von Mering, C., Bork, P., 2017. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. Mol. Biol. Evol. 34, 2115–2122. https://doi.org/10.1093/molbev/msx148

Ikeda, Y., Komura, M., Watanabe, M., Minami, C., Koike, H., Itoh, S., Kashino, Y., Satoh, K., 2008. Photosystem I complexes associated with fucoxanthin-chlorophyll-binding proteins from a marine centric diatom, Chaetoceros gracilis. Biochim. Biophys. Acta 1777, 351–361. https://doi.org/10.1016/j.bbabio.2008.01.011

Irisarri, I., Strassert, J.F.H., Burki, F., 2022. Phylogenomic Insights into the Origin of Primary Plastids. Syst. Biol. 71, 105–120. https://doi.org/10.1093/sysbio/syab036

Jarvis, P., Soll, J., 2002. Toc, tic, and chloroplast protein import. Biochim. Biophys. Acta 1590, 177–189. https://doi.org/10.1016/s0167-4889(02)00176-3

Joyard, J., Ferro, M., Masselon, C., Seigneurin-Berny, D., Salvi, D., Garin, J., Rolland, N., 2010.

Chloroplast proteomics highlights the subcellular compartmentation of lipid metabolism. Prog. Lipid Res. 49, 128–158. https://doi.org/10.1016/j.plipres.2009.10.003

Kanehisa, M., Sato, Y., Morishima, K., 2016. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. J. Mol. Biol. 428, 726–731. https://doi.org/10.1016/j.jmb.2015.11.006

Katoh, K., Standley, D.M., 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. 30, 772–780. https://doi.org/10.1093/molbev/mst010

Keeling, P.J., 2010. The endosymbiotic origin, diversification and fate of plastids. Philos. Trans. R. Soc. B Biol. Sci. 365, 729–748. https://doi.org/10.1098/rstb.2009.0103

Kelly, S., 2021. The economics of organellar gene loss and endosymbiotic gene transfer. Genome Biol. 22, 345. https://doi.org/10.1186/s13059-021-02567-w

Kim, M., Chen, Y., Xi, J., Waters, C., Chen, R., Wang, D., 2015. An antimicrobial peptide essential for bacterial survival in the nitrogen-fixing symbiosis. Proc. Natl. Acad. Sci. U. S. A. 112, 15238–15243. https://doi.org/10.1073/pnas.1500123112

Lefebvre-Legendre, L., Rappaport, F., Finazzi, G., Ceol, M., Grivet, C., Hopfgartner, G., Rochaix, J.-D., 2007. Loss of phylloquinone in Chlamydomonas affects plastoquinone pool size and photosystem II synthesis. J. Biol. Chem. 282, 13250–13263. https://doi.org/10.1074/jbc.M610249200

Léonard, R.R., Leleu, M., Van Vlierberghe, M., Cornet, L., Kerff, F., Baurain, D., 2021. ToRQuEMaDA: tool for retrieving queried Eubacteria, metadata and dereplicating assemblies. PeerJ 9, e11348. https://doi.org/10.7717/peerj.11348

Letunic, I., Bork, P., 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 47, W256–W259. https://doi.org/10.1093/nar/gkz239

Li, W., 2016. Bringing Bioactive Compounds into Membranes: The UbiA Superfamily of Intramembrane Aromatic Prenyltransferases. Trends Biochem. Sci. 41, 356–370. https://doi.org/10.1016/j.tibs.2016.01.007

Li, W., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinforma. Oxf. Engl. 22, 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Loddenkötter, B., Kammerer, B., Fischer, K., Flügge, U.I., 1993. Expression of the functional mature chloroplast triose phosphate translocator in yeast internal membranes and purification of the histidine-tagged protein by a single metal-affinity chromatography step. Proc. Natl. Acad. Sci. U. S. A. 90, 2155–2159. https://doi.org/10.1073/pnas.90.6.2155

Lu, C., Lei, L., Peng, B., Tang, L., Ding, H., Gong, S., Li, Z., Wu, Y., Zhong, G., 2013. Chlamydia trachomatis GlgA Is Secreted into Host Cell Cytoplasm. PLOS ONE 8, e68764. https://doi.org/10.1371/journal.pone.0068764

Mai, U., Mirarab, S., 2018. TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. BMC Genomics 19, 272. https://doi.org/10.1186/s12864-018-4620-2

Marin, B., Nowack, E.C.M., Melkonian, M., 2005. A plastid in the making: evidence for a second primary endosymbiosis. Protist 156, 425–432. https://doi.org/10.1016/j.protis.2005.09.001

Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., Leister, D., Stoebe, B., Hasegawa, M., Penny, D., 2002. Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc. Natl. Acad. Sci. 99, 12246–12251. https://doi.org/10.1073/pnas.182432999

Masson, F., Zaidman-Rémy, A., Heddi, A., 2016. Antimicrobial peptides and cell processes tracking endosymbiont dynamics. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 371, 20150298. https://doi.org/10.1098/rstb.2015.0298

McCutcheon, J.P., Moran, N.A., 2011. Extreme genome reduction in symbiotic bacteria. Nat. Rev. Microbiol. 10, 13–26. https://doi.org/10.1038/nrmicro2670

McFadden, G.I., 2014. Origin and evolution of plastids and photosynthesis in eukaryotes. Cold Spring Harb. Perspect. Biol. 6, a016105. https://doi.org/10.1101/cshperspect.a016105

McFadden, G.I., 2001. Primary and Secondary Endosymbiosis and the Origin of Plastids. J. Phycol. 37, 951–959. https://doi.org/10.1046/j.1529-8817.2001.01126.x

Menardo, F., Loiseau, C., Brites, D., Coscolla, M., Gygli, S.M., Rutaihwa, L.K., Trauner, A., Beisel, C., Borrell, S., Gagneux, S., 2018. Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity. BMC Bioinformatics 19, 164. https://doi.org/10.1186/s12859-018-2164-8

Mergaert, P., 2018. Role of antimicrobial peptides in controlling symbiotic bacterial populations. Nat. Prod. Rep. 35, 336–356. https://doi.org/10.1039/c7np00056a

Mergaert, P., Kikuchi, Y., Shigenobu, S., Nowack, E.C.M., 2017. Metabolic Integration of Bacterial Endosymbionts through Antimicrobial Peptides. Trends Microbiol. 25, 703–712. https://doi.org/10.1016/j.tim.2017.04.007

Mirarab, S., Reaz, R., Bayzid, Md.S., Zimmermann, T., Swenson, M.S., Warnow, T., 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30, i541–i548. https://doi.org/10.1093/bioinformatics/btu462

Moran, N.A., 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc. Natl. Acad. Sci. U. S. A. 93, 2873–2878. https://doi.org/10.1073/pnas.93.7.2873

Moreira, D., Deschamps, P., 2014. What Was the Real Contribution of Endosymbionts to the Eukaryotic Nucleus? Insights from Photosynthetic Eukaryotes. Cold Spring Harb. Perspect. Biol. 6. https://doi.org/10.1101/cshperspect.a016014

Moreira, D., Le Guyader, H., Philippe, H., 2000. The origin of red algae and the evolution of chloroplasts. Nature 405, 69–72. https://doi.org/10.1038/35011054

Moustafa, A., Reyes-Prieto, A., Bhattacharya, D., 2008. Chlamydiae Has Contributed at Least 55 Genes to Plantae with Predominantly Plastid Functions. PLoS ONE 3. https://doi.org/10.1371/journal.pone.0002205

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol. Biol. Evol. 32, 268–274. https://doi.org/10.1093/molbev/msu300

Nowack, E.C.M., Melkonian, M., Glöckner, G., 2008. Chromatophore genome sequence of Paulinella sheds light on acquisition of photosynthesis by eukaryotes. Curr. Biol. CB 18, 410–418. https://doi.org/10.1016/j.cub.2008.02.051

Nowicka, B., Kruk, J., 2010. Occurrence, biosynthesis and function of isoprenoid quinones. Biochim. Biophys. Acta BBA - Bioenerg. 1797, 1587–1605. https://doi.org/10.1016/j.bbabio.2010.06.007

Nozaki, H., Maruyama, S., Matsuzaki, M., Nakada, T., Kato, S., Misawa, K., 2009. Phylogenetic positions of Glaucophyta, green plants (Archaeplastida) and Haptophyta (Chromalveolata) as deduced from slowly evolving nuclear genes. Mol. Phylogenet. Evol. 53, 872–880. https://doi.org/10.1016/j.ympev.2009.08.015

Ochoa de Alda, J.A.G., Esteban, R., Diago, M.L., Houmard, J., 2014. The plastid ancestor originated among one of the major cyanobacterial lineages. Nat. Commun. 5, 4937. https://doi.org/10.1038/ncomms5937

Omsland, A., Sixt, B.S., Horn, M., Hackstadt, T., 2014. Chlamydial metabolism revisited: interspecies metabolic variability and developmental stage-specific physiologic activities. FEMS Microbiol. Rev. 38, 779–801. https://doi.org/10.1111/1574-6976.12059

Oostende, C. van, Widhalm, J.R., Basset, G.J.C., 2008. Detection and quantification of vitamin K(1) quinol in leaf tissues. Phytochemistry 69, 2457–2462.

https://doi.org/10.1016/j.phytochem.2008.07.006

Ponce-Toledo, R.I., Deschamps, P., López-García, P., Zivanovic, Y., Benzerara, K., Moreira, D., 2017. An early-branching freshwater cyanobacterium at the origin of plastids. Curr. Biol. CB 27, 386–391. https://doi.org/10.1016/j.cub.2016.11.056

Ponce-Toledo, R.I., López-García, P., Moreira, D., 2019. Horizontal and endosymbiotic gene transfer in early plastid evolution. New Phytol. 224, 618–624. https://doi.org/10.1111/nph.15965

Poole, P., Ramachandran, V., Terpolilli, J., 2018. Rhizobia: from saprophytes to endosymbionts. Nat. Rev. Microbiol. 16, 291–303. https://doi.org/10.1038/nrmicro.2017.171

Price, D.C., Chan, C.X., Yoon, H.S., Yang, E.C., Qiu, H., Weber, A.P.M., Schwacke, R., Gross, J., Blouin, N.A., Lane, C., Reyes-Prieto, A., Durnford, D.G., Neilson, J.A.D., Lang, B.F., Burger, G., Steiner, J.M., Löffelhardt, W., Meuser, J.E., Posewitz, M.C., Ball, S., Arias, M.C., Henrissat, B., Coutinho, P.M., Rensing, S.A., Symeonidi, A., Doddapaneni, H., Green, B.R., Rajah, V.D., Boore, J., Bhattacharya, D., 2012. Cyanophora paradoxa genome elucidates origin of photosynthesis in algae and plants. Science 335, 843–847. https://doi.org/10.1126/science.1213561

Price, D.C., Goodenough, U.W., Roth, R., Lee, J.-H., Kariyawasam, T., Mutwil, M., Ferrari, C., Facchinelli, F., Ball, S.G., Cenci, U., Chan, C.X., Wagner, N.E., Yoon, H.S., Weber, A.P.M., Bhattacharya, D., 2019. Analysis of an improved Cyanophora paradoxa genome assembly. DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes. https://doi.org/10.1093/dnares/dsz009

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 Plant Sci. 18, 680–687. https://doi.org/10.1016/j.tplants.2013.09.007

Raven, J.A., Allen, J.F., 2003. Genomics and chloroplast evolution: what did cyanobacteria do for plants? Genome Biol. 4, 209. https://doi.org/10.1186/gb-2003-4-3-209

Reumann, S., 2013. Biosynthesis of vitamin K1 (phylloquinone) by plant peroxisomes and its integration into signaling molecule synthesis pathways. Subcell. Biochem. 69, 213–229. https://doi.org/10.1007/978-94-007-6889-5_12

Reyes-Prieto, A., Weber, A.P.M., Bhattacharya, D., 2007. The origin and establishment of the plastid in algae and plants. Annu. Rev. Genet. 41, 147–168. https://doi.org/10.1146/annurev.genet.41.110306.130134

Rodríguez-Ezpeleta, N., Brinkmann, H., Burey, S.C., Roure, B., Burger, G., Löffelhardt, W., Bohnert, H.J., Philippe, H., Lang, B.F., 2005. Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. Curr. Biol. CB 15, 1325–1330. https://doi.org/10.1016/j.cub.2005.06.040

Roure, B., Rodriguez-Ezpeleta, N., Philippe, H., 2007. SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. BMC Evol. Biol. 7 Suppl 1, S2. https://doi.org/10.1186/1471-2148-7-S1-S2

Sagan, L., 1967. On the origin of mitosing cells. J. Theor. Biol. 14, 225-IN6. https://doi.org/10.1016/0022-5193(67)90079-3

Sato, N., 2021. Are Cyanobacteria an Ancestor of Chloroplasts or Just One of the Gene Donors for Plants and Algae? Genes 12, 823. https://doi.org/10.3390/genes12060823

Sato, N., 2019. Phylogenetic Evidence for the Endosymbiotic Origin of Organelles, in: Sato, N. (Ed.), Endosymbiotic Theories of Organelles Revisited: Retrospects and Prospects. Springer, Singapore, pp. 97–120. https://doi.org/10.1007/978-981-15-1161-5_6

Sharma, P., Teixeira de Mattos, M.J., Hellingwerf, K.J., Bekker, M., 2012. On the function of the various quinone species in Escherichia coli. FEBS J. 279, 3364–3373. https://doi.org/10.1111/j.1742-4658.2012.08608.x

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinforma. Oxf. Engl. 30, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stephens, R.S., Kalman, S., Lammel, C., Fan, J., Marathe, R., Aravind, L., Mitchell, W., Olinger, L., Tatusov, R.L., Zhao, Q., Koonin, E.V., Davis, R.W., 1998. Genome Sequence of an Obligate Intracellular Pathogen of Humans: Chlamydia trachomatis. Science 282, 754–759. https://doi.org/10.1126/science.282.5389.754

Stephens, T.G., Bhattacharya, D., Ragan, M.A., Chan, C.X., 2016. PhySortR: a fast, flexible tool for sorting phylogenetic trees in R. PeerJ 4. https://doi.org/10.7717/peerj.2038

Strassert, J.F.H., Irisarri, I., Williams, T.A., Burki, F., 2021. A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. Nat. Commun. 12, 1879. https://doi.org/10.1038/s41467-021-22044-z

Turner, S., Pryer, K.M., Miao, V.P., Palmer, J.D., 1999. Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. J. Eukaryot. Microbiol. 46, 327–338.

van Dooren, G.G., Schwartzbach, S.D., Osafune, T., McFadden, G.I., 2001. Translocation of proteins across the multiple membranes of complex plastids. Biochim. Biophys. Acta 1541, 34–53. https://doi.org/10.1016/s0167-4889(01)00154-9

Van Vlierberghe, M., Philippe, H., Baurain, D., 2021. Broadly sampled orthologous groups of eukaryotic proteins for the phylogenetic study of plastid-bearing lineages. BMC Res. Notes 14, 143. https://doi.org/10.1186/s13104-021-05553-4

Vos, M., Esposito, G., Edirisinghe, J.N., Vilain, S., Haddad, D.M., Slabbaert, J.R., Van Meensel, S., Schaap, O., De Strooper, B., Meganathan, R., Morais, V.A., Verstreken, P., 2012. Vitamin K2 is a mitochondrial electron carrier that rescues pink1 deficiency. Science 336, 1306–1310. https://doi.org/10.1126/science.1218632

Vries, J. de, Archibald, J.M., 2018. Plastid genomes. Curr. Biol. 28, R336–R337. https://doi.org/10.1016/j.cub.2018.01.027

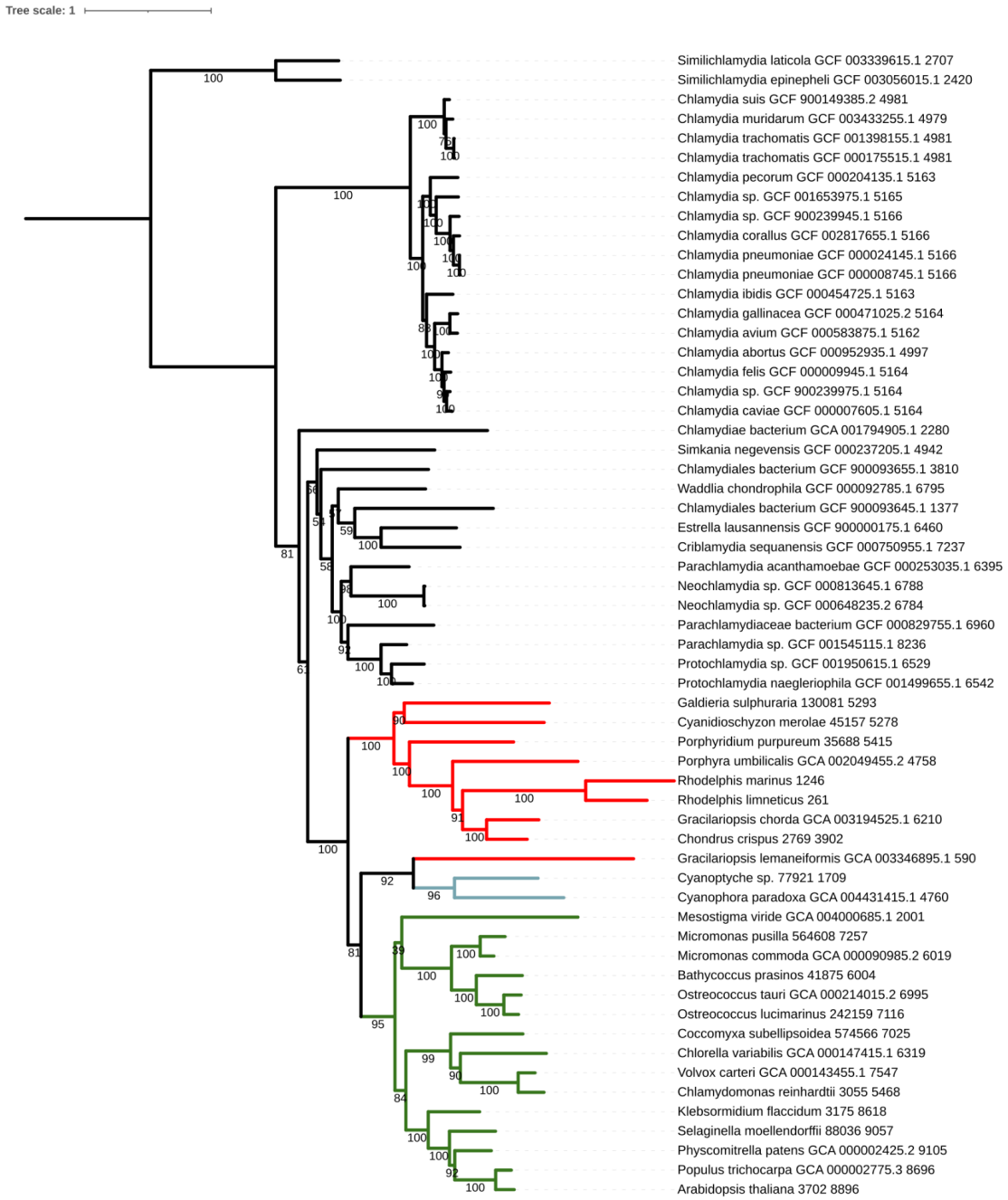
Weber, A.P.M., Linka, M., Bhattacharya, D., 2006. Single, ancient origin of a plastid metabolite translocator family in Plantae from an endomembrane-derived ancestor. Eukaryot. Cell 5, 609–612. https://doi.org/10.1128/EC.5.3.609-612.2006

Wickham, H., 2009. Getting started with qplot, in: Wickham, H. (Ed.), Ggplot2: Elegant Graphics for Data Analysis, Use R. Springer, New York, NY, pp. 9–26. https://doi.org/10.1007/978-0-387-98141-3_2

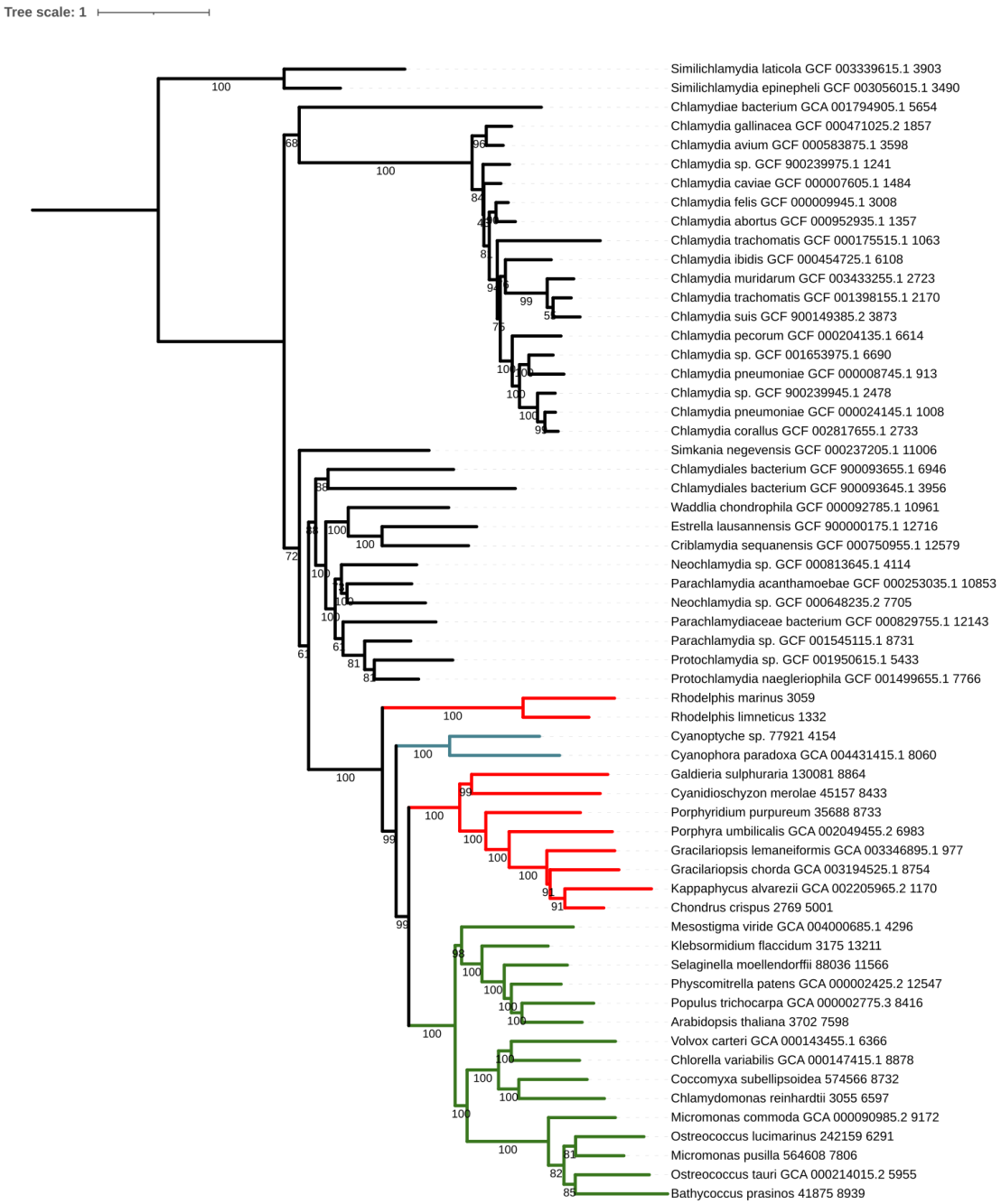
Yoon, H.S., Hackett, J.D., Ciniglia, C., Pinto, G., Bhattacharya, D., 2004. A Molecular Timeline for the Origin of Photosynthetic Eukaryotes. Mol. Biol. Evol. 21, 809–818. https://doi.org/10.1093/molbev/msh075

Yoshida, E., Nakamura, A., Watanabe, T., 2003. Reversed-phase HPLC determination of chlorophyll a' and naphthoquinones in photosystem I of red algae: existence of two menaquinone-4 molecules in photosystem I of Cyanidium caldarium. Anal. Sci. Int. J. Jpn. Soc. Anal. Chem. 19, 1001–1005.

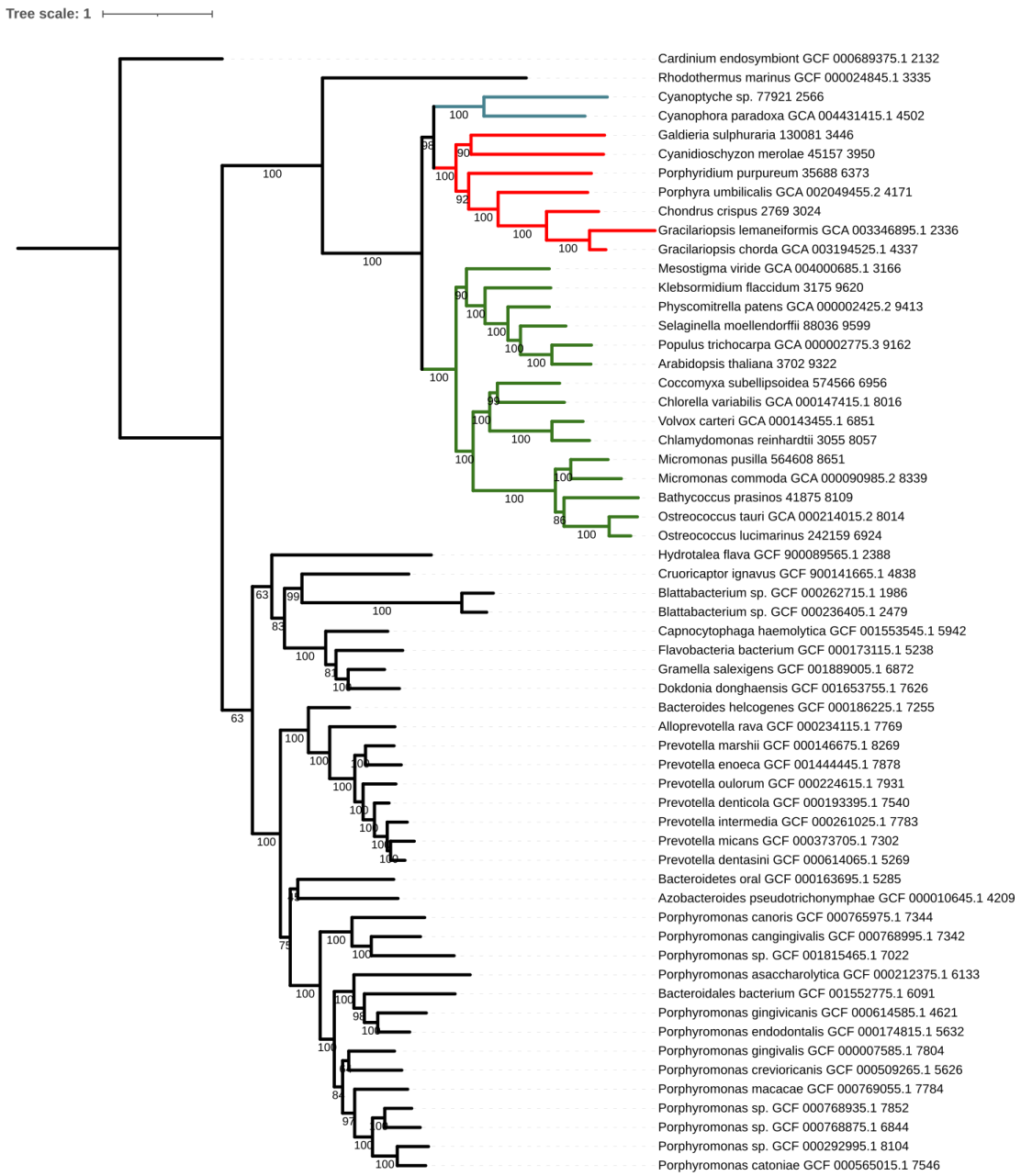
Zhi, X.-Y., Yao, J.-C., Tang, S.-K., Huang, Y., Li, H.-W., Li, W.-J., 2014. The Futalosine Pathway Played an Important Role in Menaquinone Biosynthesis during Early Prokaryote Evolution. Genome Biol. Evol. 6, 149–160. https://doi.org/10.1093/gbe/evu007



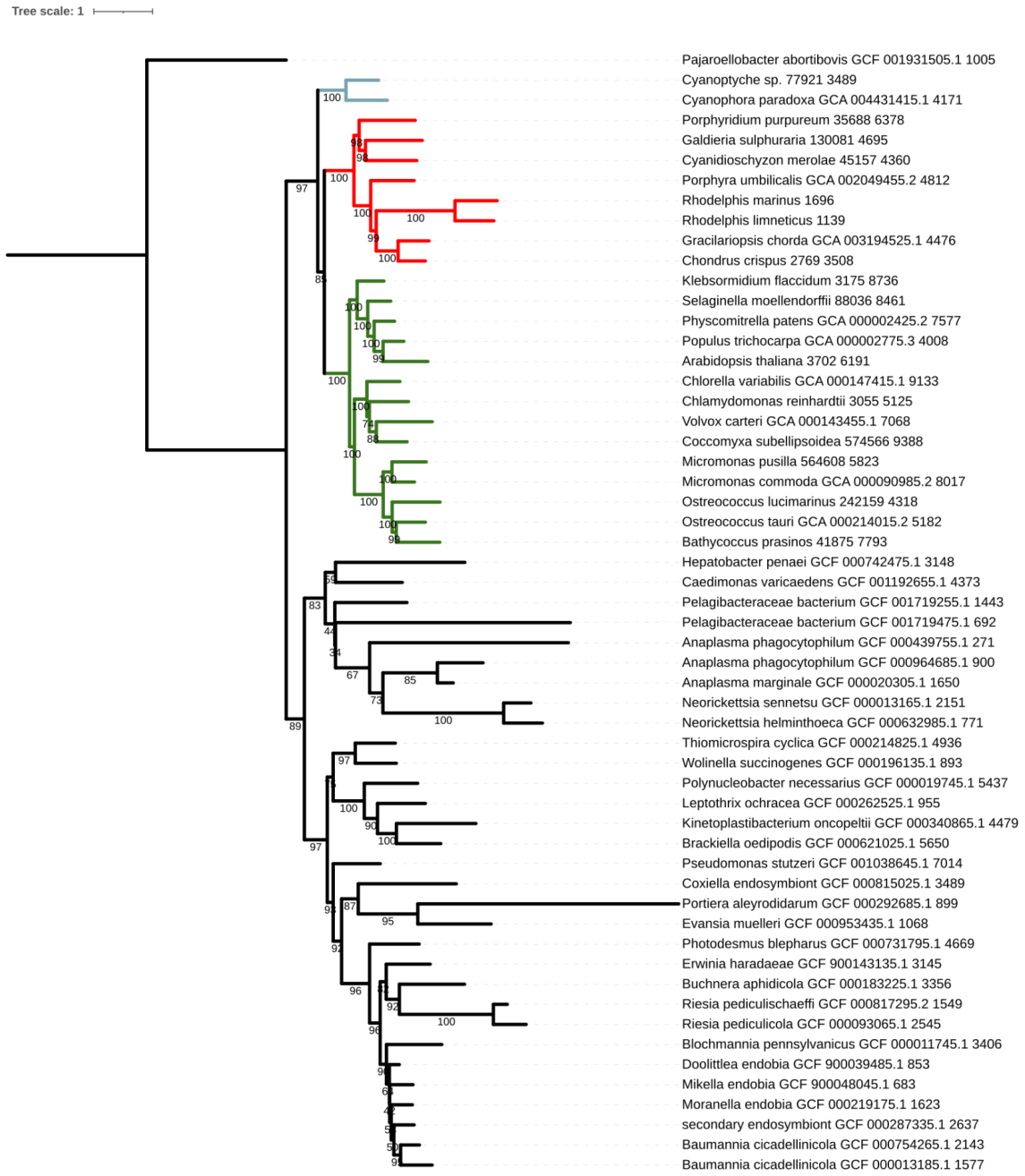
Appendix 1: Phylogenetic tree from the concatenation of the 26 genes selected by the manual analysis of the chlamydial pipeline, under a C60 model. The tree was obtained by IQ-TREE, under a C60 model, after concatenation of the chlamydial and archaeplastida sequences of the genes selected by the pipeline. Bootstrap values are shown on the branches. Rooting is manual on Similichlamydia. The scale counts in number of amino acid substitutions per site. In red: Rhodophyta and Rhodelphea, in green: Viridiplantae and in blue: Glaucophyta.



Appendix 2: Phylogenetic tree from the concatenation of the 57 genes selected by the automatic chlamydial pipeline, under a C60 model. The tree was obtained by IQ-TREE, under a C60 model, after concatenation of the chlamydial and archaeplastid sequences of the genes selected by the pipeline. Bootstrap values are shown on the branches. Rooting is manual on Similichlamydia. The scale counts in number of amino acid substitutions per site. In red: Rhodophyta and Rhodelphea, in green: Viridiplantae and in blue: Glaucophyta.



Appendix 3: Phylogenetic tree from the concatenation of the 44 genes selected by the automatic Bacteroidetes pipeline, under a C60 model. The tree was obtained by IQ-TREE, under a C60 model, after concatenation of the chlamydial and archaeplastida sequences of the genes selected by the pipeline. Bootstrap values are shown on the branches. Rooting is manual on basal Bacteroidetes. The scale counts in number of amino acid substitutions per site. In red: Rhodophyta and Rhodelphea, in green: Viridiplantae and in blue: Glaucophyta.



Appendix 4: Phylogenetic tree from the concatenation of the 39 genes selected by the Proteobacteria automatic pipeline, under a C60 model. The tree was obtained by IQ-TREE, under a C60 model, after concatenation of Proteobacteria and Archaeplastida sequences of the genes selected by the pipeline. Bootstrap values are shown on the branches. Rooting is manual on basal Proteobacteria. The scale counts in number of amino acid substitutions per site. In red: Rhodophyta and Rhodelphea, in green: Viridiplantae and in blue: Glaucophyta.

Appendix 5: Configuration file for the quality assessment of the chlamydial proteomes with 42.

```
# ===Run mode===
# Two values are available: 'phylogenomic' and 'metagenomic'.
# The phylogenomic mode is designed to enrich multiple sequence alignements
# (ALIs) with orthologues for subsequent phylogenomic analysis. In contrast,
# the metagenomic mode is designed to probe contamination in transcriptomic
# data using reference ribosomal protein ALIs; it produces a taxonomic report
# per ALI listing the lineage of each identified orthologous sequence.
# When not specified, 'run_mode' internally defaults to 'phylogenomic'.
run_mode: metagenomic

# ===Suffix to append to infile basenames for deriving outfile names===
# When not specified 'outsuffix' internally defaults to '-42'.
# Use a bare 'out_suffix:' to reuse the ALI name and to preserve the original
# file by appending a .bak extension to its name.
out_suffix: -42-chlam

# ===Orgs from where to select BLAST queries===
# Depending on availability at least one query by family and by org will be
# picked for the 'homologues' and 'references' BLAST rounds.
query_orgs:
        - Bacillus anthracis_1392
        - Brucella suis_645170
        - Burkholderia mallei_13373
        - Chlamydia pneumoniae_83558
        - Corynebacterium pseudotuberculosis_1719
        - Escherichia coli_83333
        - Flavobacterium psychrophilum_96345
        - Francisella philomiragia_28110
        - Helicobacter pylori_210
        - Listeria monocytogenes_1639
        - Mycobacterium tuberculosis_1773
        - Neisseria meningitidis_487
        - Pseudomonas aeruginosa_287
        - Staphylococcus aureus_1074919
        - Streptococcus agalactiae_1311
        - Sulfolobus solfataricus_2287
        - Thermoproteus uzoniensis_999630
        - Vulcanisaeta moutnovskia_985053
        - Xanthomonas citri_611301
        - Yersinia pestis_632

# ===Optional args for each BLAST step===
# Any valid command-line option can be specified (see NCBIBLAST+ docs).
# Note the hyphens (-) before option names (departing from API consistency).
# -query, -db, -out, -outfmt, -db_gencode, -query_gencode will be ignored as
# they are directly handled by forty-two itself. -max_target_seqs may be
# specified at step 'homologues' to speed up things.
blast_args:
        # TBLASTN vs banks
        homologues:
        -evalue: 1e-05
        -seg: yes
        -num_threads: 1
        -max_target_seqs: 10000
        # BLASTP vs ref banks (for transitive BRH ; actually two steps)
        references:
        -evalue: 1e-05
        -num_threads: 1
        # BLASTX vs ref banks (for transitive BRH)
        orthologues:
        -evalue: 1e-05
```

```
        -num_threads: 1
        # BLASTX vs ALI (for tax filters and alignment)
        templates:
        -evalue: 1e-05
        -seg: no
        -num_threads: 1


# ===BRH mode for assessing orthology===
# Two values are available: 'on' and 'off'.
# If set to 'on', a candidate seq must be in BRH with all reference proteomes
# to be considered as an orthologous seq. In contrast, all candidate seqs are
# considered as orthologous seqs when this parameter is set to 'off'.
# When not specified, 'ref_brh_mode' internally defaults to 'on'.
# To limit the number of candidate seqs, use the '-max_target_seqs' option of
# the BLAST executable(s) at the 'homologues' step.
ref_brh_mode: on


# ===Path to dir holding complete proteome BLAST databases===
# Only required when setting 'ref_brh_mode' to 'on'.
ref_bank_dir: /media/vol2/home/mleleu/Forty-Two/ref_banks/Prokaryotes

# ===Basenames of complete proteome BLAST databases (keyed by org name)===
# Only required when setting 'ref_brh_mode' to 'on'.
# You can list as many databases as needed here.
# Only those specified as 'ref_orgs' below will actually be used for BRH.
ref_org_mapper:
        Acidobacterium capsulatum_GCA_000022565.1:
Acidobacterium_capsulatum_atcc_51196.GCA_000022565.1.30.pep.all_taxid
        Korarchaeum cryptofilum_GCA_000019605.1:
Candidatus_korarchaeum_cryptofilum_opf8.GCA_000019605.1.30.pep.all_taxid
        Nitrosopumilus adriaticus_GCA_000956175.1:
Candidatus_nitrosopumilus_sp_nf5.GCA_000956175.1.30.pep.all_taxid
        Chitinophaga pinensis_GCA_000024005.1:
Chitinophaga_pinensis_dsm_2588.GCA_000024005.1.30.pep.all_taxid
        Chlamydia trachomatis_GCA_000026905.1:
Chlamydia_trachomatis_b_tz1a828_ot.GCA_000026905.1.30.pep.all_taxid
        Escherichia coli_GCA_000026545.1:
Escherichia_coli_o127_h6_str_e2348_69.GCA_000026545.1.30.pep.all_taxid
        Ignisphaera aggregans_GCA_000145985.1:
Ignisphaera_aggregans_dsm_17230.GCA_000145985.1.30.pep.all_taxid
        Ilyobacter polytropus_GCA_000165505.1:
Ilyobacter_polytropus_dsm_2926.GCA_000165505.1.30.pep.all_taxid
        Nanoarchaeum equitans_GCA_000008085.1:
Nanoarchaeum_equitans_kin4_m.GCA_000008085.1.30.pep.all_taxid
        Nitrososphaera viennensis_GCA_000698785.1:
Nitrososphaera_viennensis_en76.GCA_000698785.1.30.pep.all_taxid
        Palaeococcus pacificus_GCA_000725425.1:
Palaeococcus_pacificus_dy20341.GCA_000725425.1.30.pep.all_taxid
        Propionibacterium freudenreichii_GCA_000091725.1:
Propionibacterium_freudenreichii_subsp_shermanii_cirm_bia1.GCA_000091725.1.30.pep.all_taxid
        Staphylococcus aureus_GCA_000011505.1:
Staphylococcus_aureus_subsp_aureus_mrsa252.GCA_000011505.1.30.pep.all_taxid
        Sulfolobus acidocaldarius_GCA_000012285.1:
Sulfolobus_acidocaldarius_dsm_639.GCA_000012285.1.30.pep.all_taxid
        Synechococcus sp._GCA_000014585.1: Synechococcus_sp_cc9311.GCA_000014585.1.30.pep.all_taxid

# ===Orgs to be used for BRH checks===
# Only required when setting 'ref_brh_mode' to 'on'.
# To be considered as an orthologue, a candidate seq must be in transitive BRH
# for all listed orgs (and not for only one of them).
# Listing more orgs thus increases the stringency of the BRH check. Note that
# 'ref_orgs' may but DO NOT NEED to match 'query_orgs'.
```

```
ref_orgs:
        - Acidobacterium capsulatum_GCA_000022565.1
        - Korarchaeum cryptofilum_GCA_000019605.1
        - Nitrosopumilus adriaticus_GCA_000956175.1
        - Chitinophaga pinensis_GCA_000024005.1
        - Chlamydia trachomatis_GCA_000026905.1
        - Escherichia coli_GCA_000026545.1
        - Ignisphaera aggregans_GCA_000145985.1
        - Ilyobacter polytropus_GCA_000165505.1
        - Nanoarchaeum equitans_GCA_000008085.1
        - Nitrososphaera viennensis_GCA_000698785.1
        - Palaeococcus pacificus_GCA_000725425.1
        - Propionibacterium freudenreichii_GCA_000091725.1
        - Staphylococcus aureus_GCA_000011505.1
        - Sulfolobus acidocaldarius_GCA_000012285.1
        - Synechococcus sp._GCA_000014585.1

# ===Fraction of ref_orgs to really use when assessing orthology===
# Only meaningful when setting 'ref_brh_mode' to 'on'.
# This parameter introduces some flexibility when using reference proteomes.
# If set to a fractional value (below 1), only the best proteomes will be
# considered during BRHs. The best proteomes are those against which the
# queries have the highest average scores. This helps discarding ref_orgs that
# might hinder orthology assessment because they lack the orthologous gene(s).
# When not specified, 'reg_org_mul' internally defaults to 1.0, which is the
# strictest mode since all reference proteomes are used during BRHs.
ref_org_mul: 0.3

# ===Bit score reduction allowed when including non-1st hits among best hits===
# Only meaningful when setting 'ref_brh_mode' to 'on'.
# This parameter applies when collecting best hits for queries to complete
# proteomes, so that close in-paralogues can all be included in the set of
# best hits. The allowed bit score reduction of any hit is expressed relatively
# to the score of the previous hit. During BRH checks, only the very first hit
# for the candidate seq is actually tested for inclusion in this set but for
# all complete proteomes. By default at most 10 hits are considered. To change
# this, use the '-max_target_seqs' option of the BLAST executable(s) at the
# 'reference' step.
# When not specified 'ref_score_mul' internally defaults to 1.0, which is the
# strictest mode since only equally-best hits are retained.
ref_score_mul: 0.99



# ===Hit trimming switch===
# Two values are available: 'on' and 'off'.
# If set to 'on', each candidate seq is first trimmed to the range covered by
# the HSPs that retrieved it. This helps exonerate to splice genes correctly.
# The details of this trimming step can be fine-tuned by editing the other
# hit_* parameters of this configuration file.
# When not specified, 'trimming_mode' internally defaults to 'on'.
trimming_mode: off



# ===Action to perform when a preexisting lengthened seq is identified===
# Currently, two values are available: 'remove' and 'keep'.
# When not specified, 'ls_action' internally defaults to 'keep'.
ls_action: keep


# ===Engine to be used for aligning new seqs===
# Four values are available: 'blast', 'exonerate', 'exoblast' and 'off'.
```

```
# If the alignment engine is 'off', new seqs are added 'as is' to the ALI.
# Consequently, they will be full length but not aligned to existing seqs.
# This mode is meant for protein seqs only and thus cannot be used when adding
# transcripts from nucleotide banks.
# The 'exonerate' mode sometimes fails to align orthologous seqs due to a bug
# in exonerate executable. This causes new seqs to be lost. To automatically
# retry aligning them with BLAST in case of failure, use the 'exoblast' mode.
# When not specified, 'aligner_mode' internally defaults to 'blast'.
aligner_mode: off



# ===Taxonomic report switch===
# Two values are available: 'on' and 'off'.
# If set to 'on', the lineage of new seqs is inferred by analyzing the taxonomy
# of their ALI closest relatives and one '.tax-report' file is generated for
# each ALI processed (see 'run_mode' above).
# The details of this taxonomic analysis can be fine-tuned by editing the other
# tax_* parameters of this configuration file.
# When not specified, 'tax_reports' internally defaults to 'off'. Yet, the YAML
# generator automatically sets it to 'on' if 'run_mode' is 'metagenomic'.
tax_reports: on

# ===Path to dir holding NCBI Taxonomy database===
# Only required when enabling 'tax_reports' or specifying 'tax_filter'.
# It can be installed using setup-taxdir.pl.
tax_dir: /media/vol2/home/mleleu/taxdump

# ===Min number of relatives to use when inferring taxonomy of new seqs===
# Only meaningful when enabling 'tax_reports' or specifying 'tax_filter'.
# This parameter is a lower bound. The real number will depend both on the
# four thresholds below ('tax_min_ident', 'tax_min_len', 'tax_min_score' and
# 'tax_score_mul') and on the ability of 42 to deduce the taxonomy of each
# individual relative to compute the LCA of the new seq.
# When not specified, 'tax_min_hits' internally defaults to 1.
tax_min_hits: 1

# ===Max number of relatives to use when inferring taxonomy of new seqs===
# Only meaningful when enabling 'tax_reports' or specifying 'tax_filter'.
# As for 'tax_min_hits' above, this parameter is a upper bound.
# When not specified, 'tax_max_hits' internally defaults to unlimited.
tax_max_hits: 100

# ===Min identity of relatives to use when inferring taxonomy of new seqs===
# Only meaningful when enabling 'tax_reports' or specifying 'tax_filter'.
# This parameter is the traditional BLAST 'percent identity' statistics except
# that it is specified as a fractional number (between 0 and 1). It is
# evaluated on the first HSP of potential relatives.
# When not specified, 'tax_min_ident' internally defaults to 0.
tax_min_ident: 0
# ===Min length of relatives to use when inferring taxonomy of new seqs===
# Only meaningful when enabling 'tax_reports' or specifying 'tax_filter'.
# This parameter is the traditional BLAST 'alignment length' statistics. It is
# evaluated on the first HSP of potential relatives.
# When not specified, 'tax_min_len' internally defaults to 0.
tax_min_len: 0
# ===Min bit score of relatives to use when inferring taxonomy of new seqs===
# Only meaningful when enabling 'tax_reports' or specifying 'tax_filter'.
# This parameter is the traditional BLAST 'bit score' statistics. It is
# evaluated on the first HSP of potential relatives.
# When not specified, 'tax_min_score' internally defaults to 0.
tax_min_score: 80
```

```
# ===Bit score reduction allowed when inferring taxonomy of new seqs===
# Only meaningful when enabling 'tax_reports' or specifying 'tax_filter'.
# The allowed bit score reduction of any relative is expressed relatively
# to the score of the FIRST relative (as in MEGAN algorithm).
# When not specified, 'tax_score_mul' internally defaults to 0.
tax_score_mul: 0.95

# ===Path to dir holding transcript BLAST databases===
bank_dir: /media/vol2/home/mleleu/Busco/Chlamydia/proteomes

# ===Default args applying to all orgs unless otherwise specified===
# Some of these args can be thus specified on a per-org basis below if needed.
# This especially makes sense for 'code'.
defaults:
        # ===Genetic code for translated BLAST searches===
        # When not specified 'code' internally defaults to 1 (standard).
        # See ftp://ftp.ncbi.nih.gov/entrez/misc/data/gc.prt for other codes.
        code: 1

# ===Org-specific args===
# The only mandatory args are 'org' and 'banks'. All other args are taken from
# the 'defaults:' section described above.
# This part can be concatenated on a per-run basis to the previous part, which
# would be the same for several runs. In the future, forty-two might support
# two different configuration files to reflect this conceptual distinction.
orgs:
 - org: Chlamydiae bacterium_GCA_001794905.1
        banks:
        - GCA_001794905.1_ASM179490v1_protein.faa

 - org: Chlamydia caviae_GCF_000007605.1
        banks:
        - GCF_000007605.1_ASM760v1_protein.faa

 - org: Chlamydia pneumoniae_GCF_000008745.1
        banks:
        - GCF_000008745.1_ASM874v1_protein.faa

 - org: Chlamydia felis_GCF_000009945.1
        banks:
        - GCF_000009945.1_ASM994v1_protein.faa

 - org: Chlamydia pneumoniae_GCF_000024145.1
        banks:
        - GCF_000024145.1_ASM2414v1_protein.faa

 - org: Waddlia chondrophila_GCF_000092785.1
        banks:
        - GCF_000092785.1_ASM9278v1_protein.faa

 - org: Chlamydia trachomatis_GCF_000175515.1
        banks:
        - GCF_000175515.1_ASM17551v1_protein.faa

 - org: Chlamydia pecorum_GCF_000204135.1
        banks:
        - GCF_000204135.1_ASM20413v1_protein.faa

 - org: Simkania negevensis_GCF_000237205.1
        banks:
        - GCF_000237205.1_ASM23720v1_protein.faa

 - org: Parachlamydia acanthamoebae_GCF_000253035.1
        banks:
```

- GCF_000253035.1_ASM25303v1_protein.faa

- org: Chlamydia ibidis_GCF_000454725.1
  banks:
  - GCF_000454725.1_ibidis.assembly_protein.faa

- org: Chlamydia gallinacea_GCF_000471025.2
  banks:
  - GCF_000471025.2_ASM47102v2_protein.faa

- org: Chlamydia avium_GCF_000583875.1
  banks:
  - GCF_000583875.1_ASM58387v1_protein.faa

- org: Neochlamydia sp._GCF_000648235.2
  banks:
  - GCF_000648235.2_ASM64823v2_protein.faa

- org: Criblamydia sequanensis_GCF_000750955.1
  banks:
  - GCF_000750955.1_CS_CRIB18-1_protein.faa

- org: Neochlamydia sp._GCF_000813645.1
  banks:
  - GCF_000813645.1_TUME1_v1_protein.faa

- org: Parachlamydiaceae bacterium_GCF_000829755.1
  banks:
  - GCF_000829755.1_ASM82975v1_protein.faa

- org: Chlamydia abortus_GCF_000952935.1
  banks:
  - GCF_000952935.1_CAAB7_protein.faa

- org: Chlamydia trachomatis_GCF_001398155.1
  banks:
  - GCF_001398155.1_7501_6_49_protein.faa

- org: Protochlamydia naegleriophila_GCF_001499655.1
  banks:
  - GCF_001499655.1_PNK1_protein.faa

- org: Parachlamydia sp._GCF_001545115.1
  banks:
  - GCF_001545115.1_Protochlamydia_greubae_protein.faa

- org: Chlamydia sp._GCF_001653975.1
  banks:
  - GCF_001653975.1_ASM165397v1_protein.faa

- org: Protochlamydia sp._GCF_001950615.1
  banks:
  - GCF_001950615.1_ASM195061v1_protein.faa

- org: Chlamydia corallus_GCF_002817655.1
  banks:
  - GCF_002817655.1_ASM281765v1_protein.faa

- org: Similichlamydia epinepheli_GCF_003056015.1
  banks:
  - GCF_003056015.1_ASM305601v1_protein.faa

- org: Similichlamydia laticola_GCF_003339615.1
  banks:

```
        - GCF_003339615.1_ASM333961v1_protein.faa

 - org: Chlamydia muridarum_GCF_003433255.1
        banks:
        - GCF_003433255.1_ASM343325v1_protein.faa

 - org: Estrella lausannensis_GCF_900000175.1
        banks:
        - GCF_900000175.1_ASM90000017v1_protein.faa

 - org: Chlamydiales bacterium_GCF_900093645.1
        banks:
        - GCF_900093645.1_AB751O23_protein.faa

 - org: Chlamydiales bacterium_GCF_900093655.1
        banks:
        - GCF_900093655.1_SCG7086_protein.faa

 - org: Chlamydia suis_GCF_900149385.2
        banks:
        - GCF_900149385.2_4-29b_chromosome_protein.faa

 - org: Chlamydia sp._GCF_900239945.1
        banks:
        - GCF_900239945.1_Chlamydia_sp._nov._H15-1957-10C_protein.faa

 - org: Chlamydia sp._GCF_900239975.1
        banks:
        - GCF_900239975.1_Chlamydia_sp._S15-834K_protein.faa


#
# This config file has been generated automatically on 09:53:46 16-Apr-2019.
# We advise not to modify directly this file manually but rather to modify
# the yaml-generator command instead for traceability and reproducibility.
#
#yaml-generator-42.pl --run_mode=metagenomic --out_suffix=-42-chlam \
#--queries /media/vol2/home/mvanvlierberghe/databases/ribo_prots/prokaryotes/queries.idl \
#--evalue=1e-05 --homologues_seg=yes --max_target_seqs=10000 --templates_seg=no \
#--bank_dir /media/vol2/home/mleleu/Busco/Chlamydia/proteomes --bank_suffix=.psq --bank_mapper
/media/vol2/home/mleleu/Busco/Chlamydia/proteomes/chlamydia-bank-mapper.idm \
#--ref_brh_mode=on --ref_bank_dir /media/vol2/home/mleleu/Forty-Two/ref_banks/Prokaryotes
--ref_bank_suffix=.psq --ref_bank_mapper
/media/vol2/home/mvanvlierberghe/databases/ref_banks/prokaryotes/proka_ref_bank_mapper.idm \
#--ref_org_mul=0.3 --ref_score_mul=0.99 \
#--trimming_mode=off \
#--ls_action=keep --aligner_mode=off \
#--tax_reports=on --tax_dir /media/vol2/home/mleleu/taxdump \
#--megan_like \
#--tol_check=off
```

Appendix 6: Configuration file for the enrichissement of orthologous groups with 42.

```
# ===Run mode===
# Two values are available: 'phylogenomic' and 'metagenomic'.
# The phylogenomic mode is designed to enrich multiple sequence alignements
# (ALIs) with orthologues for subsequent phylogenomic analysis. In contrast,
# the metagenomic mode is designed to probe contamination in transcriptomic
# data using reference ribosomal protein ALIs; it produces a taxonomic report
# per ALI listing the lineage of each identified orthologous sequence.
# When not specified, 'run_mode' internally defaults to 'phylogenomic'.
run_mode: phylogenomic

# ===Suffix to append to infile basenames for deriving outfile names===
# When not specified 'outsuffix' internally defaults to '-42'.
# Use a bare 'out_suffix:' to reuse the ALI name and to preserve the original
# file by appending a .bak extension to its name.
out_suffix: -g


# ===Orgs from where to select BLAST queries===
# Depending on availability at least one query by family and by org will be
# picked for the 'homologues' and 'references' BLAST rounds.
query_orgs:
        - Arabidopsis thaliana_3702
        - Bathycoccus prasinos_41875
        - Chlamydomonas reinhardtii_3055
        - Chondrus crispus_2769
        - Coccomyxa subellipsoidea_574566
        - Cyanidioschyzon merolae_45157
        - Galdieria sulphuraria_130081
        - Klebsormidium flaccidum_3175
        - Micromonas pusilla_564608
        - Ostreococcus lucimarinus_242159
        - Porphyridium purpureum_35688
        - Selaginella moellendorffii_88036


# ===Optional args for each BLAST step===
# Any valid command-line option can be specified (see NCBIBLAST+ docs).
# Note the hyphens (-) before option names (departing from API consistency).
# -query, -db, -out, -outfmt, -db_gencode, -query_gencode will be ignored as
# they are directly handled by forty-two itself. -max_target_seqs may be
# specified at step 'homologues' to speed up things.
blast_args:
        # TBLASTN vs banks
        homologues:
        -evalue: 1e-05
        -seg: yes
        -num_threads: 1
        -max_target_seqs: 10000
        # BLASTP vs ref banks (for transitive BRH ; actually two steps)
        references:
        -evalue: 1e-05
        -num_threads: 1
        # BLASTX vs ref banks (for transitive BRH)
        orthologues:
        -evalue: 1e-05
        -num_threads: 1
        # BLASTX vs ALI (for tax filters and alignment)
        templates:
        -evalue: 1e-05
        -seg: no
        -num_threads: 1
```

```
# ===BRH switch for assessing orthology===
# Two values are available: 'on' and 'off'.
# If set to 'on', a candidate seq must be in BRH with all reference proteomes
# to be considered as an orthologous seq. In contrast, all candidate seqs are
# considered as orthologous seqs when this parameter is set to 'off'.
# When not specified, 'ref_brh' internally defaults to 'on'.
# To limit the number of candidate seqs, use the '-max_target_seqs' option of
# the BLAST executable(s) at the 'homologues' step.
ref_brh: on


# ===Path to dir holding complete proteome BLAST databases===
# Only required when setting 'ref_brh' to 'on'.
ref_bank_dir: /media/vol2/home/mleleu/Forty-Two/OrthologousGroups/ref_banks

# ===Basenames of complete proteome BLAST databases (keyed by org name)===
# Only required when setting 'ref_brh' to 'on'.
# You can list as many databases as needed here. Only those specified as
# 'ref_orgs' below will actually be used for BRH.
ref_org_mapper:
        Arabidopsis thaliana_3702: Arabidopsis_thaliana_3702_abbr_d99.faa
        Bathycoccus prasinos_41875: Bathycoccus_prasinos_41875_abbr_d99.faa
        Chlamydomonas reinhardtii_3055: Chlamydomonas_reinhardtii_3055_abbr_d99.faa
        Chondrus crispus_2769: Chondrus_crispus_2769_abbr_d99.faa
        Coccomyxa subellipsoidea_574566: Coccomyxa_subellipsoidea_574566_abbr_d99.faa
        Cyanidioschyzon merolae_45157: Cyanidioschyzon_merolae_45157_abbr_d99.faa
        Galdieria sulphuraria_130081: Galdieria_sulphuraria_130081_abbr_d99.faa
        Klebsormidium flaccidum_3175: Klebsormidium_flaccidum_3175_abbr_d99.faa
        Micromonas pusilla_564608: Micromonas_pusilla_564608_abbr_d99.faa
        Ostreococcus lucimarinus_242159: Ostreococcus_lucimarinus_242159_abbr_d99.faa
        Porphyridium purpureum_35688: Porphyridium_purpureum_35688_abbr_d99.faa
        Selaginella moellendorffii_88036: Selaginella_moellendorffii_88036_abbr_d99.faa

# ===Orgs to be used for BRH checks===
# Only required when setting 'ref_brh' to 'on'.
# To be considered as an orthologue, a candidate seq must be in transitive BRH
# for all listed orgs (and not for only one of them). Listing more orgs thus
# increases the stringency of the BRH check. Note that 'ref_orgs' may but DO
# NOT NEED to match 'query_orgs'.
ref_orgs:
        - Arabidopsis thaliana_3702
        - Bathycoccus prasinos_41875
        - Chlamydomonas reinhardtii_3055
        - Chondrus crispus_2769
        - Coccomyxa subellipsoidea_574566
        - Cyanidioschyzon merolae_45157
        - Galdieria sulphuraria_130081
        - Klebsormidium flaccidum_3175
        - Micromonas pusilla_564608
        - Ostreococcus lucimarinus_242159
        - Porphyridium purpureum_35688
        - Selaginella moellendorffii_88036

# ===Fraction of ref_orgs to really use when assessing orthology===
# Only meaningful when setting 'ref_brh' to 'on'.
# This parameter introduces some flexibility when using reference proteomes.
# If set to a fractional value (below 1), only the best proteomes will be
# considered during BRHs. The best proteomes are those against which the
# queries have the highest average scores. This helps discarding ref_orgs that
# might hinder orthology assessment because they lack the orthologous gene(s).
# When not specified, 'reg_org_mul' internally defaults to 1.0, which is the
# strictest mode since all reference proteomes are used during BRHs.
```

```
ref_org_mul: 0.2

# ===Bit score reduction allowed when including non-1st hits among best hits===
# Only meaningful when setting 'ref_brh' to 'on'.
# This parameter applies when collecting best hits for queries to complete
# proteomes, so that close in-paralogues can all be included in the set of
# best hits. The allowed bit score reduction of any hit is expressed
# relatively to the score of the previous hit. During BRH checks, only the
# very first hit for the candidate seq is actually tested for inclusion in
# this set but for all complete proteomes. By default at most 10 hits are
# considered. To change this, use the '-max_target_seqs' option of the BLAST
# executable(s) at the 'reference' step.
# When not specified 'ref_score_mul' internally defaults to 1.0, which is the
# strictest mode since only equally-best hits are retained.
ref_score_mul: 0.99



# ===Homologues trimming switch===
# Two values are available: 'on' and 'off'.
# If set to 'on', each candidate seq is first trimmed to the range covered by
# the HSPs that retrieved it. This makes the orthology assessment more robust
# and helps exonerate to splice genes correctly. The details of this trimming
# step can be fine-tuned by editing the other trim_* parameters of this
# configuration file.
# When not specified, 'trim_homologues' internally defaults to 'on'.
trim_homologues: off



# ===Orthologues merging switch===
# Two values are available: 'on' and 'off'.
# If set to 'on', each batch of orthologous seqs from the same org is first
# fed to CAP3 in an attempt to merge some of them into contigs. Successfully
# merged orthologous seqs are identified by a trailing +N tag where N is the
# number of orthologous seqs removed in the merging process. The contig itself
# is named after the longest orthologous seq composing it.
# The details of this merging step can be fine-tuned by editing the other
# merge_* parameters of this configuration file.
# When not specified, 'merge_orthologues' internally defaults to 'off'.
merge_orthologues: off



# ===Engine to be used for aligning new seqs===
# Four values are available: 'blast', 'exonerate', 'exoblast' and 'off'.
# If the alignment engine is 'off', new seqs are added 'as is' to the ALI.
# Consequently, they will be full length but not aligned to existing seqs.
# This mode is meant for protein seqs only and thus cannot be used when adding
# transcripts from nucleotide banks.
# The 'exonerate' mode sometimes fails to align orthologous seqs due to a bug
# in exonerate executable. This causes new seqs to be lost. To automatically
# retry aligning them with BLAST in case of failure, use the 'exoblast' mode.
# When not specified, 'aligner_mode' internally defaults to 'blast'.
aligner_mode: exoblast


# ===Self-template selection switch for aligning new seqs===
# Only meaningful when setting 'aligner_mode' to a value other than 'off'.
# Two values are available: 'on' and 'off'.
# If set to 'on', closest relatives belonging to the same org as the new seqs
# will not be selected as templates, thus allowing the latter to align better.
# When not specified, 'ali_skip_self' internally defaults to 'off'.
```

```
ali_skip_self: off

# ===Coverage improvement required to consider non-1st hits as templates===
# Only meaningful when setting 'aligner_mode' to a value other than 'off'.
# This parameter applies when collecting templates for aligning the new seqs.
# Templates get considered as long as query coverage improves at least of this
# value (relatively to the previous template). The exact effect of this
# parameter depends on the 'aligner_mode' engine: 'exonerate' will try to use
# the longest template for alignment while 'blast' will use each hit in turn
# (as a fall-back with 'exoblast'). New seqs can thus be added more than once
# to the ALI (with ids *.H1.N, *.H2.N etc).
# When not specified 'ali_cover_mul' internally defaults to 1.1., which means
# that if the BLAST alignment with the second template is at least 110% of the
# BLAST alignment with the first template, both templates are retained.
ali_cover_mul: 1.1

# ===Preservation switch for '#NEW#' tags from preexisting sequences===
# Two values are available: 'on' and 'off'.
# If set to 'on' (default), #NEW# tags will be preserved. Note that
# preexisting new sequences are invisible to 42 (they cannot be used as
# queries etc).
ali_keep_old_new_tags: off

# ===Action to perform when a preexisting lengthened seq is identified===
# Currently, two values are available: 'remove' and 'keep'.
# The option is quite self-explanatory. It is useful when one runs 42 multiple
# times on the sames ALIs to repeatedly enrich the same orgs, assuming that
# org banks are updated between runs.
# When not specified, 'ali_keep_lengthened_seqs' internally defaults to
# 'keep'.
ali_keep_lengthened_seqs: keep


# ===Taxonomic report switch===
# Two values are available: 'on' and 'off'.
# If set to 'on', the lineage of new seqs is inferred by analyzing the
# taxonomy of their ALI closest relatives and one '.tax-report' file is
# generated for each ALI processed (see 'run_mode' above).
# The details of this taxonomic analysis can be fine-tuned by editing the
# other tax_* parameters of this configuration file.
# When not specified, 'tax_reports' internally defaults to 'off'. Yet, the
# YAML generator automatically sets it to 'on' if 'run_mode' is 'metagenomic'.
tax_reports: on


# ===Path to dir holding transcript BLAST databases===
bank_dir: /media/vol2/home/mleleu/Monophylie-Chl/OF57-IqTree/42-genomes/genomes

# ===Default args applying to all orgs unless otherwise specified===
# Some of these args can be thus specified on a per-org basis below if needed.
# This especially makes sense for 'code'.
defaults:
        # ===Genetic code for translated BLAST searches===
        # When not specified 'code' internally defaults to 1 (standard).
        # See ftp://ftp.ncbi.nih.gov/entrez/misc/data/gc.prt for other codes.
        code: 1

# ===Org-specific args===
# The only mandatory args are 'org' and 'banks'. All other args are taken from
# the 'defaults:' section described above.
# This part can be concatenated on a per-run basis to the previous part, which
# would be the same for several runs. In the future, forty-two might support
# two different configuration files to reflect this conceptual distinction.
orgs:
```

```
 - org: Cyanoptyche sp._77921
          banks:
          - Cyanoptyche_sp._MMETSP1086_77921_d95_abbr

 - org: Kappaphycus alvarezii_GCA_002205965.2
          banks:
          - GCA_002205965.2_ASM220596v2_genomic.fna

 - org: Gracilariopsis lemaneiformis_GCA_003346895.1
          banks:
          - GCA_003346895.1_Glem_v01_genomic.fna

 - org: Mesostigma viride_GCA_004000685.1
          banks:
          - GCA_004000685.1_MeVI296_assembly3_genomic.fna


#
# This config file has been generated automatically on 15:05:51 30-Nov-2020.
# We advise not to modify directly this file manually but rather to modify the
# yaml-generator command instead for traceability and reproducibility.
#
#yaml-generator-42.pl --run_mode=phylogenomic --out_suffix=-g \
#--queries /media/vol2/home/mleleu/Forty-Two/OrthologousGroups/ref_banks/queries.idl \
#--evalue=1e-05 --homologues_seg=yes --max_target_seqs=10000 --templates_seg=no \
#--bank_dir /media/vol2/home/mleleu/Monophylie-Chl/OF57-IqTree/42-genomes/genomes --bank_suffix=.nsq
--bank_mapper
/media/vol2/home/mleleu/Monophylie-Chl/OF57-IqTree/42-genomes/genomes/iqtree-genomes-bank-mapper.id
m \
#--ref_brh=on --ref_bank_dir /media/vol2/home/mleleu/Forty-Two/OrthologousGroups/ref_banks
--ref_bank_suffix=.psq --ref_bank_mapper
/media/vol2/home/mleleu/Forty-Two/OrthologousGroups/ref_banks/ref_bank_mapper-inverse.idm \
#--ref_org_mul=0.2 --ref_score_mul=0.99 \
#--trim_homologues=off \
#--merge_orthologues=off \
#--aligner_mode=exoblast --ali_skip_self=off --ali_cover_mul=1.1 --ali_keep_old_new_tags=off
--ali_keep_lengthened_seqs=keep \
#--tax_reports=on \
#--tax_min_score=0 --tax_score_mul= --tax_min_ident=0 --tax_min_len=0 \
#--tol_check=off
```

Appendix 7: Summary table of the potential LGT between Chlamydia and Archaeplastida selected by different methods. This table represents all orthologous groups (OG) selected by our methods presenting a potential LGT between Chlamydia and Archaeplastida. 1 = Selected, 0 = not selected. semi-auto = semi-automatic pipeline , Man-2-3p = confirmed by 2 or 3 observers in the manual analysis of the trees , Man-1p = confirmed by only 1 observer in the manual analysis of the trees. auto = automatic pipeline. Literature = identified in previous studies (Ball et al., 2013; Becker et al., 2008; Cenci et al., 2018, 2017; Huang and Gogarten, 2007; Moustafa et al., 2008). OG are ranked from most to least convincing LGT in a MATH context.

Pipeline Chlamydien - Sélections					
OG	semi-auto	Man-2-3p	Man-1p	auto	Litterature
OG0000904	1	1	0	1	1
OG0000674	1	1	0	1	1
OG0000913	1	1	0	1	1
OG0001059	1	1	0	1	1
OG0001293	1	1	0	1	1
OG0001851	1	1	0	1	1
OG0002498	1	1	0	1	1
OG0002584	1	1	0	1	1
OG0004493	1	1	0	1	1
OG0004954	1	1	0	1	1
OG0005232	1	1	0	1	1
OG0005255	1	1	0	1	1
OG0005374	1	1	0	1	1
OG0005382	1	1	0	1	1
OG0005581	1	1	0	1	1
OG0007168	1	1	0	1	1
OG0008425	1	1	0	1	1
OG0008974	1	1	0	1	1
OG0024221	1	1	0	1	1
OG0001078	1	1	0	1	0
OG0008763	1	1	0	1	0
OG0001468	1	0	1	1	1
OG0000013	1	0	1	1	1
OG0017872	1	0	1	1	1
OG0002300	1	0	1	1	0
OG0008957	1	0	1	1	0
OG0003312	0	0	0	1	1
OG0003961	0	0	0	1	1
OG0005053	0	0	0	1	1
OG0000812	0	0	0	1	1
OG0000134	0	0	0	1	1
OG0000649	0	0	0	1	1
OG0002222	0	0	0	1	1
OG0002395	0	0	0	1	1
OG0002591	0	0	0	1	1
OG0003272	0	0	0	1	1
OG0003449	0	0	0	1	1

OG0003873	0	0	0	1	1
OG0004746	0	0	0	1	1
OG0004766	0	0	0	1	1
OG0005231	0	0	0	1	1
OG0006000	0	0	0	1	1
OG0009869	0	0	0	1	1
OG0014617	0	0	0	1	1
OG0028045	0	0	0	1	1
OG0001950	0	0	0	1	0
OG0005097	0	0	0	1	0
OG0001410	0	0	0	1	0
OG0000479	0	0	0	1	0
OG0001000	0	0	0	1	0
OG0002167	0	0	0	1	0
OG0003309	0	0	0	1	0
OG0003383	0	0	0	1	0
OG0004281	0	0	0	1	0
OG0004382	0	0	0	1	0
OG0005308	0	0	0	1	0
OG0017499	0	0	0	1	0
OG0000254	1	1	0	0	1
OG0000691	1	1	0	0	1
OG0001105	1	1	0	0	0
OG0007172	1	1	0	0	0
OG0007527	1	1	0	0	0
OG0000627	1	0	1	0	1
OG0000049	1	0	1	0	1
OG0003741	1	0	1	0	1
OG0006758	1	0	1	0	1
OG0011723	1	0	1	0	1
OG0000021	1	0	1	0	0
OG0000603	1	0	1	0	0
OG0004322	1	0	1	0	0
OG0000576	1	0	0	0	1
OG0000894	1	0	0	0	1
OG0004082	1	0	0	0	1
OG0000125	0	0	0	0	1
OG0000521	0	0	0	0	1
OG0000712	0	0	0	0	1
OG0000849	0	0	0	0	1
OG0000857	0	0	0	0	1
OG0000971	0	0	0	0	1
OG0001148	0	0	0	0	1
OG0001287	0	0	0	0	1
OG0001777	0	0	0	0	1
OG0002686	0	0	0	0	1

OG0003114	0	0	0	0	1
OG0005215	0	0	0	0	1
OG0009120	0	0	0	0	1
OG0000944	0	0	0	0	1
OG0000035	0	0	0	0	1
OG0000057	0	0	0	0	1
OG0000065	0	0	0	0	1
OG0000426	0	0	0	0	1
OG0000685	0	0	0	0	1
OG0000752	0	0	0	0	1
OG0003397	0	0	0	0	1
OG0004170	0	0	0	0	1
OG0014950	0	0	0	0	1
OG0000010	0	0	0	0	1
OG0000016	0	0	0	0	1
OG0000018	0	0	0	0	1
OG0000069	0	0	0	0	1
OG0000078	0	0	0	0	1
OG0000104	0	0	0	0	1
OG0000107	0	0	0	0	1
OG0000113	0	0	0	0	1
OG0000152	0	0	0	0	1
OG0000193	0	0	0	0	1
OG0000233	0	0	0	0	1
OG0000399	0	0	0	0	1
OG0000413	0	0	0	0	1
OG0000436	0	0	0	0	1
OG0000561	0	0	0	0	1
OG0000595	0	0	0	0	1
OG0000601	0	0	0	0	1
OG0000606	0	0	0	0	1
OG0000615	0	0	0	0	1
OG0000624	0	0	0	0	1
OG0000644	0	0	0	0	1
OG0000722	0	0	0	0	1
OG0000742	0	0	0	0	1
OG0000755	0	0	0	0	1
OG0000806	0	0	0	0	1
OG0000867	0	0	0	0	1
OG0000906	0	0	0	0	1
OG0000916	0	0	0	0	1
OG0000937	0	0	0	0	1
OG0000970	0	0	0	0	1
OG0001021	0	0	0	0	1
OG0001103	0	0	0	0	1
OG0001146	0	0	0	0	1

OG0001174	0	0	0	0	1
OG0001178	0	0	0	0	1
OG0001192	0	0	0	0	1
OG0001373	0	0	0	0	1
OG0001442	0	0	0	0	1
OG0001712	0	0	0	0	1
OG0002088	0	0	0	0	1
OG0002313	0	0	0	0	1
OG0002860	0	0	0	0	1
OG0003031	0	0	0	0	1
OG0004008	0	0	0	0	1
OG0004316	0	0	0	0	1
OG0004595	0	0	0	0	1
OG0006414	0	0	0	0	1
OG0009625	0	0	0	0	1
OG0010637	0	0	0	0	1
OG0015224	0	0	0	0	1
OG0015752	0	0	0	0	1
OG0000681	1	0	0	0	0
OG0002959	1	0	0	0	0
OG0000221	1	0	0	0	0
OG0000447	1	0	0	0	0
OG0000985	1	0	0	0	0
OG0009586	1	0	0	0	0
OG0000654	1	0	0	0	0
OG0001134	1	0	0	0	0
OG0001289	1	0	0	0	0
OG0003133	1	0	0	0	0
OG0003870	1	0	0	0	0
OG0011552	1	0	0	0	0

Appendix 8: Summary table of the functional annotation of all trees selected by the bacterial automatic pipeline. The annotation of each selected clan, performed by eggNOG mapper and BlastKOALA, was generalized if at least 50% of the sequences were annotated in the same way. The tropism of each clan is indicated in the second column. R: Rhodophyta, V: Viridiplantae, G: Glaucophyta.

Pipeline Chlamydia				
OG	**Tropisme**	**Identification protéine**	**N°Kegg**	**Fonction**
OG0000013	RVG	NLRC3, NOD3; NLR family CARD domain-containing protein 3	K22614	Signaling and cellular processes
OG0000134	R	aqpZ; aquaporin Z	K06188	Transporters
OG0000479	RVG	aroDE, DHQ-SDH; 3-dehydroquinate dehydratase / shikimate dehydrogenase [EC:4.2.1.10 1.1.1.25]	K13832	Amino acid metabolism
OG0000649	VG	ksgA; 16S rRNA (adenine1518-N6/adenine1519-N6)-dimethyltransferase [EC:2.1.1.182]	K02528	Ribosome biogenesis
OG0000674	V	TC.PIT; inorganic phosphate transporter, PiT family	K03306	Transporters
OG0000812	VG	K07146; UPF0176 protein	K07146	
OG0000904	V	rluB; 23S rRNA pseudouridine2605 synthase [EC:5.4.99.22]	K06178	Ribosome biogenesis
OG0000913	RVG	fabF, OXSM, CEM1; 3-oxoacyl-[acyl-carrier-protein] synthase II [EC:2.3.1.179]	K09458	Lipid metabolism - Metabolism of cofactors and vitamins
OG0001000	V	NSF, SEC18; vesicle-fusing ATPase [EC:3.6.4.6]	K06027	Membrane trafficking
OG0001059	V	E1.1.1.82; malate dehydrogenase (NADP+) [EC:1.1.1.82]	K00024	Carbohydrate metabolism - Energy metabolism - Amino acid metabolism
OG0001078	RV	ribBA; 3,4-dihydroxy 2-butanone 4-phosphate synthase / GTP cyclohydrolase II [EC:4.1.99.12 3.5.4.25]	K14652	Metabolism of cofactors and vitamins
OG0001293	VG	glgA; starch synthase [EC:2.4.1.21]	K00703	Carbohydrate metabolism
OG0001410	RVG	RP-L24, MRPL24, rplX; large subunit ribosomal protein L24	K02895	Ribosome
OG0001468	R	trpC; indole-3-glycerol phosphate synthase [EC:4.1.1.48]	K01609	Amino acid metabolism
OG0001851	RV	mraW, rsmH; 16S rRNA (cytosine1402-N4)-methyltransferase [EC:2.1.1.199]	K03438	Ribosome biogenesis
OG0001950	RV	trxB, TRR; thioredoxin reductase (NADPH) [EC:1.8.1.9]	K00384	Metabolism of other amino acids
OG0002167	RVG	PHYH; phytanoyl-CoA hydroxylase [EC:1.14.11.18]	K00477	Peroxisome
OG0002222	RVG	SAL; 3'(2'), 5'-bisphosphate nucleotidase / inositol polyphosphate 1-phosphatase [EC:3.1.3.7 3.1.3.57]	K15422	Carbohydrate metabolism - Energy metabolism
OG0002300	R	PTH1, pth, spoVC; peptidyl-tRNA hydrolase, PTH1 family [EC:3.1.1.29]	K01056	Translation factors
OG0002395	RVG	truA, PUS1; tRNA pseudouridine38-40 synthase [EC:5.4.99.12]	K06173	Transfer RNA biogenesis
OG0002498	RV	YARS, tyrS; tyrosyl-tRNA synthetase [EC:6.1.1.1]	K01866	Transfer RNA biogenesis
OG0002584	RV	pnp, PNPT1; polyribonucleotide nucleotidyltransferase [EC:2.7.7.8]	K00962	Messenger RNA biogenesis
OG0002591	RVG			
OG0003272	RVG	trpD; anthranilate phosphoribosyltransferase [EC:2.4.2.18]	K00766	Amino acid metabolism
OG0003309	RG	rnhB; ribonuclease HII [EC:3.1.26.4]	K03470	DNA replication proteins
OG0003312	RVG	tyrP; tyrosine-specific transport protein	K03834	Transporters
OG0003383	RV	dicarboxylate transporter - 2-oxoglutarate/malate transporter	K03319	Transporters
OG0003449	RVG	TC.AAA; ATP:ADP antiporter, AAA family	K03301	Transporters
OG0003873	RVG	ISA, treX; isoamylase [EC:3.2.1.68]	K01214	Carbohydrate metabolism
OG0003961	V	fabI; enoyl-[acyl-carrier protein] reductase I [EC:1.3.1.9 1.3.1.10]	K00208	Lipid metabolism - Metabolism of cofactors and vitamins
OG0004281	RV	K09858; SEC-C motif domain protein	K09858	
OG0004382	RVG			
OG0004493	RVG	Na H antiporter		Transporters
OG0004746	RVG	uhpC; MFS transporter, OPA family, sugar phosphate sensor protein UhpC	K07783	Transporters
OG0004766	RVG	ispE; 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase [EC:2.7.1.148]	K00919	Metabolism of terpenoids and polyketides
OG0004954	RVG			
OG0005053	V	gcpE, ispG; (E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase [EC:1.17.7.1 1.17.7.3]	K03526	Metabolism of terpenoids and polyketides
OG0005097	RVG			
OG0005231	RVG	E2.6.1.83; LL-diaminopimelate aminotransferase [EC:2.6.1.83]	K10206	Amino acid metabolism
OG0005232	RVG	ispD; 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase [EC:2.7.7.60]	K00991	Metabolism of terpenoids and polyketides
OG0005255	RVG			
OG0005308	RV	CHS; chalcone synthase [EC:2.3.1.74]	K00660	Biosynthesis of other secondary metabolites
OG0005374	R	rlmH; 23S rRNA (pseudouridine1915-N3)-methyltransferase [EC:2.1.1.177]	K00783	Ribosome biogenesis
OG0005382	RVG		K03215	Ribosome biogenesis
OG0005581	RVG	ATS1; glycerol-3-phosphate O-acyltransferase [EC:2.3.1.15]	K00630	Lipid metabolism
OG0006000	V	kdsB; 3-deoxy-manno-octulosonate cytidylyltransferase (CMP-KDO synthetase) [EC:2.7.7.38]	K00979	Glycan biosynthesis and metabolism
OG0007168	RVG			
OG0008425	VG	UGP3; UTP---glucose-1-phosphate uridylyltransferase [EC:2.7.7.9]	K22920	Lipid metabolism
OG0008763	V			
OG0008957	V			
OG0008974	RV	ddl; D-alanine-D-alanine ligase [EC:6.3.2.4]	K01921	Metabolism of other amino acids - Glycan biosynthesis and metabolism
OG0009869	RG	apbE; FAD:protein FMN transferase [EC:2.7.1.180]	K03734	
OG0014617	V	dnaQ; DNA polymerase III subunit epsilon [EC:2.7.7.7]	K02342	DNA replication proteins
OG0017499	RG	wbpA; UDP-N-acetyl-D-glucosamine dehydrogenase [EC:1.1.1.136]	K13015	Carbohydrate metabolism - Glycan biosynthesis and metabolism
OG0017872	R		K01046	Glycerolipid metabolism

OG0024221	V	murB; UDP-N-acetylmuramate dehydrogenase [EC:1.3.1.98]	K00075	Carbohydrate metabolism - Glycan biosynthesis and metabolism
OG0028045	RG	queD, ptpS, PTS; 6-pyruvoyltetrahydropterin/6-carboxytetrahydropterin synthase [EC:4.2.3.12 4.1.2.50]	K01737	Metabolism of cofactors and vitamins

Pipeline Bacteroidetes

OG	Tropisme	Identification protéine	N°Kegg	Fonction
OG0000010	RV			
OG0000118	RVG			
OG0000574	RVG			
OG0000610	RV	msrB; peptide-methionine (R)-S-oxide reductase [EC:1.8.4.12]	K07305	
OG0000634	V	engB; GTP-binding protein	K03978	genetic information processing
OG0000928	V			
OG0001033	RVG	plsC; 1-acyl-sn-glycerol-3-phosphate acyltransferase [EC:2.3.1.51]	K00655	Lipid metabolism
OG0001445	V	aspA; aspartate ammonia-lyase [EC:4.3.1.1]	K01744	Amino acid metabolism
OG0001493	RV	TRMU, SLM3; tRNA-5-taurinomethyluridine 2-sulfurtransferase [EC:2.8.1.14]	K21027	genetic information processing
OG0001651	V	era, ERAL1; GTPase	K03595	genetic information processing
OG0001696	V	truB, PUS4, TRUB1; tRNA pseudouridine55 synthase [EC:5.4.99.25]	K03177	genetic information processing
OG0001774	RV	hemC, HMBS; hydroxymethylbilane synthase [EC:2.5.1.61]	K01749	Metabolism of cofactors and vitamins
OG0001893	V			
OG0001937	RVG			
OG0002073	RV			
OG0002300	R	PTH1, pth, spoVC; peptidyl-tRNA hydrolase, PTH1 family [EC:3.1.1.29]	K01056	genetic information processing
OG0002448	V			
OG0002628	V	nadC, QPRT; nicotinate-nucleotide pyrophosphorylase (carboxylating) [EC:2.4.2.19]	K00767	Metabolism of cofactors and vitamins
OG0003226	RVG			
OG0003314	RV	AMY, amyA, malS; alpha-amylase [EC:3.2.1.1]	K01176	Carbohydrate metabolism
OG0003315	V	FUCA; alpha-L-fucosidase [EC:3.2.1.51]	K01206	Glycan biosynthesis and metabolism - signaling and cellular processes
OG0003969	V	FAB2, SSI2, desA1; acyl-[acyl-carrier-protein] desaturase [EC:1.14.19.2 1.14.19.11 1.14.19.26]	K03921	Lipid metabolism
OG0004493	RVG			
OG0005051	RV	fabZ; 3-hydroxyacyl-[acyl-carrier-protein] dehydratase [EC:4.2.1.59]	K02372	Lipid metabolism - Metabolism of cofactors and vitamins
OG0005056	RV	polA; DNA polymerase I [EC:2.7.7.7]	K02335	genetic information processing
OG0005129	V	cutC; copper homeostasis protein	K06201	
OG0005228	RV	DVR; divinyl chlorophyllide a 8-vinyl-reductase [EC:1.3.1.75]	K19073	Metabolism of cofactors and vitamins
OG0005374	VG	rlmH; 23S rRNA (pseudouridine1915-N3)-methyltransferase [EC:2.1.1.177]	K00783	genetic information processing
OG0005785	RVG	malQ; 4-alpha-glucanotransferase [EC:2.4.1.25]	K00705	Carbohydrate metabolism
OG0005794	V	FATA; fatty acyl-ACP thioesterase A [EC:3.1.2.14]	K10782	Lipid metabolism
OG0006048	RVG			
OG0006163	RVG	K07137; uncharacterized protein	K07137	
OG0006273	VG	cph1; two-component system, chemotaxis family, sensor kinase Cph1 [EC:2.7.13.3]	K11354	signaling and cellular processes
OG0007955	VG	crtZ; beta-carotene 3-hydroxylase [EC:1.14.15.24]	K15746	Metabolism of terpenoids and polyketides
OG0008343	RG			
OG0008372	R	queA; S-adenosylmethionine:tRNA ribosyltransferase-isomerase [EC:2.4.99.17]	K07568	genetic information processing
OG0008391	V	ybgC; acyl-CoA thioester hydrolase [EC:3.1.2.-]	K07107	
OG0008499	RV	pepD; dipeptidase D [EC:3.4.13.-]	K01270	Metabolism of other amino acids
OG0008718	V	rng, cafA; ribonuclease G [EC:3.1.26.-]	K08301	genetic information processing
OG0009601	RV	purM; phosphoribosylformylglycinamidine cyclo-ligase [EC:6.3.3.1]	K01933	Nucleotide metabolism
OG0009615	V	murG; UDP-N-acetylglucosamine--N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase [EC:2.4.1.227]	K02563	Glycan biosynthesis and metabolism
OG0010637	RVG	rumA; 23S rRNA (uracil1939-C5)-methyltransferase [EC:2.1.1.190]	K03215	genetic information processing
OG0011258	VG			
OG0017013	V	E4.4.1.11; methionine-gamma-lyase [EC:4.4.1.11]	K01761	Amino acid metabolism - Metabolism of other amino acids

Pipeline Proteobacteria

OG	Tropisme	Identification protéine	N°Kegg	Fonction
OG0000574	RVG			
OG0000615	V	CARS, cysS; cysteinyl-tRNA synthetase [EC:6.1.1.16]	K01883	genetic information processing
OG0000637	R	RP-L14, MRPL14, rplN; large subunit ribosomal protein L14	K02874	genetic information processing
OG0000704	V	PGK, pgk; phosphoglycerate kinase [EC:2.7.2.3]	K00927	Carbohydrate metabolism - Energy metabolism - Exosome
OG0000852	RVG	HIBCH; 3-hydroxyisobutyryl-CoA hydrolase [EC:3.1.2.4]	K05605	Carbohydrate metabolism - Amino acid metabolism
OG0001118	VG	relA; GTP pyrophosphokinase [EC:2.7.6.5]	K00951	Nucleotide metabolism

OG0001135	RV	ridA, tdcF, RIDA; 2-iminobutanoate/2-iminopropanoate deaminase [EC:3.5.99.10]	K09022	
OG0001282	RVG	ISCA1; iron-sulfur cluster assembly 1	K22063	genetic information processing
OG0001293	RVG	glgA; starch synthase [EC:2.4.1.21]	K00703	Carbohydrate metabolism
OG0001349	RVG	pgsA, PGS1; CDP-diacylglycerol---glycerol-3-phosphate 3-phosphatidyltransferase [EC:2.7.8.5]	K00995	Lipid metabolism
OG0001387	R	NFU1, HIRIP5; NFU1 iron-sulfur cluster scaffold homolog, mitochondrial	K22074	genetic information processing
OG0001473	V	E2.2.1.2, talA, talB; transaldolase [EC:2.2.1.2]	K00616	Carbohydrate metabolism
OG0001961	RVG	MTFMT, fmt; methionyl-tRNA formyltransferase [EC:2.1.2.9]	K00604	Metabolism of cofactors and vitamins
OG0002061	V	ATPeF0C, ATP5G, ATP9; F-type H+-transporting ATPase subunit c	K02128	Energy metabolism
OG0002095	V	guaA, GMPS; GMP synthase (glutamine-hydrolysing) [EC:6.3.5.2]	K01951	Nucleotide metabolism
OG0002258	RVG	cysE; serine O-acetyltransferase [EC:2.3.1.30]	K00640	Energy metabolism - Amino acid metabolism
OG0002349	RVG	E2.5.1.54, aroF, aroG, aroH; 3-deoxy-7-phosphoheptulonate synthase [EC:2.5.1.54]	K01626	Amino acid metabolism
OG0002664	RVG	RP-L33, MRPL33, rpmG; large subunit ribosomal protein L33	K02913	genetic information processing
OG0002722	V	argG, ASS1; argininosuccinate synthase [EC:6.3.4.5]	K01940	Amino acid metabolism - Exosome
OG0003130	RVG	NDUFS4; NADH dehydrogenase (ubiquinone) Fe-S protein 4	K03937	Energy metabolism
OG0003133	V	ribE, RIB5; riboflavin synthase [EC:2.5.1.9]	K00793	Metabolism of cofactors and vitamins
OG0003243	RVG	pepN; aminopeptidase N [EC:3.4.11.2]	K01256	Metabolism of other amino acids
OG0003312	RVG	tyrP; tyrosine-specific transport protein	K03834	Transporters
OG0003449	RVG	TC.AAA; ATP:ADP antiporter, AAA family	K03301	Transporters
OG0003649	RV	purB, ADSL; adenylosuccinate lyase [EC:4.3.2.2]	K01756	Nucleotide metabolism - Amino acid metabolism
OG0003970	RVG	dxs; 1-deoxy-D-xylulose-5-phosphate synthase [EC:2.2.1.7]	K01662	Metabolism of cofactors and vitamins - Metabolism of terpenoids and polyketides
OG0004247	RVG			
OG0004267	RVG	prmA; ribosomal protein L11 methyltransferase [EC:2.1.1.-]	K02687	genetic information processing
OG0004720	V			
OG0004900	V	topA; DNA topoisomerase I [EC:5.6.2.1]	K03168	genetic information processing
OG0005026	V	trmA; tRNA (uracil-5-)-methyltransferase [EC:2.1.1.35]	K00557	genetic information processing
OG0007156	V	K09919; uncharacterized protein	K09919	
OG0008000	RV	ycfD; 50S ribosomal protein L16 3-hydroxylase [EC:1.14.11.47]	K18850	genetic information processing
OG0008420	V	gshA; glutamate--cysteine ligase [EC:6.3.2.2]	K01919	Amino acid metabolism
OG0009004	V	murE; UDP-N-acetylmuramoyl-L-alanyl-D-glutamate--2,6-diaminopimelate ligase [EC:6.3.2.13]	K01928	Amino acid metabolism - Glycan biosynthesis and metabolism
OG0010452	VG			
OG0011797	RV			
OG0017102	V	RP-S6, MRPS6, rpsF; small subunit ribosomal protein S6	K02990	genetic information processing
OG0024207	V	mltB; membrane-bound lytic murein transglycosylase B [EC:4.2.2.-]	K08305	

Pipeline Actinobacteria

OG	Tropisme	Identification protéine	N°Kegg	Fonction
OG0001118	RVG	relA; GTP pyrophosphokinase	K00951	Nucleotid metabolism
OG0003160	RVG	terC; tellurite resistance protein TerC	K05794	
OG0007174	RVG	RIBF; FAD synthetase	K22949	Metabolism of cofactors and vitamins
OG0007944	V	malQ; 4-alpha-glucanotransferase	K00705	Carbohydrate metabolism
OG0014591	RVG	trpB; tryptophan synthase beta chain	K06001	Amino acid metabolism

Pipeline Firmicutes

OG	Tropisme	Identification protéine	N°Kegg	Fonction
OG0000574	RVG			
OG0002787	V	pel; pectate lyase	K01728	Carbohydrate metabolism
OG0003701	RVG			
OG0003702	RV			
OG0004076	RV	gidB, rsmG; 16S rRNA (guanine527-N7)-methyltransferase	K03501	
OG0006278	V	MTN; 5'-methylthioadenosine nucleosidase	K01244	Amino acid metabolism
OG0008532	RVG			
OG0010015	V			
OG0017847	RV			

Appendix 9: Summary table of all species selected for our study. x = the species is entering the corresponding pipeline.

Species	Taxonomique group	Pipeline Chlamydia	Pipeline Bacteroidetes	Pipeline Proteobacteria	Pipeline Actinobacteria	Pipeline Firmicutes	Pipeline Cyanobacteria	Pipeline Amoebozoa	Pipeline Fungi	Pipeline Glaucophyte
Cryptosporidium parvum_353152	Alveolata	x	x	x	x	x	x	x	x	x
Perkinsus marinus_423536	Alveolata	x	x	x	x	x	x	x	x	x
Stylonychia lemnae_755200	Alveolata	x	x	x	x	x	x	x	x	x
Symbiodinium minutum_1202447	Alveolata	x	x	x	x	x	x	x	x	x
Tetrahymena thermophila_5911	Alveolata	x	x	x	x	x	x	x	x	x
Toxoplasma gondii_5811	Alveolata	x	x	x	x	x	x	x	x	x
Vitrella brassicaformis_1169540	Alveolata	x	x	x	x	x	x	x	x	x
Acanthamoeba castellanii_5755	Amoebozoa	x	x	x	x	x	x	x	x	x
Dictyostelium discoideum_44689	Amoebozoa	x	x	x	x	x	x	x	x	x
Polysphondylium pallidum_13642	Amoebozoa	x	x	x	x	x	x	x	x	x
Bodo saltans_75058	Euglenozoa	x	x	x	x	x	x	x	x	x
Leishmania infantum_435258	Euglenozoa	x	x	x	x	x	x	x	x	x
Trypanosoma brucei_679716	Euglenozoa	x	x	x	x	x	x	x	x	x
Cyanophora paradoxa	Glaucocystophyceae	x	x	x	x	x	x	x	x	x
Cyanoptyche sp._77921	Glaucocystophyceae	x	x	x	x	x	x	x	x	x
Agaricus bisporus_5341	Opisthokonta	x	x	x	x	x	x	x	x	x
Allomyces macrogynus_28583	Opisthokonta	x	x	x	x	x	x	x	x	x
Aspergillus niger_5061	Opisthokonta	x	x	x	x	x	x	x	x	x
Batrachochytrium dendrobatidis_684364	Opisthokonta	x	x	x	x	x	x	x	x	x
Capsaspora owczarzaki_595528	Opisthokonta	x	x	x	x	x	x	x	x	x
Fonticula alba_691883	Opisthokonta	x	x	x	x	x	x	x	x	x
Homo sapiens_9606	Opisthokonta	x	x	x	x	x	x	x	x	x
Komagataella phaffii_460519	Opisthokonta	x	x	x	x	x	x	x	x	x
Lottia spp_72691	Opisthokonta	x	x	x	x	x	x	x	x	x
Monosiga brevicollis_81824	Opisthokonta	x	x	x	x	x	x	x	x	x
Mucor circinelloides_747725	Opisthokonta	x	x	x	x	x	x	x	x	x
Nematostella spp_45350	Opisthokonta	x	x	x	x	x	x	x	x	x
Puccinia graminis_418459	Opisthokonta	x	x	x	x	x	x	x	x	x
Rhizophagus irregularis_588596	Opisthokonta	x	x	x	x	x	x	x	x	x
Rozella allomycis_281847	Opisthokonta	x	x	x	x	x	x	x	x	x
Saccharomyces cerevisiae_559292	Opisthokonta	x	x	x	x	x	x	x	x	x
Trichoplax adhaerens_10228	Opisthokonta	x	x	x	x	x	x	x	x	x
Ustilago maydis_5270	Opisthokonta	x	x	x	x	x	x	x	x	x
Chondrus crispus_2769	Rhodophyta	x	x	x	x	x	x	x	x	x
Cyanidioschyzon merolae_45157	Rhodophyta	x	x	x	x	x	x	x	x	x
Galdieria sulphuraria_130081	Rhodophyta	x	x	x	x	x	x	x	x	x
Porphyridium purpureum_35688	Rhodophyta	x	x	x	x	x	x	x	x	x
Kappaphycus alvarezii_GCA_002205965.2	Rhodophyta	x	x	x	x	x	x	x	x	x
Gracilariopsis lemaneiformis_GCA_003346895.1	Rhodophyta	x	x	x	x	x	x	x	x	x
Porphyra umbilicalis_GCA_002049455.2	Rhodophyta	x	x	x	x	x	x	x	x	x
Gracilariopsis chorda_GCA_003194525.1	Rhodophyta	x	x	x	x	x	x	x	x	x
Rhodelphis marinus	Rhodophyta	x	x	x	x	x	x	x	x	x
Rhodelphis limneticus	Rhodophyta	x	x	x	x	x	x	x	x	x
Albugo candida_65357	Stramenopiles	x	x	x	x	x	x	x	x	x
Aplanochytrium kerguelense_87111	Stramenopiles	x	x	x	x	x	x	x	x	x
Aurantiochytrium limacinum_717989	Stramenopiles	x	x	x	x	x	x	x	x	x
Aureococcus anophagefferens_44056	Stramenopiles	x	x	x	x	x	x	x	x	x
Blastocystis sp_1529846	Stramenopiles	x	x	x	x	x	x	x	x	x
Ectocarpus siliculosus_2880	Stramenopiles	x	x	x	x	x	x	x	x	x
Fragilariopsis cylindrus_186039	Stramenopiles	x	x	x	x	x	x	x	x	x
Phaeodactylum tricornutum_2850	Stramenopiles	x	x	x	x	x	x	x	x	x
Phytophthora infestans_4787	Stramenopiles	x	x	x	x	x	x	x	x	x
Pseudo-nitzschia multiseries_37319	Stramenopiles	x	x	x	x	x	x	x	x	x
Pythium ultimum_65071	Stramenopiles	x	x	x	x	x	x	x	x	x
Saprolegnia parasitica_695850	Stramenopiles	x	x	x	x	x	x	x	x	x
Schizochytrium aggregatum_876976	Stramenopiles	x	x	x	x	x	x	x	x	x
Thalassiosira pseudonana_35128	Stramenopiles	x	x	x	x	x	x	x	x	x
Arabidopsis thaliana_3702	Viridiplantae	x	x	x	x	x	x	x	x	x
Bathycoccus prasinos_41875	Viridiplantae	x	x	x	x	x	x	x	x	x
Chlamydomonas reinhardtii_3055	Viridiplantae	x	x	x	x	x	x	x	x	x
Coccomyxa subellipsoidea_574566	Viridiplantae	x	x	x	x	x	x	x	x	x
Klebsormidium flaccidum_3175	Viridiplantae	x	x	x	x	x	x	x	x	x
Micromonas pusilla_564608	Viridiplantae	x	x	x	x	x	x	x	x	x
Ostreococcus lucimarinus_242159	Viridiplantae	x	x	x	x	x	x	x	x	x
Selaginella moellendorffii_88036	Viridiplantae	x	x	x	x	x	x	x	x	x
Mesostigma viride_GCA_004000685.1	Viridiplantae	x	x	x	x	x	x	x	x	x
Physcomitrella patens_GCA_000002425.2	Viridiplantae	x	x	x	x	x	x	x	x	x
Populus trichocarpa_GCA_000002775.3	Viridiplantae	x	x	x	x	x	x	x	x	x
Micromonas commoda_GCA_000090985.2	Viridiplantae	x	x	x	x	x	x	x	x	x
Volvox carteri_GCA_000143455.1	Viridiplantae	x	x	x	x	x	x	x	x	x
Chlorella variabilis_GCA_000147415.1	Viridiplantae	x	x	x	x	x	x	x	x	x
Ostreococcus tauri_GCA_000214015.2	Viridiplantae	x	x	x	x	x	x	x	x	x
Gloeobacter violaceus_GCA_000011385.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Pseudanabaena biceps_GCA_000332215.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Prochlorococcus marinus_GCF_000007925.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Prochlorococcus marinus_GCF_000011465.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Prochlorococcus marinus_GCF_000011485.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Synechococcus sp._GCF_000012505.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Synechococcus sp._GCF_000013205.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Synechococcus sp._GCF_000014585.1	Cyanobacteria	x	x	x	x	x	x	x	x	x

Prochlorococcus marinus_GCF_000015645.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Prochlorococcus marinus_GCF_000018585.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Atelocyanobacterium thalassa_GCF_000025125.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Synechococcus sp._GCF_000063505.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Synechococcus sp._GCF_000153285.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Synechococcus sp._GCF_000153805.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Synechococcus sp._GCF_000153825.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Synechococcus sp._GCF_000155595.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Synechococcus sp._GCF_000161795.2	Cyanobacteria	x	x	x	x	x	x	x	x	x
Synechococcus sp._GCF_000316685.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Geitlerinema sp._GCF_000317045.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Dactylococcopsis salina_GCF_000317615.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Synechococcus sp._GCF_000332275.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Rubidibacter lacunae_GCF_000473895.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Synechococcus sp._GCF_000515235.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Synechococcus spongiarum_GCF_000586015.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Synechococcus sp._GCF_000737595.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Prochlorococcus sp._GCF_000757865.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Prochlorococcus sp._GCF_000760175.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Prochlorococcus sp._GCF_000760195.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Synechococcus sp._GCF_001182765.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Gloeomargarita lithophora_GCF_001870225.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Prochlorococcus sp._GCF_001989435.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Prochlorococcus sp._GCF_002026175.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Halomicronema hongdechloris_GCF_002075285.2	Cyanobacteria	x	x	x	x	x	x	x	x	x
Leptolyngbya sp._GCF_002286735.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Synechococcus sp._GCF_002356215.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Calothrix sp._GCF_002368395.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Synechococcus lividus_GCF_002754935.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Synechococcus lacustris_GCF_003011125.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Synechocystis sp._GCF_003015225.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
cyanobacterium endosymbiont_GCF_003574135.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Synechococcus elongatus_GCF_003957805.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Thermosynechococcus vulcanus_GCF_003990665.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Synechococcus sp._GCF_004209775.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Synechococcus sp._GCF_900177825.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Synechococcus sp._GCF_900473935.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Synechococcus sp._GCF_900473955.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Synechococcus sp._GCF_900473965.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Synechococcus sp._GCF_900474245.1	Cyanobacteria	x	x	x	x	x	x	x	x	x
Chlamydiae bacterium_GCA_001794905.1	Chlamydiae	x					x	x	x	x
Chlamydia caviae_GCF_000007605.1	Chlamydiae	x					x	x	x	x
Chlamydia pneumoniae_GCF_000008745.1	Chlamydiae	x					x	x	x	x
Chlamydia felis_GCF_000009945.1	Chlamydiae	x					x	x	x	x
Chlamydia pneumoniae_GCF_000024145.1	Chlamydiae	x					x	x	x	x
Waddlia chondrophila_GCF_000092785.1	Chlamydiae	x					x	x	x	x
Chlamydia trachomatis_GCF_000175515.1	Chlamydiae	x					x	x	x	x
Chlamydia pecorum_GCF_000204135.1	Chlamydiae	x					x	x	x	x
Simkania negevensis_GCF_000237205.1	Chlamydiae	x					x	x	x	x
Parachlamydia acanthamoebae_GCF_000253035.1	Chlamydiae	x					x	x	x	x
Chlamydia ibidis_GCF_000454725.1	Chlamydiae	x					x	x	x	x
Chlamydia gallinacea_GCF_000471025.2	Chlamydiae	x					x	x	x	x
Chlamydia avium_GCF_000583875.1	Chlamydiae	x					x	x	x	x
Neochlamydia sp._GCF_000648235.2	Chlamydiae	x					x	x	x	x
Criblamydia sequanensis_GCF_000750955.1	Chlamydiae	x					x	x	x	x
Neochlamydia sp._GCF_000813645.1	Chlamydiae	x					x	x	x	x
Parachlamydiaceae bacterium_GCF_000829755.1	Chlamydiae	x					x	x	x	x
Chlamydia abortus_GCF_000952935.1	Chlamydiae	x					x	x	x	x
Chlamydia trachomatis_GCF_001398155.1	Chlamydiae	x					x	x	x	x
Protochlamydia naegleriophila_GCF_001499655.1	Chlamydiae	x					x	x	x	x
Parachlamydia sp._GCF_001545115.1	Chlamydiae	x					x	x	x	x
Chlamydia sp._GCF_001653975.1	Chlamydiae	x					x	x	x	x
Protochlamydia sp._GCF_001950615.1	Chlamydiae	x					x	x	x	x
Chlamydia corallus_GCF_002817655.1	Chlamydiae	x					x	x	x	x
Similichlamydia epinepheli_GCF_003056015.1	Chlamydiae	x					x	x	x	x
Similichlamydia laticola_GCF_003339615.1	Chlamydiae	x					x	x	x	x
Chlamydia muridarum_GCF_003433255.1	Chlamydiae	x					x	x	x	x
Estrella lausannensis_GCF_900000175.1	Chlamydiae	x					x	x	x	x
Chlamydiales bacterium_GCF_900093645.1	Chlamydiae	x					x	x	x	x
Chlamydiales bacterium_GCF_900093655.1	Chlamydiae	x					x	x	x	x
Chlamydia suis_GCF_900149385.2	Chlamydiae	x					x	x	x	x
Chlamydia sp._GCF_900239945.1	Chlamydiae	x					x	x	x	x
Chlamydia sp._GCF_900239975.1	Chlamydiae	x					x	x	x	x
Porphyromonas gingivalis_GCF_000007585.1	Bacteroidetes		x							
Azobacteroides pseudotrichonymphae_GCF_000010645.1	Bacteroidetes		x							
Rhodothermus marinus_GCF_000024845.1	Bacteroidetes		x							
Prevotella marshii_GCF_000146675.1	Bacteroidetes		x							
Bacteroidetes oral_GCF_000163695.1	Bacteroidetes		x							
Flavobacteria bacterium_GCF_000173115.1	Bacteroidetes		x							
Porphyromonas endodontalis_GCF_000174815.1	Bacteroidetes		x							
Bacteroides helcogenes_GCF_000186225.1	Bacteroidetes		x							
Prevotella denticola_GCF_000193395.1	Bacteroidetes		x							
Porphyromonas asaccharolytica_GCF_000212375.1	Bacteroidetes		x							

Prevotella oulorum_GCF_000224615.1	Bacteroidetes	x			
Alloprevotella rava_GCF_000234115.1	Bacteroidetes	x			
Blattabacterium sp._GCF_000236405.1	Bacteroidetes	x			
Prevotella intermedia_GCF_000261025.1	Bacteroidetes	x			
Blattabacterium sp._GCF_000262715.1	Bacteroidetes	x			
Porphyromonas sp._GCF_000292995.1	Bacteroidetes	x			
Prevotella micans_GCF_000373705.1	Bacteroidetes	x			
Porphyromonas crevioricanis_GCF_000509265.1	Bacteroidetes	x			
Porphyromonas catoniae_GCF_000565015.1	Bacteroidetes	x			
Prevotella dentasini_GCF_000614065.1	Bacteroidetes	x			
Porphyromonas gingivicanis_GCF_000614585.1	Bacteroidetes	x			
Cardinium endosymbiont_GCF_000689375.1	Bacteroidetes	x			
Walczuchella monophlebidarum_GCF_000709555.1	Bacteroidetes	x			
Porphyromonas canoris_GCF_000765975.1	Bacteroidetes	x			
Porphyromonas sp._GCF_000768875.1	Bacteroidetes	x			
Porphyromonas sp._GCF_000768935.1	Bacteroidetes	x			
Porphyromonas cangingivalis_GCF_000768995.1	Bacteroidetes	x			
Porphyromonas macacae_GCF_000769055.1	Bacteroidetes	x			
Prevotella enoeca_GCF_001444445.1	Bacteroidetes	x			
Sulcia muelleri_GCF_001447915.1	Bacteroidetes	x			
Bacteroidales bacterium_GCF_001552775.1	Bacteroidetes	x			
Capnocytophaga haemolytica_GCF_001553545.1	Bacteroidetes	x			
Dokdonia donghaensis_GCF_001653755.1	Bacteroidetes	x			
Porphyromonas sp._GCF_001815465.1	Bacteroidetes	x			
Gramella salexigens_GCF_001889005.1	Bacteroidetes	x			
Hydrotalea flava_GCF_900089565.1	Bacteroidetes	x			
Cruoricaptor ignavus_GCF_900141665.1	Bacteroidetes	x			
Blochmannia pennsylvanicus_GCF_000011745.1	Proteobacteria		x		
Neorickettsia sennetsu_GCF_000013165.1	Proteobacteria		x		
Baumannia cicadellinicola_GCF_000013185.1	Proteobacteria		x		
Polynucleobacter necessarius_GCF_000019745.1	Proteobacteria		x		
Anaplasma marginale_GCF_000020305.1	Proteobacteria		x		
Anaplasma centrale_GCF_000024505.1	Proteobacteria		x		
Helicobacter mustelae_GCF_000091985.1	Proteobacteria		x		
Riesia pediculicola_GCF_000093065.1	Proteobacteria		x		
Buchnera aphidicola_GCF_000183225.1	Proteobacteria		x		
Wolinella succinogenes_GCF_000196135.1	Proteobacteria		x		
Thiomicrospira cyclica_GCF_000214825.1	Proteobacteria		x		
Moranella endobia_GCF_000219175.1	Proteobacteria		x		
Leptothrix ochracea_GCF_000262525.1	Proteobacteria		x		
Carsonella ruddii_GCF_000287295.1	Proteobacteria		x		
secondary endosymbiont_GCF_000287335.1	Proteobacteria		x		
Portiera aleyrodidarum_GCF_000292685.1	Proteobacteria		x		
Kinetoplastibacterium oncopeltii_GCF_000340865.1	Proteobacteria		x		
Anaplasma phagocytophilum_GCF_000439755.1	Proteobacteria		x		
Brackiella oedipodis_GCF_000621025.1	Proteobacteria		x		
Neorickettsia helminthoeca_GCF_000632985.1	Proteobacteria		x		
Photodesmus blepharus_GCF_000731795.1	Proteobacteria		x		
Hepatobacter penaei_GCF_000742475.1	Proteobacteria		x		
Baumannia cicadellinicola_GCF_000754265.1	Proteobacteria		x		
Coxiella endosymbiont_GCF_000815025.1	Proteobacteria		x		
Riesia pediculischaeffi_GCF_000817295.2	Proteobacteria		x		
Evansia muelleri_GCF_000953435.1	Proteobacteria		x		
Anaplasma phagocytophilum_GCF_000964685.1	Proteobacteria		x		
Pseudomonas stutzeri_GCF_001038645.1	Proteobacteria		x		
Caedimonas varicaedens_GCF_001192655.1	Proteobacteria		x		
Pelagibacteraceae bacterium_GCF_001719255.1	Proteobacteria		x		
Pelagibacteraceae bacterium_GCF_001719475.1	Proteobacteria		x		
Pajaroellobacter abortibovis_GCF_001931505.1	Proteobacteria		x		
Doolittlea endobia_GCF_900039485.1	Proteobacteria		x		
Mikella endobia_GCF_900048045.1	Proteobacteria		x		
Tremblaya princeps_GCF_900080145.1	Proteobacteria		x		
Erwinia haradaeae_GCF_900143135.1	Proteobacteria		x		
Atopobium parvulum_GCF_000024225.1	Actinobacteria			x	
Actinomyces coleocanis_GCF_000159015.1	Actinobacteria			x	
Atopobium vaginae_GCF_000159235.2	Actinobacteria			x	
Atopobium vaginae_GCF_000179715.1	Actinobacteria			x	
Tropheryma whipplei_GCF_000196075.1	Actinobacteria			x	
Aquiluna sp._GCF_000257665.1	Actinobacteria			x	
Scardovia wiggsiae_GCF_000269605.1	Actinobacteria			x	
Mycobacterium intracellulare_GCF_000298095.1	Actinobacteria			x	
Atopobium sp._GCF_000411555.1	Actinobacteria			x	
Atopobium fossor_GCF_000483125.1	Actinobacteria			x	
Rhodoluna lacicola_GCF_000699505.1	Actinobacteria			x	
Gardnerella vaginalis_GCF_001042655.1	Actinobacteria			x	
Scardovia inopinata_GCF_001042695.1	Actinobacteria			x	
Atopobium deltae_GCF_001552785.1	Actinobacteria			x	
Coriobacteriales bacterium_GCF_001552935.1	Actinobacteria			x	
Alloscardovia sp._GCF_001813415.1	Actinobacteria			x	
Rhodoluna planktonica_GCF_001854225.1	Actinobacteria			x	
Olegusella massiliensis_GCF_900078545.1	Actinobacteria			x	
Atopobium minutum_GCF_900105895.1	Actinobacteria			x	
Propionimicrobium sp._GCF_900155645.1	Actinobacteria			x	

Syntrophothermus lipocalidus_GCF_000092405.1	Firmicutes					x				
Abiotrophia defectiva_GCF_000160075.2	Firmicutes					x				
Eubacterium saphenum_GCF_000161975.1	Firmicutes					x				
Listeria monocytogenes_GCF_000168595.2	Firmicutes					x				
Facklamia languida_GCF_000245795.1	Firmicutes					x				
Mogibacterium sp._GCF_000293155.1	Firmicutes					x				
Lactobacillus florum_GCF_000304715.1	Firmicutes					x				
Levyella massiliensis_GCF_000308275.2	Firmicutes					x				
Fructobacillus fructosus_GCF_001047095.1	Firmicutes					x				
Fructobacillus pseudoficulneus_GCF_001047115.1	Firmicutes					x				
Selenomonas sp._GCF_001189555.1	Firmicutes					x				
Lactobacillus brantae_GCF_001436115.1	Firmicutes					x				
Lactobacillus thailandensis_GCF_001436135.1	Firmicutes					x				
Weissella halotolerans_GCF_001436865.1	Firmicutes					x				
Lactobacillus secaliphilus_GCF_001437055.1	Firmicutes					x				
Streptococcus sp._GCF_001578885.1	Firmicutes					x				
Lactobacillus fermentum_GCF_001742205.1	Firmicutes					x				
Lactobacillus delbrueckii_GCF_001888985.1	Firmicutes					x				
Lactobacillus amylophilus_GCF_001936335.1	Firmicutes					x				
Colibacter massiliensis_GCF_900095855.1	Firmicutes					x				
Ndongobacter massiliensis_GCF_900120375.1	Firmicutes					x				
Negativicoccus massiliensis_GCF_900155405.1	Firmicutes					x				
Filamoeba nolandi_136452	Amoebozoa							x		
Sexangularia sp._1486928	Amoebozoa							x		
Mayorella sp._242540	Amoebozoa							x		
Stygamoeba regulata_1042828	Amoebozoa							x		
Paramoeba aestuarina_180227	Amoebozoa							x		
Trichosphaerium sp._498011	Amoebozoa							x		
Paramoeba sp._200890	Amoebozoa							x		
Vannella sp._95228	Amoebozoa							x		
Pessonella sp._498007	Amoebozoa							x		
Vexillifera sp._218659	Amoebozoa							x		
Acidobacterium capsulatum_GCF_000022565.1	Acidobacteria	x	x	x	x	x	x	x	x	x
Pyrinomonas methylaliphatogenes_GCF_000820845.2	Acidobacteria	x	x	x	x	x	x	x	x	x
Sulfurihydrogenibium azorense_GCF_000021545.1	Aquificae	x	x	x	x	x	x	x	x	x
Persephonella marina_GCF_000021565.1	Aquificae	x	x	x	x	x	x	x	x	x
Hydrogenivirga sp._GCF_000171895.1	Aquificae	x	x	x	x	x	x	x	x	x
Thermovibrio ammonificans_GCF_000185805.1	Aquificae	x	x	x	x	x	x	x	x	x
Caldisericum exile_GCF_000284335.1	Caldiserica/Cryosericota gro	x	x	x	x	x	x	x	x	x
Caldithrix abyssi_GCF_000241815.1	Calditrichaeota	x	x	x	x	x	x	x	x	x
Desulfurispirillum indicum_GCF_000177635.2	Chrysiogenetes	x	x	x	x	x	x	x	x	x
Deferribacter desulfuricans_GCF_000010985.1	Deferribacteres	x	x	x	x	x	x	x	x	x
Denitrovibrio acetiphilus_GCF_000025725.1	Deferribacteres	x	x	x	x	x	x	x	x	x
Calditerrivibrio nitroreducens_GCF_000183405.1	Deferribacteres	x	x	x	x	x	x	x	x	x
Flexistipes sinusarabici_GCF_000218625.1	Deferribacteres	x	x	x	x	x	x	x	x	x
Elusimicrobium minutum_GCF_000020145.1	Elusimicrobia	x	x	x	x	x	x	x	x	x
Chitinophaga pinensis_GCF_000024005.1	Bacteroidetes	x		x	x	x	x	x	x	x
Paludibacter propionicigenes_GCF_000183135.1	Bacteroidetes	x		x	x	x	x	x	x	x
Porphyromonas asaccharolytica_GCF_000212375.1	Bacteroidetes	x		x	x	x	x	x	x	x
Rhodothermus marinus_GCF_000224745.1	Bacteroidetes	x		x	x	x	x	x	x	x
Fulvivirga imtechensis_GCF_000331535.1	Bacteroidetes	x		x	x	x	x	x	x	x
Walczuchella monophlebidarum_GCF_000709555.1	Bacteroidetes	x		x	x	x	x	x	x	x
Sulcia muelleri_GCF_001447915.1	Bacteroidetes	x		x	x	x	x	x	x	x
Chlorobium chlorochromatii_GCF_000012585.1	Chlorobi	x	x	x	x	x	x	x	x	x
Chlorobium phaeovibrioides_GCF_000016085.1	Chlorobi	x	x	x	x	x	x	x	x	x
Chlorobaculum parvum_GCF_000020505.1	Chlorobi	x	x	x	x	x	x	x	x	x
Chloroherpeton thalassium_GCF_000020525.1	Chlorobi	x	x	x	x	x	x	x	x	x
Fibrobacter succinogenes_GCF_000146505.1	Fibrobacteria	x	x	x	x	x	x	x	x	x
Gemmatimonas aurantiaca_GCF_000010305.1	Gemmatimonadetes	x	x	x	x	x	x	x	x	x
Ignavibacterium album_GCF_000258405.1	Ignavibacteriae	x	x	x	x	x	x	x	x	x
Melioribacter roseus_GCF_000279145.1	Ignavibacteriae	x	x	x	x	x	x	x	x	x
Sebaldella termitidis_GCF_000024405.1	Fusobacteria	x	x	x	x	x	x	x	x	x
Fusobacterium mortiferum_GCF_000158195.2	Fusobacteria	x	x	x	x	x	x	x	x	x
Ilyobacter polytropus_GCF_000165505.1	Fusobacteria	x	x	x	x	x	x	x	x	x
Fusobacterium necrophorum_GCF_000292975.1	Fusobacteria	x	x	x	x	x	x	x	x	x
Thermodesulfovibrio yellowstonii_GCF_000020985.1	Nitrospirae	x	x	x	x	x	x	x	x	x
Leptospirillum ferrooxidans_GCF_000284315.1	Nitrospirae	x	x	x	x	x	x	x	x	x
Caulobacter vibrioides_GCF_000006905.1	Proteobacteria	x	x		x	x	x	x	x	x
Neorickettsia sennetsu_GCF_000013165.1	Proteobacteria	x	x		x	x	x	x	x	x
Magnetococcus marinus_GCF_000014865.1	Proteobacteria	x	x		x	x	x	x	x	x
Polynucleobacter necessarius_GCF_000019745.1	Proteobacteria	x	x		x	x	x	x	x	x
Anaplasma marginale_GCF_000020305.1	Proteobacteria	x	x		x	x	x	x	x	x
Geobacter bemidjiensis_GCF_000020725.1	Proteobacteria	x	x		x	x	x	x	x	x
Escherichia coli_GCF_000026545.1	Proteobacteria	x	x		x	x	x	x	x	x
Sorangium cellulosum_GCF_000067165.1	Proteobacteria	x	x		x	x	x	x	x	x
Neisseria meningitidis_GCF_000083565.1	Proteobacteria	x	x		x	x	x	x	x	x
Arcobacter nitrofigilis_GCF_000092245.1	Proteobacteria	x	x		x	x	x	x	x	x
Nitrosomonas sp._GCF_000175095.2	Proteobacteria	x	x		x	x	x	x	x	x
Bordetella pertussis_GCF_000195715.1	Proteobacteria	x	x		x	x	x	x	x	x
Moranella endobia_GCF_000219175.1	Proteobacteria	x	x		x	x	x	x	x	x
Leptothrix ochracea_GCF_000262525.1	Proteobacteria	x	x		x	x	x	x	x	x
Carsonella ruddii_GCF_000287295.1	Proteobacteria	x	x		x	x	x	x	x	x
secondary endosymbiont_GCF_000287335.1	Proteobacteria	x	x		x	x	x	x	x	x

Kinetoplastibacterium oncopeltii_GCF_000340865.1	Proteobacteria	x	x		x	x	x	x	x	x
Anaplasma phagocytophilum_GCF_000439755.1	Proteobacteria	x	x		x	x	x	x	x	x
Neorickettsia helminthoeca_GCF_000632985.1	Proteobacteria	x	x		x	x	x	x	x	x
Baumannia cicadellinicola_GCF_000754265.1	Proteobacteria	x	x		x	x	x	x	x	x
Coxiella endosymbiont_GCF_000815025.1	Proteobacteria	x	x		x	x	x	x	x	x
Helicobacter pylori_GCF_000982695.1	Proteobacteria	x	x		x	x	x	x	x	x
Pelagibacteraceae bacterium_GCF_001719255.1	Proteobacteria	x	x		x	x	x	x	x	x
Pelagibacteraceae bacterium_GCF_001719475.1	Proteobacteria	x	x		x	x	x	x	x	x
Doolittlea endobia_GCF_900039485.1	Proteobacteria	x	x		x	x	x	x	x	x
Mikella endobia_GCF_900048045.1	Proteobacteria	x	x		x	x	x	x	x	x
Tremblaya princeps_GCF_900080145.1	Proteobacteria	x	x		x	x	x	x	x	x
Chlamydia trachomatis_GCF_000175515.1	Chlamydiae		x	x	x	x				
Chlamydia pecorum_GCF_000204135.1	Chlamydiae		x	x	x	x				
Chlamydia trachomatis_GCF_001398155.1	Chlamydiae		x	x	x	x				
Lentisphaera araneosa_GCF_000170755.1	Lentisphaerae	x	x	x	x	x	x	x	x	x
Gimesia maris_GCF_000181475.1	Planctomycetes	x	x	x	x	x	x	x	x	x
Jettenia caeni_GCF_000296795.1	Planctomycetes	x	x	x	x	x	x	x	x	x
Opitutus terrae_GCF_000019965.1	Verrucomicrobia	x	x	x	x	x	x	x	x	x
Verrucomicrobiae bacterium_GCF_000155695.1	Verrucomicrobia	x	x	x	x	x	x	x	x	x
Pedosphaera parvula_GCF_000172555.1	Verrucomicrobia	x	x	x	x	x	x	x	x	x
Xiphinematobacter sp._GCF_001318295.1	Verrucomicrobia	x	x	x	x	x	x	x	x	x
Leptospira interrogans_GCF_000092565.1	Spirochaetes	x	x	x	x	x	x	x	x	x
Treponema paraluiscuniculi_GCF_000217655.1	Spirochaetes	x	x	x	x	x	x	x	x	x
Leptonema illini_GCF_000243335.1	Spirochaetes	x	x	x	x	x	x	x	x	x
Turneriella parva_GCF_000266885.1	Spirochaetes	x	x	x	x	x	x	x	x	x
Leptospira meyeri_GCF_000347075.1	Spirochaetes	x	x	x	x	x	x	x	x	x
Thermanaerovibrio acidaminovorans_GCF_000024905.1	Synergistetes	x	x	x	x	x	x	x	x	x
Aminobacterium colombiense_GCF_000025885.1	Synergistetes	x	x	x	x	x	x	x	x	x
Aminomonas paucivorans_GCF_000165795.1	Synergistetes	x	x	x	x	x	x	x	x	x
Dethiosulfovibrio peptidovorans_GCF_000172975.1	Synergistetes	x	x	x	x	x	x	x	x	x
Thermovirga lienii_GCF_000233775.1	Synergistetes	x	x	x	x	x	x	x	x	x
Synergistes sp._GCF_000238615.1	Synergistetes	x	x	x	x	x	x	x	x	x
Acetomicrobium mobile_GCF_000266925.1	Synergistetes	x	x	x	x	x	x	x	x	x
Rubrobacter xylanophilus_GCF_000014185.1	Actinobacteria	x	x	x		x	x	x	x	x
Acidothermus cellulolyticus_GCF_000015025.1	Actinobacteria	x	x	x		x	x	x	x	x
Atopobium parvulum_GCF_000024225.1	Actinobacteria	x	x	x		x	x	x	x	x
Propionibacterium freudenreichii_GCF_000091725.1	Actinobacteria	x	x	x		x	x	x	x	x
Actinomyces coleocanis_GCF_000159015.1	Actinobacteria	x	x	x		x	x	x	x	x
Atopobium vaginae_GCF_000159235.2	Actinobacteria	x	x	x		x	x	x	x	x
Tropheryma whipplei_GCF_000196075.1	Actinobacteria	x	x	x		x	x	x	x	x
Streptomyces coelicolor_GCF_000203835.1	Actinobacteria	x	x	x		x	x	x	x	x
Aquiluna sp._GCF_000257665.1	Actinobacteria	x	x	x		x	x	x	x	x
Scardovia wiggsiae_GCF_000269605.1	Actinobacteria	x	x	x		x	x	x	x	x
Atopobium fossor_GCF_000483125.1	Actinobacteria	x	x	x		x	x	x	x	x
Rhodoluna lacicola_GCF_000699505.1	Actinobacteria	x	x	x		x	x	x	x	x
Scardovia inopinata_GCF_001042695.1	Actinobacteria	x	x	x		x	x	x	x	x
Atopobium rimae_GCF_001438885.1	Actinobacteria	x	x	x		x	x	x	x	x
Atopobium deltae_GCF_001552785.1	Actinobacteria	x	x	x		x	x	x	x	x
Olegusella massiliensis_GCF_900078545.1	Actinobacteria	x	x	x		x	x	x	x	x
Atopobium minutum_GCF_900105895.1	Actinobacteria	x	x	x		x	x	x	x	x
Herpetosiphon aurantiacus_GCF_000018565.1	Chloroflexi	x	x	x	x	x	x	x	x	x
Thermomicrobium roseum_GCF_000021685.1	Chloroflexi	x	x	x	x	x	x	x	x	x
Ktedonobacter racemifer_GCF_000178855.1	Chloroflexi	x	x	x	x	x	x	x	x	x
Caldilinea aerophila_GCF_000281175.1	Chloroflexi	x	x	x	x	x	x	x	x	x
Dehalococcoides mccartyi_GCF_002007825.1	Chloroflexi	x	x	x	x	x	x	x	x	x
Deinococcus deserti_GCF_000020685.1	Deinococcus-Thermus	x	x	x	x	x	x	x	x	x
Oceanithermus profundus_GCF_000183745.1	Deinococcus-Thermus	x	x	x	x	x	x	x	x	x
Thermus scotoductus_GCF_000187005.1	Deinococcus-Thermus	x	x	x	x	x	x	x	x	x
Marinithermus hydrothermalis_GCF_000195335.1	Deinococcus-Thermus	x	x	x	x	x	x	x	x	x
Staphylococcus aureus_GCF_000011505.1	Firmicutes	x	x	x	x		x	x	x	x
Heliobacterium modesticaldum_GCF_000019165.1	Firmicutes	x	x	x	x		x	x	x	x
Clostridium saccharolyticum_GCF_000144625.1	Firmicutes	x	x	x	x		x	x	x	x
Acetohalobium arabaticum_GCF_000144695.1	Firmicutes	x	x	x	x		x	x	x	x
Erysipelatoclostridium ramosum_GCF_000154485.1	Firmicutes	x	x	x	x		x	x	x	x
Thermodesulfobium narugense_GCF_000212395.1	Firmicutes	x	x	x	x		x	x	x	x
Desulfotomaculum nigrificans_GCF_000214435.1	Firmicutes	x	x	x	x		x	x	x	x
Pelosinus fermentans_GCF_000271485.1	Firmicutes	x	x	x	x		x	x	x	x
Bacillus subtilis_GCF_000344745.1	Firmicutes	x	x	x	x		x	x	x	x
Weissella halotolerans_GCF_001436865.1	Firmicutes	x	x	x	x		x	x	x	x
Streptococcus pyogenes_GCF_001635935.1	Firmicutes	x	x	x	x		x	x	x	x
Negativicoccus massiliensis_GCF_900155405.1	Firmicutes	x	x	x	x		x	x	x	x
Mycoplasma penetrans_GCF_000011225.1	Tenericutes	x	x	x	x	x	x	x	x	x
Acholeplasma laidlawii_GCF_000018785.1	Tenericutes	x	x	x	x	x	x	x	x	x
Mycoplasma haemominutum_GCF_000319365.1	Tenericutes	x	x	x	x	x	x	x	x	x
Spiroplasma melliferum_GCF_000328865.1	Tenericutes	x	x	x	x	x	x	x	x	x
Spiroplasma chrysopicola_GCF_000400935.1	Tenericutes	x	x	x	x	x	x	x	x	x
Mycoplasma pneumoniae_GCF_001509195.1	Tenericutes	x	x	x	x	x	x	x	x	x
Thermobaculum terrenum_GCF_000025005.1	unclassified Terrabacteria gr	x	x	x	x	x	x	x	x	x
Thermodesulfobacterium geofontis_GCF_000215975.1	Thermodesulfobacteria	x	x	x	x	x	x	x	x	x
Thermodesulfatator indicus_GCF_000217795.1	Thermodesulfobacteria	x	x	x	x	x	x	x	x	x
Petrotoga mobilis_GCF_000018605.1	Thermotogae	x	x	x	x	x	x	x	x	x
Thermotoga neapolitana_GCF_000018945.1	Thermotogae	x	x	x	x	x	x	x	x	x
Thermosipho africanus_GCF_000021285.1	Thermotogae	x	x	x	x	x	x	x	x	x

Pseudothermotoga thermarum_GCF_000217815.1	Thermotogae	x	x	x	x	x	x	x	x	x
Thermotoga sp._GCF_000832145.1	Thermotogae	x	x	x	x	x	x	x	x	x