Impact des violations des modèles d'annotation et d'évolution de séquences en phylogénomique : application à l'étude des eucaryotes photosynthétiques

Arnaud Di Franco

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 3 - Paul Sabatier

Présentée et soutenue par

Arnaud DI FRANCO

Le 13 mai 2019

Impact des violations des modèles d'annotation et d'évolution de séquences en phylogénomique: Application à l'étude des eucaryotes photosynthétiques

Ecole doctorale : **SEVAB - Sciences Ecologiques, Vétérinaires, Agronomiques et Bioingenieries**

Spécialité : **Ecologie, biodiversité et évolution**

Unité de recherche :
SETE - Station d'Ecologie Théorique et Expérimentale

Thèse dirigée par
Hervé PHILIPPE

Jury

M. Vincent Daubin, Rapporteur
Mme Purificación López-García, Rapporteure
M. Jérôme Murienne, Examinateur
M. Frédéric Delsuc, Examinateur
M. Hervé PHILIPPE, Directeur de thèse

Résumé

L'oxygène a façonné la vie sur Terre. L'enrichissement de sa forme gazeuse dans l'eau puis dans l'atmosphère a permis le développement de la vie multicellulaire et terrestre menant à la biodiversité actuelle. Celui-ci a pu avoir lieu grâce la mise en place du processus de la photosynthèse oxydative chez les organismes vivants. Ce dernier est d'abord apparu chez les bactérie avant d'être subtilisé par les formes de vie nucléées. Cette action s'est réalisé en asservissant la bactérie et en la maintenant à l'intérieur de l'organisme nucléé par un phénomène nommé endosymbiose. Différentes endosymbioses ont eu lieu dans l'histoire des organismes photosynthétiques, attribuant la capacité de photosynthèse à un large panel d'êtres vivants.

L'objectif de cette thèse est d'étudier la transmission de la photosynthèse chez les organismes eucaryotes. Ces derniers présentent une grande diversité de chloroplaste, l'organite réalisant la photosynthèse et témoin de l'intégration d'un organisme étranger à l'intérieur de leur cellule. L'inférence de la phylogénie, i.e. l'estimation des relations de parentés entre les organismes, révèle des discordances entre l'histoire racontée par le génome des chloroplastes et des noyaux. L'obtention de ces phylogénies et l'étude de leurs discordances sont au coeur de la compréhension de l'historique de l'acquisition de la photosynthèse. Cependant, l'inférence de la phylogénie est un procédé complexe influencé par la nature et la qualité des données ainsi que par les techniques employées. Les considérations de ce manuscrit de thèse se focalise sur l'impact de ces éléments sur la résolution de l'arbre des eucaryotes, avec pour objectif une meilleure compréhension de l'histoire des hôtes photosynthétiques et de leur endosymbionte.

Premièrement, nous avons développé un logiciel améliorant la qualité des inférences phylogénétiques par le retrait de segments de séquences déterminés comme non relatif à l'évolution des organismes. Nous démontrons l'efficacité de la méthode et son impact comparé aux autres méthodes de filtrage de séquence couramment employées.

Secondairement, nous créons un jeu de données phylogénomique en vue d'inférer la phylogénie des eucaryotes. Celui-ci est réalisé de manière semi-automatique et vise à retirer le maximum de signal phylogénétique tout en évitant l'intégration de séquences ne permettant pas de retracer les liens de parentés entre organisme. Nous obtenons un arbre des eucaryotes comprenant la plus grande diversité en organismes à ce jour et discutons l'impact de l'échantillonnage sur les soutiens apportés à la topologie de l'arbre.

Dernièrement, nous avons étudié l'impact du choix du modèle d'évolution des séquences sur la congruence des phylogénies obtenues entre les génomes des différents compartiments présents chez les stramenopiles photosynthétiques. Nos résultats sont en faveur de la présence d'un faible signal phylogénétique pour résoudre les noeuds à la base de ce groupe, ce dernier pouvant être facilement dépassé par le signal non phylogénétique produits par les violations de modèles.

Au final, cette thèse met en évidence l'importance du développement des méthodes bioinformatique liées à la phylogénie afin de répondre avec assurance aux questions évolutives relatifs à des événements anciens.

Sommaire

Introduction

La photosynthèse à la base de l'évolution des eucaryotes

Les mystères entourant l'origine de la vie sont parmi les questionnements existentiels les plus fascinants. Au cours des siècles, les religions et courants philosophiques ont fourni leur vision du déroulement de la genèse afin de répondre aux interrogations de l'être humain. À notre époque, on s'attend souvent à ce que les différents domaines scientifiques nous fournissent à leur tour des réponses. Mais, de son côté, la communauté scientifique peaufine encore ses propositions. En effet, peu de choses demeurent incontestables sur cette version du "Jardin d'Eden", que l'on s'intéresse à sa situation (mares, évents hydrothermaux, sols) [Martin et al. 2008 ; Pearce et al. 2017] ou à son type d'environnement (atmosphère neutre ou réductrice, chaude ou non) [Lazcano et Miller 1996 ; Orgel 1998]. Une première hypothèse suppose qu'il contenait une "soupe primitive" [Oparin 1924 ; Haldane 1929], composée des molécules essentielles à l'apparition du vivant, de laquelle émergea le premier organisme, l' "Adam" scientifique. L'origine de la vie selon la science repose donc sur des phénomènes chimiques mettant en scène les éléments inorganiques présents dans l'environnement.

C'est au domaine de la chimie prébiotique que revient la tâche d'élucider les étapes ayant permis à la soupe d'engendrer les premiers organismes vivants. Evidemment, l'idée que la vie est apparue par chance au milieu d'un mélange inorganique ne plait pas aux chimistes et ceux-ci tentent encore d'améliorer l'hypothèse de départ. Deux grandes hypothèses coexistent aujourd'hui débattant sur l'ordre d'ajout des ingrédients. La première donne la priorité à l'encodage de l'information et suppose l'arrivée rapide des acides ribonucléiques, principalement à cause de leur capacité de catalyse et surtout d'auto-réplication (Hypothèse RNA-world, [Gilbert 1986]). La seconde considère que ces molécules ne purent apparaître et devenir autonomes sans la mise en place d'un métabolisme primitif générant les éléments plus complexes de la soupe (Hypothèse Metabolism-first, [Wächtershäuser 1990 ; Martin et Russell 2003]). Au final, on s'accorde à dire que ces deux éléments devaient être présents pour constituer les premiers être vivants tels que nous les connaissons, délimités par une membrane cellulaire. L'addition de ces mécanismes clefs a mené à l'apparition d'une ou plusieurs populations, parmi lesquels l'une d'entre elles, que nous

nommons LUCA (Last Universal Common Ancestor ou dernier ancêtre commun universel), fut à l'origine des êtres vivants qui nous sont familiers.

Selon la théorie de l'évolution [Darwin 1859], LUCA est l'organisme dont découle l'entièreté des êtres vivants actuels sur Terre. Il est donc à la base des trois domaines du vivant actuellement décrits que sont les bactéries (Bacteria), les archées (Archeae) et les eucaryotes (Eucarya ou Eukaryota) [Woese et al. 1990]. Cela signifie que des liens de spéciation connectent les espèces contemporaines entre elles et ultimement à LUCA. La phylogénie correspond au domaine de la science dédié à l'étude de ces liens de parenté. C'est par leur connaissance qu'il devient possible de proposer et de tester des hypothèses sur le déroulement de l'évolution des organismes. On peut alors s'intéresser aux mécanismes génétiques ayant séparé ces espèces (évolution moléculaire) ou encore tenter d'associer cette information à celles d'autres domaines de la science (paléontologie, géologie) dans le but d'étudier leur impact sur l'évolution. La phylogénie contribue ainsi à retracer l'évolution du vivant sur la Terre.

Si les astrophysiciens et géologues estiment la formation de la Terre à environ 4,565 milliards d'années, la date des premières traces de vie sur la planète bleue est moins précise. Actuellement, le registre fossile estime l'apparition de la vie sur Terre durant l'ère Archéen (4-2,5 Ga), avec des microfossiles datant d'il y a 3,5 Ga retrouvés en Australie [Kudryavtsev et al. 2007 ; Schopf et al. 2018]. Ces fossiles constituent actuellement les traces de vie anciennes les plus fiables. On estime d'ailleurs que la vie n'aurait pas pu apparaître beaucoup plus tôt (i.e. <3,9 Ga) à cause du grand bombardement tardif (Late Heavy Bombardment ; période durant laquelle les planètes du système solaire fut marqué par de nombreux impacts de météorites et comètes). Cependant, de nouveaux résultats pourraient faire reculer les traces de vie à plus de 4,2 Ga, lors de l'Hadéen [Dodd et al. 2017].

Trouver un fossile valide reste une tâche difficile mais c'est également le cas lorsqu'il s'agit de l'attribuer avec certitude à un des domaines du vivant. Dernièrement, les fossiles les plus anciens, précédemment évoqués, ont été affiliés au domaine des Archaea [Schopf et al. 2018]. Ici, nous nous intéressons plus particulièrement aux signes d'organismes eucaryotes. Si certains biomarqueurs (stéranes) valideraient leur présence dès 2,7 Ga [Brocks et al. 1999 ; Brocks et al. 2003], les premiers fossiles qui leur sont attribués de manière non ambiguë n'apparaissent qu'au environ de 1,7 Ga (*Valeria lophostriata*, [Javaux 2011 ; Javaux et Lepot 2018]). On retrouve peu de diversité parmi les fossiles eucaryotes même s'ils présentent déjà de nombreux traits propres aux eucaryotes contemporains [Knoll et al. 2006]. Plusieurs fossiles sont notamment affiliés directement à des eucaryotes photosynthétiques, comme les algues rouges (*Bangiomorpha pubescens*, 1,2 Ga [Butterfield 2000] ; *Rafatazmia* et *Ramathallus*, 1,6 Ga, Bengtson et al. 2017), supposant une ap-

parition assez rapide de la photosynthèse dans ce domaine du vivant. Cependant, les apparitions de traces eucaryotes restent limitées durant cette période baptisée "Boring billion" (1,8-0,8 Ga) pour sa stabilité géologique, un frein longtemps supposé pour le développement des formes de vie complexes [Mukherjee et al. 2018]. Le registre fossile s'agrandit à la sortie d'une importante succession d'ère glaciaire (Cryogénien, 720-635 Ma) en présentant une grande diversité d'organismes pluricellulaires (Ediacarien, 635–541 Ma). La majorité de ces organismes seront cependant remplacés par l'apparition massive des ancêtres des animaux, lors de l'explosion cambrienne (541 Ma - Figure 1).

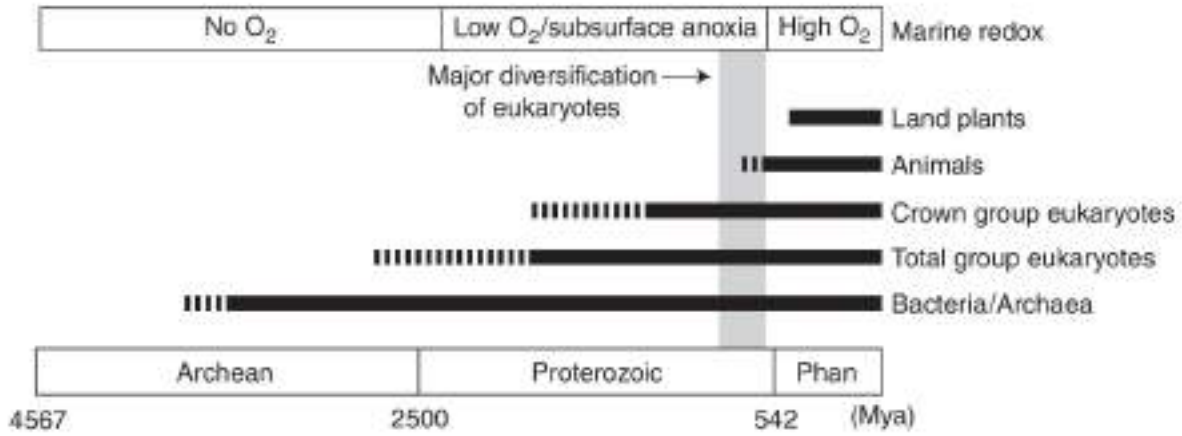


Figure 1 – Résumé du début de l'évolution des eucaryotes en fonction de la présence d'oxygène Knoll 2014.

Plusieurs théories ont essayé d'expliquer la soudaine explosion de diversité apparue au Cambrien. Parmi elles, l'oxygénation des océans semble jouer un rôle clé [Glass et al. 2009 ; Lenton et al. 2014 ; Brocks et al. 2017]. Outre l'approvisionnement de la respiration oxygénique des animaux, le passage à un milieu oxygénique a également affecté la présence de plusieurs éléments dans le milieu, comme la mobilisation du molybdène [Glass et al. 2009] et la diminution du phosphate et des éléments ferreux et sulfureux [Lenton et al. 2014]. Si ces changements n'ont peut-être pas directement affecté les espèces du Cambrien, ils ont favorisé la domination des eucaryotes photosynthétiques [Falkowski et al. 2004 ; Brocks et al. 2017] en tant que producteurs primaires qui fourniront aux autres eucaryotes une nourriture plus riche que les procaryotes majoritaires de l'époque.

Les eucaryotes photosynthétiques réalisent la photosynthèse oxygénique, un processus bioénergétique qui, à partir de l'énergie lumineuse, d'eau et de dioxyde de carbone (CO_2), produit des polysaccharides (molécules énergétiques) et rejette du dioxygène (O_2). Ce type de photosynthèse fut d'abord réalisé par les ancêtres des cyanobactéries actuelles [Shih 2015]. C'est par leur action que le milieu terrestre subit un premier enrichissement en O_2, la Grande Oxygénation (GOE : Great Oxydation Event, 2,4 Ga [Bekker

et al. 2004]). Les taux d'O_2 resteront stables jusqu'au transfert de cette capacité aux eucaryotes et leur développement subséquent juste avant l'explosion cambrienne. En effet, tout au long de l'histoire de la vie sur terre, différents organismes photosynthétiques ont dominé la production d'O_2 dans les océans [Falkowski et al. 2004 ; Martin et Quigg 2013]. Les cyanobactéries ont d'abord laissé leur place à des algues arborant des pigments verts à l'entrée dans le Cambrien (541 Ma, [Falkowski et al. 2004]). Ces dernières ont ensuite été remplacées au Mésozoïque (251-65 Ma, [Falkowski et al. 2004]) par les algues à pigments rouges qui dominent encore le phytoplancton actuel. Ces deux changements écologiques majeurs ont donc eu lieu en même temps que d'importants événements ; l'explosion cambrienne et, plus tard, le passage du Permien au Trias caractérisé par la plus grande extinction de l'histoire de la vie sur Terre.

Mon sujet de thèse s'inscrit dans ce contexte de coévolution liant l'évolution des eucaryotes photosynthétiques à celles des autres organismes eucaryotes ainsi qu'aux modifications environnementales ayant eu lieu sur Terre. Cette thèse avait pour objectif d'améliorer notre compréhension de l'évolution des organismes photosynthétiques et des contraintes environnementales qui l'ont façonnée. La suite de l'introduction de ce manuscrit sera divisée en plusieurs points. Dans un premier temps, j'introduirai les principes fondamentaux de la phylogénie moléculaire qui me permettront de mieux expliquer les prochains points. J'enchainerai alors avec une brève présentation des mécanismes liés à l'acquisition et la transmission de la photosynthèse oxygénique. Ensuite, je détaillerai la diversité des organismes la réalisant en m'aidant de la phylogénie des marqueurs liés à la photosynthèse et les repositionnerai par rapport à l'ensemble des eucaryotes grâce aux phylogénies obtenues à partir de leur génome nucléaire. Finalement, je reviendrai sur le processus de l'inférence phylogénétique et sur ces limitations au niveau de la phylogénomique. Ces points constitueront ensuite les sujets de discussions principaux des chapitres de ce manuscrit. Dans cette thèse uniquement la photosynthèse oxygénique sera abordée. Il existe en effet des photosynthèses non oxygéniques (ne produisant pas d'oxygène), comme celle réalisée par les bactéries pourpres sulfureuses. À partir de maintenant, toute évocation de la photosynthèse visera exclusivement la photosynthèse oxygénique.

La phylogénie moléculaire : déterminer les liens de parenté entre organismes grâce aux macromolécules biologiques.

L'inférence des liens de parenté entre espèces, qui sont représentés par la phylogénie, se réalise par la comparaison de caractères entre individus. Pour rendre cette comparaison

valable, les caractères choisis doivent être homologues, c'est-à-dire être hérités d'un même ancêtre commun [Fitch 2000]. Le critère d'homologie n'oblige cependant pas le caractère en question à être identique, ni même semblable, entre les espèces. Un exemple simple en anatomie est celui des membres chez les vertébrés. Si un bras, une patte, une aile ou une nageoire ne se ressemblent pas beaucoup, les os qui les composent ont bel et bien la même origine. La diversité des formes que peut prendre un caractère rend donc parfois l'établissement de l'homologie ainsi que la comparaison des différents états d'un caractère difficile, voire impossible quand l'échelle évolutive devient trop grande. Si les premières phylogénies étaient réalisées sur la base de caractères morphologiques, cela fait plusieurs années que ce type de phylogénie a été supplanté par celles fondées sur les données moléculaires afin de faciliter l'identification et la comparaison de caractères homologues.

Les données moléculaires d'un organisme correspondent aux séquences d'éléments constitutifs de ses macromolécules biologiques. Elles sont acquises par des séquenceurs qui digitalisent le contenu de ces molécules en chaînes de caractères informatiques. Les molécules les plus communément récupérées sont les acides désoxy-ribonucléiques (ADN), constituant le génome de l'individu, les acides ribonucléiques (ARN) puis les protéines. Grâce aux connaissances de la biologie moléculaire et de la biochimie, on peut élargir l'ensemble des données moléculaires à d'autres types d'informations comme la composition complète d'un chromosome ou le positionnement d'un gène sur ce dernier. Plusieurs types de données sont utilisables dans le but de réaliser une phylogénie moléculaire comme l'ordre des gènes présents dans une région chromosomique ou encore la fréquence d'apparition de sous-éĺéments courts (kmer). Cependant, la majorité des inférences en phylogénies moléculaires sont basées sur la comparaison des séquences de gènes codants pour des protéines, que ce soit via les nucléotides ou les acides aminés. Dans la suite de cette partie, je vais revenir sur les principes de bases de la phylogénie moléculaire tel que hérité de la morphologie et appliqué à un marqueur unique.

Le concept d'homologie en biologie moléculaire

Comme pour les caractères morphologiques, les caractères moléculaires ne peuvent correctement mesurer les liens de parentés entre espèces que si ceux-ci sont homologues. Cependant, on distingue plusieurs catégories d'homologies entre séquences moléculaires, l'orthologie, la paralogie et la xénologie [Fitch 1970]. On considère les gènes issus d'évènements de spéciation comme orthologues (les gènes α dérivés de la copie α chez l'ancêtre commun, de même pour les gènes β dérivés de la copie β, Figure 2A) et ceux issus d'évènements de duplication comme paralogues (les copies α et β chez un même organisme, par exemple). Le dernier cas des gènes xénologues est comparable à celui des paralogues

par leur incapacité à représenter les liens de spéciation. Ils représentent des transferts de matériel génétique en provenance d'une autre espèce, pour lequel le lien de parenté n'est pas direct. Ainsi, l'inférence de la phylogénie des espèces, dans ce contexte hérité de la morphologie de comparaison de caractères homologues, requiert l'analyse de gènes strictement orthologues, seuls témoins des événements de spéciation. L'insertion de gènes paralogues et/ou xénologues dans ce type d'analyse est à éviter à tout prix si on veut reconstruire la phylogénie des espèces car ils faussent le résultat de l'inférence (Figure 2B) [Roy 2009 ; Philippe et al. 2011b ; Struck 2013].

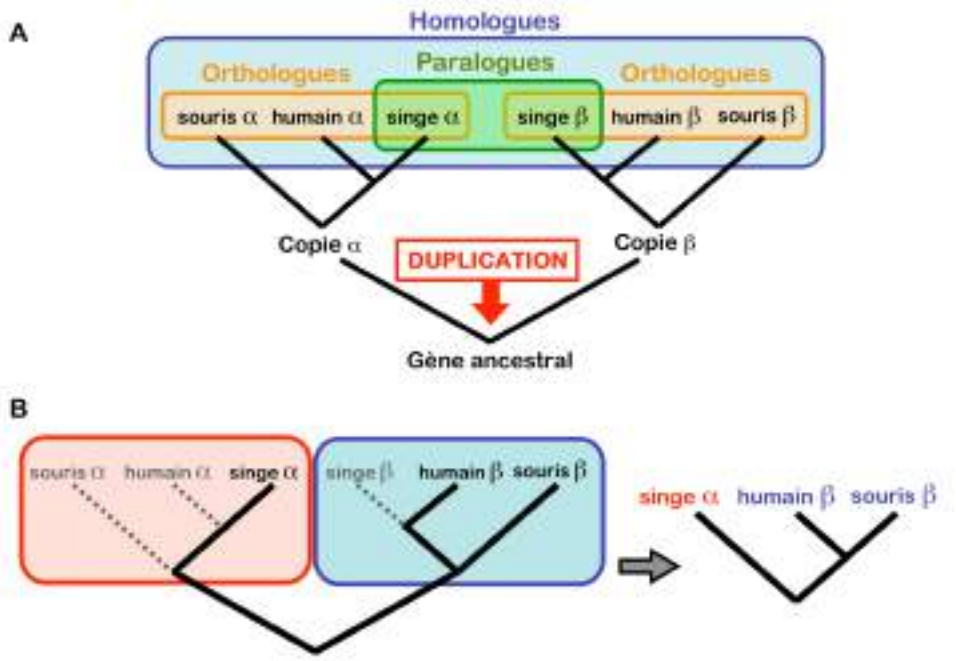


Figure 2 – A : Différenciation entre deux cas d'homologie, l'orthologie et la paralogie. B : Inférence d'un arbre contenant des copies paralogues menant à un arbre ne représentant pas la phylogénie des espèces.

Une fois les séquences orthologues rassemblées, le critère d'homologie doit également être vérifié pour chaque résidu des séquences sélectionnées. Cette seconde étape est réalisée par l'alignement des séquences entre elles. Il s'agit de positionner chaque acide aminé ou nucléotide face aux homologues correspondants afin de prendre en compte les possibles modifications subies par les séquences au cours du temps. En effet, les séquences moléculaires peuvent être affectées par divers types de changements, comme l'ajout ou le retrait de résidus (insertion et délétion) ou leur remplacement (substitution). L'alignement des séquences peut être réalisé manuellement mais tend à être confié exclusivement à des logiciels informatiques de nos jours [Chatzou et al. 2016]. Il constitue une étape cruciale pour permettre la comparaison de caractères précédant l'obtention d'une phylogénie.

Représentation de la phylogénie

Les liens de parenté entre organismes sont illustrés par un schéma spécifique appelé arbre phylogénétique. Ce graphe est composé de noeuds, représentant les espèces (actuelles ou ancestrales), et de branches, connectant les espèces entre elles. On distingue deux types de noeuds, les noeuds terminaux ou feuilles (connectés par une seule branche) représentant les organismes actuels et les noeuds internes, représentant des organismes ancestraux aux autres. Les arbres peuvent être racinés ou non en fonction de la présence d'un noeud particulier, la racine, qui détermine la position de l'ancêtre commun de l'ensemble des espèces étudiées. Finalement, les branches ont pour objectif de quantifier la distance génétique séparant les organismes. Ainsi, cette dernière peut être calculée par la somme de la taille des branches séparant deux noeuds de l'arbre. Si l'on s'intéresse uniquement aux branchements d'un arbre sans considérer la taille de ses branches, on parlera souvent de topologie.

Le positionnement des organismes dans un arbre phylogénétique permet de distinguer plusieurs types de groupement (Figure 3). Ainsi, on dit qu'un groupe est monophylétique lorsqu'il se compose de l'ancêtre commun et de l'ensemble des organismes qui en ont dérivé. Un groupe excluant certains des organismes dérivés est appelé paraphylétique, alors qu'un autre, n'incluant pas l'ancêtre commun des organismes le composant, est appelé polyphylétique.

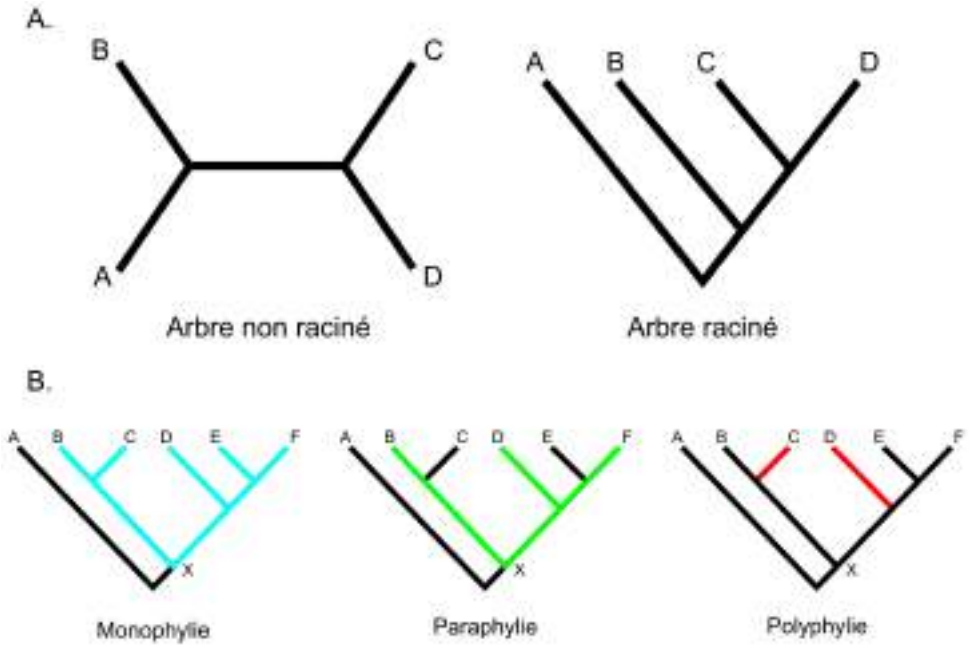


FIGURE 3 – A : Un arbre de 4 espèces non raciné suivi du même arbre raciné entre l'espèce A et les espèces B,C et D. B : Représentation d'un groupe monophylétique (bleu), d'un groupe paraphylétique (vert) et d'un groupe polyphylétique (rouge).

Si l'on étudie trois organismes, il existe une seule topologie non-racinée possible et trois to-

pologies racinées, une par branche existante. Les nombres de topologies possibles peuvent être estimés par $(2n\text{-}5)!/2^{n\text{-}3}(n\text{-}3)!$ et $(2n\text{-}3)!/2^{n\text{-}2}(n\text{-}2)!$ avec n le nombre d'organismes pour les topologies non-racinées et racinées, respectivement [Felsenstein 1978b]. Le nombre de topologies possibles dépasse le milliard dès 12 organismes et continue d'augmenter très rapidement pour devenir supérieur à $3x10^{72}$ à 50 organismes. La détermination de la bonne topologie est donc un problème exponentiellement complexe en fonction du nombre d'espèces étudiées.

Il est à noter que la représentation par un arbre phylogénétique sous-entend que chaque spéciation peut être identifiée comme une divergence à partir d'un ancêtre commun. Dans ce contexte, on néglige donc toutes formes d'hybridation, d'introgression [Arnold 1992 ; Mallet et al. 2016] ou autres transferts latéraux de matériel génétique [Martin 1999 ; Keeling et Palmer 2008]. L'utilisation de réseau phylogénétique permet de visualiser ce type d'événements mais ils ne sont, à ma connaissance, que rarement inféré à grande échelle évolutive.

Inférence phylogénétique

Il existe plusieurs types de méthodes pour inférer une phylogénie à partir d'un alignement de séquences : les méthodes de distance, les méthodes de parcimonie et les méthodes probabilistes. Les trois recherchent l'arbre qui explique le mieux les données (i.e. l'alignement) en suivant (ou non) des critères d'optimisation différents. Les méthodes de distance créent l'arbre en utilisant une matrice de distances obtenue en transformant l'alignement de séquences. La parcimonie cherche à minimiser le nombre de changements nécessaires pour expliquer l'alignement. Les méthodes probabilistes cherchent à trouver l'ensemble des paramètres (incluant la topologie) qui maximise la probabilité d'observer les données (vraisemblance). En théorie, les trois méthodes devraient estimer leur critère sur toutes les topologies possibles (et toutes les longueurs de branche pour les méthodes de distance et probabilistes, et toute les valeurs de paramètres pour les méthodes probabilistes). Comme il n'est pas possible de parcourir toutes les possibilités, on utilise un algorithme de recherche qui trouvera dans un temps raisonnable un optimum, en espérant que celui-ci soit l'optimum global. Ces algorithmes heuristiques utilisent souvent un point de départ sélectionné aléatoirement, puis optimisent la topologie par des réarrangements (locaux ou globaux, e.g. NNI, SPR et TBR) (Figure 4). On utilise d'ailleurs souvent plusieurs points de départ (replica) afin de vérifier que les optimum trouvés correspondent bien à l'optimum global (voir [Felsenstein 2004] pour plus de détails). En parcimonie, un score (nombre minimum de substitutions) est donc calculé à chaque déplacement dans l'espace des possibles et oriente le parcours vers les valeurs les plus basses. Cependant, cette méthode repose sur une supposition très forte du fonctionnement du processus d'évolution

pouvant mener à des inconsistances [Felsenstein 1978a].

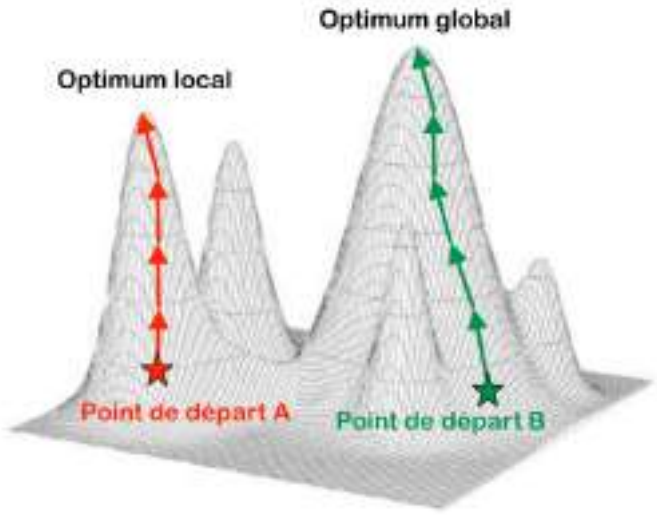


Figure 4 – Déplacement dans un espace multidimensionnel afin de déterminer l'optimum global.

On distingue deux types de méthodes probabilistes, le maximum de vraisemblance et l'inférence bayésienne. Elles utilisent les mêmes modèles d'évolution (à l'exception de l'intégration de la prior) et se basent toutes les deux sur une mesure de vraisemblance (probabilité). Cependant, cette dernière n'est pas employée pour rechercher le même optimum. Le maximum de vraisemblance tente de maximiser la probabilité suivante : L = Pr(D|M,θ,τ,ν) c'est-à-dire la probabilité d'observer les données (D) en fonction d'un modèle (M) ayant pour paramètres : les paramètres de nuisance (θ), une topologie (τ) et les longueurs de branche (ν). En pratique, on s'intéresse plutôt au logarithme de cette probabilité (pour des raisons de précision) et on suppose généralement que l'évolution de chaque site est indépendante. On peut donc calculer la valeur de vraisemblance comme la somme des logarithmes de la probabilité à chaque position de l'alignement. Le but est de parcourir l'espace afin de trouver les valeurs de θ, τ et ν c'est-à-dire l'hypothèse (H) qui maximise cette valeur.

En inférence bayésienne, le problème est pris dans un autre sens car on recherche la probabilité d'observer a posteriori une hypothèse H en connaissant les données (Pr(H|D)). Selon le théorème de Bayes, cette probabilité (posterior probability) peut être obtenu par la formule suivante : Pr(H|D)= Pr(D|H)*Pr(H)/Pr(D) dans laquelle nous retrouvons Pr(D|H), la vraisemblance et Pr(H), une distribution de probabilité a priori d'observer les différentes hypothèses, aussi appelé la prior. Le calcul de la probabilité Pr(D) implique des sommations et intégrations complexes rendant le problème non-réalisable directement. On utilise donc l'algorithme MCMC (Markov Chain Monte Carlo) qui va constituer une chaîne servant de support au parcours de l'espace de paramètres du modèle, et, comme on

se contente de comparer les rapports de probabilités obtenues pour deux séries de valeurs de paramètres, Pr(D) s'annule et n'a donc pas besoin d'être calculée. Chaque élément de la chaîne va constituer un pas se rapprochant de la meilleur approximation de la distribution des probabilités postérieures. Dans les faits, on utilise au moins deux chaînes différentes avec pour objectif qu'elles convergent toutes les deux vers la même approximation (voir [Felsenstein 2004] pour plus de détails sur le fonctionnement des diverses méthodes d'inférence).

Les bases des modèles d'évolution de séquences

Les méthodes probabilistes tentent de modéliser avec le meilleur compromis fidélité/complexité les processus évolutifs ayant généré les séquences moléculaires analysées. Cette modélisation se concentre généralement sur le processus mutationnel (ou plutôt substitutionnel) par lequel un résidu est remplacé par un autre dans la séquence de nucléotides ou d'acides aminés. Ces remplacements sont représentés par des probabilités de passer d'un état à un autre sans prendre en compte l'impact des précédents remplacements. On utilise un processus de Markov d'ordre 0 pour modéliser le processus substitutionnel en supposant que celui-ci est homogène, stationnaire et réversible, c'est-à-dire que les taux d'échange entre résidus et les fréquences des résidus sont constants au cours du temps, et que le taux d'échange instantané pour passer de A à B est équivalent à celui de B vers A, respectivement [Liò et Goldman 1998 ; Whelan et al. 2001].

Il existe deux catégories de modèles, les modèles empiriques et les modèles paramétriques. Dans les premiers, les valeurs des paramètres du modèle sont fixes et déterminées a priori en analysant un autre jeu de données, alors que, dans les seconds, elles sont déterminées lors de l'inférence. Comme le nombre d'états possibles est plus important dans un jeu de données protéiques, on préfère réaliser l'inférence avec des modèles empiriques, comme WAG [Whelan et Goldman 2001] ou LG [Le et Gascuel 2008], moins gourmands en ressources. Les modèles paramétriques sont en général réservés aux jeux de données nucléotidiques. Dans les deux cas, on différencie deux éléments : le taux instantané d'échangeabilité entre résidus et leur fréquence à l'équilibre.

En se contentant des paramètres de modèle présentés jusqu'à maintenant, on considère que le taux auquel se fixent les mutations est identique pour chaque position. Cette hypothèse est bien évidemment fausse car on observe facilement dans un alignement de séquences que certaines zones restent hautement conservées alors que d'autres ont beaucoup divergé [Uzzell et Corbin 1971]. La solution la plus courante pour modéliser cette hétérogénéité du taux de substitution entre sites consiste à discrétiser une distribution gamma en plusieurs catégories (généralement 4) et d'y attribuer les différents sites [Yang

1994]. La prise en compte de cette hétérogénéité de taux fait partie des ajouts les plus communs pour compléter les modèles classiques.

Améliorer la modélisation l'hétérogénéité du processus substitutionnel

En considérant les bases des modèles d'évolution présentées jusqu'à présent, on suppose une évolution homogène et stationnaire le long de l'arbre du vivant. Cependant, de nombreuses études ont démontré que ce processus était bien plus complexe que les hypothèses se trouvant derrière ces premiers modèles [Lockhart et al. 1992 ; Galtier et Gouy 1995 ; Mooers et Holmes 2000 ; Lopez et al. 2002 ; Lartillot et Philippe 2004 ; Roure et Philippe 2011]. On peut citer comme type d'hétérogénéité, l'hétérogénéité de composition (affectant les fréquences stationnaires le long de l'arbre) [Lockhart et al. 1992 ; Galtier et Gouy 1995 ; Mooers et Holmes 2000], l'hétérotachie ou hétérogénéité de taux au cours du temps [Lopez et al. 2002 ; Philippe et al. 2005], l'hétérogénéité du processus substitutionnel entre sites [Lartillot et Philippe 2004] et au cours du temps (hétéropécillie) [Roure et Philippe 2011].

Certaines de ces hétérogénéités ont été implémentées dans des modèles pouvant être utilisés pour réaliser des inférences phylogénétiques [Galtier et Gouy 1998 ; Galtier 2001 ; Foster 2004 ; Groussin et al. 2013 ; Crotty et al. 2017]. L'hétérogénéité du processus substitutionnel par site a notamment eu droit à plusieurs implémentations différentes via des modèle de mélange telles que LG4X et LG4M [Le et al. 2012], les modèles C20 à C60 [Si Quang et al. 2008] ou le modèle CAT [Lartillot et Philippe 2004]. Ce type d'hétérogénéité est d'ailleurs l'un des rares à être considéré dans les phylogénies à grande échelle évolutive. En effet, bien que l'existence, et dans une moindre mesure, l'impact de ces différentes violations de modèles soient connus, l'utilisation d'un modèle les prenant tous en compte reste, pour l'instant, hors de portée d'un point de vue computationnel. Nous reparlerons en fin d'introduction de ce problème et de l'impact des modèles d'évolution de séquences utilisés en phylogénie.

L'endosymbiose à l'origine de la propagation de la photosynthèse

L'évolution des eucaryotes fut grandement influencée par les concentrations en dioxygène de leur environnement [Knauth et Kennedy 2009 ; Lenton et al. 2014 ; Mills et Canfield 2014]. Deux voies métaboliques importantes en contrôlent la production et

la consommation, respectivement la photosynthèse et la respiration aérobie. Celles-ci ont un point commun dans le domaine des eucaryotes. En effet, elles se déroulent toutes les deux au sein de sous-structures propres situées dans le cytoplasme des cellules, les organites. Parmi l'ensemble des organites présents chez les eucaryotes, les deux qui nous intéressent, les plastes et les mitochondries, présentent une complexité particulière liée à leur origine évolutive. Celles-ci sont expliquées par un phénomène biologique s'appellant l'endosymbiose.

L'endosymbiose, comme la symbiose, est l'interaction mutuellement bénéfique entre deux protagonistes, un organisme hôte et un organisme symbiotique. La particularité de l'endosymbiose est l'intégration du symbionte (ici appelé endosymbionte) au sein même des cellules de l'hôte. Ce point la démarque de l'ectosymbiose dans laquelle le symbionte vit sur les tissus de l'hôte. Cette interaction particulières entre deux organismes fut proposée par Konstantin Mereschkowski comme étant à l'origine des organites, à cause des similarités structurales entre les plastes des plantes et certaines bactéries. Ce n'est que plus tard que Lynn Margulis formula la théorie de l'évolution par endosymbiose avec un procaryote en présentant des preuves biologiques [Sagan 1967].

Photosynthèse oxygénique : origine et diversité

Origine cyanobactérienne

Les cyanobactéries furent les premiers organismes capables de réaliser la photosynthèse oxygénique. Ce groupe de bactéries Gram-Négatif contient plus de 2000 espèces décrites, réparties dans 150 genres [Nabout et al. 2013]. Elles sont présentes aussi bien dans les milieux aquatiques (marin et d'eau douce) que terrestres. Deux nouveaux groupes de bactéries non-photosynthétiques (Melainabacteria et Sericytochromatia) ont été récemment découverts grâce à la méta-génomique et affiliés aux cyanobactéries [Di Rienzi et al. 2013 ; Soo et al. 2014 ; Soo et al. 2017]. Les Sericytochromatia se positionnent comme groupe-frère de l'ensemble Melainabacteria et cyanobactéries photosynthétiques, ces deux clades étant également monophylétiques. Malgré l'intégration de ces deux clades, le nom cyanobactérie reste généralement associé aux bactéries photosynthétiques maintenant renommées par certains Oxyphotobacteria [Soo et al. 2014] ou encore Cyanophyceae. Jusque dans les années 60 et la découverte de leurs traits procaryotiques, les études morphologiques positionnaient ce sous-ensemble parmi les eucaryotes [Percival et Williams 2014]. En effet, l'aspect filamenteux et visqueux de leur regroupement colonial leur donne l'apparence d'algues, ce qui mena à leur ancienne dénomination d'algues bleues ou bleu-vert, couleur propre à la chlorophylle a.

La plupart de ces bactéries photosynthétiques se caractérisent par une organisation en deux parties : une partie centrale contenant le matériel génétique, et une partie périphérique dans laquelle se concentrent les thylakoïdes (à l'exception du genre Gloeobacter qui n'en possède pas [Rippka et al. 1974]). C'est dans ces derniers que l'on trouve les pigments photosynthétiques tels que la chlorophylle a ou des pigments annexes comme la phycocyanine (bleu) et la phycoérythrine (rouge). Ils sont regroupés en complexes protéiques nommés phycobilisomes captant l'énergie lumineuse lors de la phase claire de la photosynthèse. Durant cette phase, l'énergie captée fait passer les molécules dans un état excité. Cet état se transmet de pigment en pigment à travers les photosystèmes I et II jusqu'à être converti en énergie chimique par la perte d'un électron de la chlorophylle a du centre réactionnel. Cet électron est récupéré en réalisant la photolyse de l'eau produisant le dioxygène (O2). L'énergie chimique produite durant la phase claire servira ensuite, lors de la phase sombre, à fixer le carbone du dioxyde de carbone (CO_2) dans des molécules énergétiques (Cycle de Calvin).

Apparition chez les eucaryotes

Les Cyanophyceae ont, pendant longtemps, été les seuls organismes à réaliser la photosynthèse oxygénique [Falkowski et al. 2004]. Mais un jour, il y a plus d'un milliard d'années [Yoon et al. 2004], une d'entre elles rencontra une cellule eucaryote, disposant sûrement déjà d'un noyau et d'une mitochondrie. C'est à partir de cette rencontre, et de l'endosymbiose qui en découle, qu'une partie des organismes eucaryotes furent également capables de réaliser la photosynthèse [Sagan 1967]. Cette endosymbiose est appelée primaire car elle constitue l'acquisition originelle du plaste à partir d'un procaryote. Les descendants de cet espèce forment les Archaeaplastida, un groupe d'eucaryotes photosynthétiques. Ce dernier est lui-même subdivisé en trois ensembles d'organismes : les glaucophytes, les plantes vertes et les algues rouges. Le point commun de ces trois groupes est d'avoir des plastes délimités par deux membranes plasmiques. De plus, les phylogénies basés sur les gènes plastidiaux supportent leur monophylie et leur origine au sein des cyanobactéries [Criscuolo et Gribaldo 2011 ; Mackiewicz et Gagat 2014 ; Ponce-Toledo et al. 2017]. Ces divers éléments sont en faveur d'une acquisition unique de la photosynthèse chez un des ancêtres de ces organismes.

Les glaucophytes (Glaucophyta) représentent le groupe le moins riche en espèces parmi les Archaeplastida (Figure 5). En effet, seuls 4 genres sont décrits pour un total de 15 espèces. Celles-ci sont des photoautotrophes strictes vivant en eau douce. Leur plaste a pour particularité d'avoir gardé de nombreux traits similaires à leur ancêtre cyanobactérien comme la présence d'une paroi de peptidoglycane ou une organisation des photosystèmes en phycobilisomes.

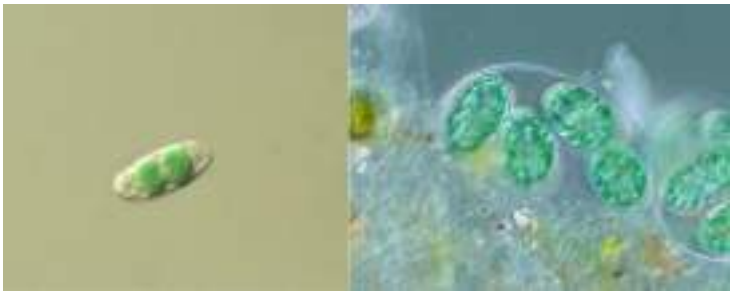


FIGURE 5 – *Cyanophora paradoxa* et *Glaucocystis sp.*

Les plantes vertes (Viridiplantae) représentent probablement le groupe d'organismes le plus associé à la photosynthèse. Il se subdivise en deux clades : les chlorophytes, comprenant la majorité des algues vertes, et les streptophytes, composées de quelques groupes groupes d'algues parmi lesquelles sont apparues les plantes terrestres [McCOURT et al. 2004 ; LAURIN-LEMAY et al. 2012 ; WODNIOK et al. 2011 ; LELIAERT et al. 2012]. Les algues vertes et les plantes terrestres possèdent des thylakoïdes lamellaires groupés en grana, contenant de la chlorophylle a et b ainsi que d'autres pigments accessoires comme les carotènes et les xanthophylles. Leurs plastes peuvent également contenir un pyrénoïde, un compartiment dans lequel le CO2 est concentré pour en faciliter la fixation par l'enzyme RuBisCO. Ces organismes ont l'amidon pour principale réserve de polysaccharides et leur cellule possède généralement une paroi constituée de cellulose. Malgré ces nombreux point communs, les plantes vertes manifestent une grande diversité écologique et morphologique. Ainsi, on y retrouve à la fois des algues unicellulaires appartenant au picoplancton que d'énormes végétaux terrestres (Figure 6).

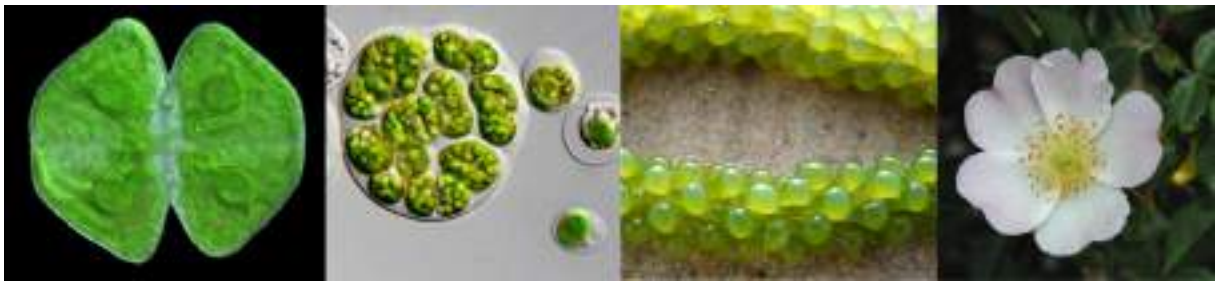


FIGURE 6 – *Cosmarium sp.* , *Chlamydomonas augustae*, *Caulerpa uva* et *Rosa canina*

Les algues rouges (Rhodophyta) forment le dernier groupe d'Archaeaplastida. La majorité d'entre elles sont des organismes pluricellulaires marins benthiques. Cependant, on peut également les retrouver en eau douce, sous des formes unicellulaires dans le plancton ou encore sur la terre ferme (Figure 7). Parmi les groupes d'unicellulaires, celui des cyanidiales est composé d'organismes vivant en milieu extrême (source chaude ou milieu pollué aux métaux lourds). L'ensemble de pigments des algues rouges se compose de chlorophylles a, de phycoérythrines et de caroténoïdes, ces deux derniers leur donnant leur couleur caractéristique. Leurs cellules possèdent généralement une paroi plutôt épaisse constituée

de cellulose et de galactanes sulfatés. Ces derniers les rendent notamment très intéressantes d'un point de vue économique. En effet, les algues rouges sont cultivées, soit directement comme nourriture (le genre *Pyropia* correspond au nori des sushi), soit pour récolter ces molécules de galactanes sulfatés (les agars et les carraghénanes sont parmi les plus connus) qui ont de nombreux usages dans les industries pharmaceutiques et alimentaires.

FIGURE 7 – *Cyanidium sp.*, *Gracilaria sp.*, *Botryocladia pseudodichotoma*, *Chondracanthus acicularis*

Un cas à part chez le genre *Paulinella*

Les Archaeplastida sont apparus suite à une endosymbiose primaire avec une cyanobactérie il y a plus d'un milliard d'années [YOON et al. 2004]. Cependant, une autre endosymbiose primaire a eu lieu plus tard entre un autre eucaryote et une autre cyanobactérie, donnant naissance à *Paulinella chromatophora* (Figure 8) [KEELING et ARCHIBALD 2008 ; NOWACK et al. 2008 ; YOON et al. 2009 ; NAKAYAMA et ARCHIBALD 2012 ; GRAY et JOHN M. ARCHIBALD 2012]. Le cas de *P. chromatophora* est considéré comme celui d'une endosymbiose en cours de réalisation. Cette organisme est donc très important pour l'étude de la genèse d'un organite par endosymbiose. Cependant, cette espèce ne représente qu'un grain de sable dans la diversité des organismes et ne sera donc pas décrite plus en détails.

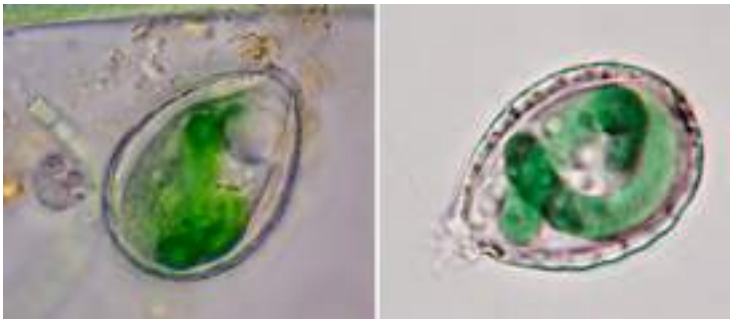


FIGURE 8 – *Paulinella chromatophora*(source : http ://arcella.nl)

Diversification au sein des eucaryotes

Suite à la diversification des Archaeplastida, la photosynthèse n'est pas restée cantonnée à ce groupe restreint d'eucaryotes. Après l'endosymbiose primaire, d'autres endosymbioses, dites complexes, ont eu lieu entre eucaryotes. En effet, ces dernières ont transformé certains eucaryotes photosynthétiques en endosymbiontes, ce qui étendit à nouveau la diversité des eucaryotes capables de réaliser la photosynthèse. On distingue deux types d'endosymbioses complexes, selon l'origine du plaste : les endosymbioses complexes vertes et les endosymbioses complexes rouges. On reconnaît les organismes ayant acquis leur plaste par une endosymbiose complexe par le nombre de membranes entourant ces derniers, toujours supérieur à deux.

Endosymbioses complexes vertes

Deux groupes d'eucaryotes photosynthétiques principaux arborent des plastes montrant des liens de parenté avec ceux des plantes vertes. Ces groupes sont ceux des euglènes et des chlorarachniophytes. Ces organismes possèdent des plastes contenant de la chlorophylle a et b, leur donnant leur couleur verte. Malgré cette caractéristique commune, les phylogénies basées sur des gènes chloroplastiques montrent que leurs plastes ne forment pas un groupe monophylétique, suggérant deux acquisitions distinctes de la photosynthèse [Rogers et al. 2007 ; Jackson et al. 2018].

Seul un sous-clade monophylétique d'euglènes (Euglenophyceae) possède la capacité de photosynthèse. Ces organismes sont des photoautotrophes unicellulaires, généralement munis d'un seul flagelle (rarement deux)(Figure 9). On les retrouve principalement en eau douce et dans les eaux saumâtres où ils s'y déplacent en nageant [Esson et Leander 2008]. Certaines espèces (e.g. Euglena sanguinea) sont à l'origine de bloom relargant des neurotoxines affectant les cultures de poissons d'eau douce [Zimba et al. 2010]. Leur plaste est entouré de trois membranes et contient des thylakoïdes toujours groupés par trois. Les euglènes possèdent également un dispositif de photoréception composé d'un micro-flagelle et d'une structure ombrageante formé de caroténoïdes leur permettant de s'orienter dans la colonne d'eau [Kuznicki et al. 1990].

Le groupe des Chlorarachniophytes est relativement petit, se composant de 14 espèces décrites réparties en 8 genres [Ishida et al. 2007]. Ce sont des organismes unicellulaires marins, à la fois phototrophes et mixotrophes, présents en climat tropical à tempéré. Ils possèdent trois stades cellulaires différents : un stade cellulaire amoeboïde présentant des réseaux de pseudopodes, un stade sphérique à paroi épaisse faisant penser à une forme d'enkystement et un stade cellulaire mobile possédant un flagelle (Figure 10). Ils ont comme grande particularité de posséder des plastes associés à un noyau réduit nommé

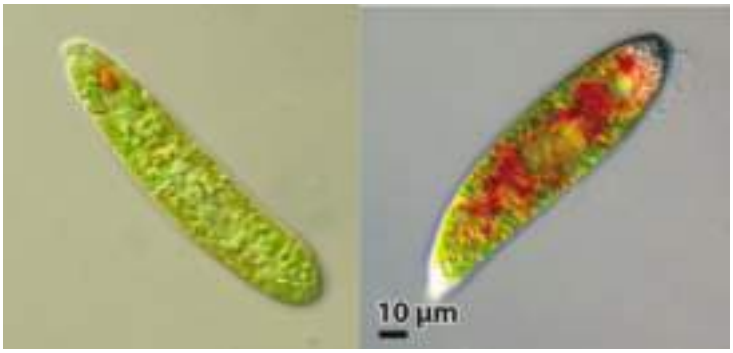


FIGURE 9 – *Euglena gracilis* et *Euglena sanguinea*

nucléomorphe, vestige du noyau de l'endosymbionte eucaryote. Chaque nucléomorphe étudié contient trois chromosomes de petite taille [ISHIDA et al. 2011]. Les chlorarachniophytes possèdent de un à plusieurs plastes entourés de quatre membranes, chacun étant associé à son propre nucléomorphe. Ce groupe est principalement étudié comme modèle d'acquisition de la photosynthèse par endosymbiose secondaire.

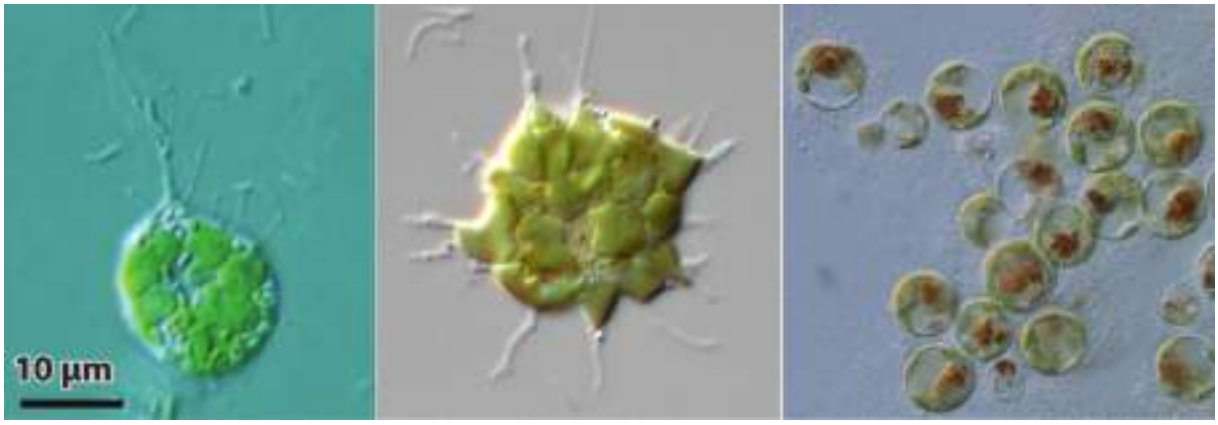


FIGURE 10 – *Chlorarachnion reptans*, *Gymnochlora stellata*, *Lotharella vacuolata*

Endosymbioses complexes rouges

Comparé à la faible diversité d'euglènes photosynthétiques et de chlorarachniophytes, de nombreux organismes possèdent un plaste affilié aux algues rouges. Ces organismes sont répartis en cinq groupes distincts que les phylogénies plastidiales considèrent monophylétiques (à une exception) [YOON et al. 2002]. Leurs plastes ont généralement pour caractéristique d'être entourés de quatre membranes, la plus extérieure étant reliée au réticulum endoplasmique. Ils contiennent généralement une troisième forme de la chlorophylle, la chlorophylle c.

Les cryptophytes (Cryptophyta) sont des unicellulaires biflagellés de petite taille (5-50µm) reconnaissables par l'asymétrie marquée de leur cellule (Figure 11). Ils possèdent notamment un organelle extrusif nommé éjectosome. Ils sont présents en grande quantité dans

les milieux aquatiques, que ce soit en mer ou en eau douce [Klaveness 1988], où ils sont d'importants composants du réseau trophique [Tirok et Gaedke 2007]. Leur plaste contient des biliprotéines formant des complexes collecteurs de lumières secondaires présents dans les thylakoides. Ces derniers permettent aux cryptophytes d'absorber des lumières de faible intensité [Hammer et al. 2002] et appartenant à un spectre d'absorption non couvert par les chlorophylles des autres algues [Doust et al. 2006]. Tout comme les chlorarachniophytes, l'espace intermembranaire du plaste contient un nucléomorphe affiliant l'endosymbionte aux algues rouges [Moore et al. 2012].

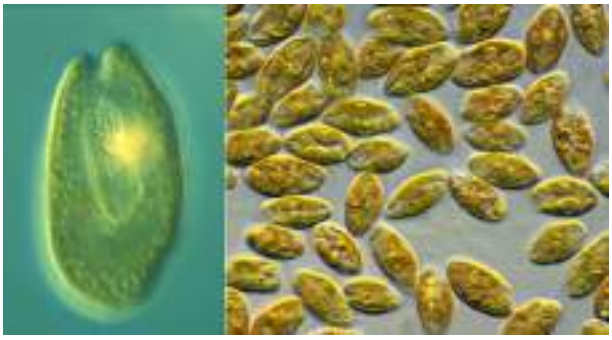


Figure 11 – *Cryptomonas sp.* et textitRhodomonas sp.

Les haptophytes (Haptophyta) sont des unicellulaires majoritairement phototrophes vivant principalement en milieu marin (Figure 12) où ils constituent une part importante du nanoplancton [Masquelier et al. 2011] et du picoplancton [Kirkham et al. 2011]. Ils sont notamment des acteurs essentiels du cycle du carbone grâce à leurs capacités de photosynthèse et de production d'écailles calcifiées. Ces dernières recouvrent d'ailleurs l'entièreté de l'organisme chez le sous-groupe des coccolithophores. Les haptophytes possèdent également deux flagelles ainsi qu'un organite filiforme positionné entre ces derniers, l'haptonème. Certaines espèces forment des blooms, parfois toxiques. Ceux-ci peuvent couvrir des distances supérieures à 200.000 km2 et comprendre des densités cellulaires allant jusqu'à 6 millions de cellules par litre [Sukhanova et Flint 1998]. Les haptophytes arborent deux plastes affichant une couleur brun-doré liés à la grande variété de pigments qu'ils possèdent [Lenning et al. 2004 ; Zapata et al. 2004].

Les ochrophytes constituent un groupe très diversifié d'algues photosynthétiques regroupant plus de 10 classes d'organismes (Figure 13). Ils sont majoritairement unicellulaires et se trouvent aussi bien en milieu aquatique que terrestre. On retrouve notamment les algues brunes (Phaeophyceae), les eustigmatophytes (Eustigmatophyceae) et les diatomées (Bacillariophyceae). Les premières constituent le seul groupe d'organismes multicellulaires parmi les ochrophytes. Elles sont essentiellement retrouvées en milieu côtier où elles peuvent atteindre des tailles de 50 mètres, formant des environnements semblables à des forêts [Dayton 1985]. Elles sont également intéressantes économiquement comme sources d'alginates utilisés comme gélifiant, notamment dans l'industrie alimentaire, ou

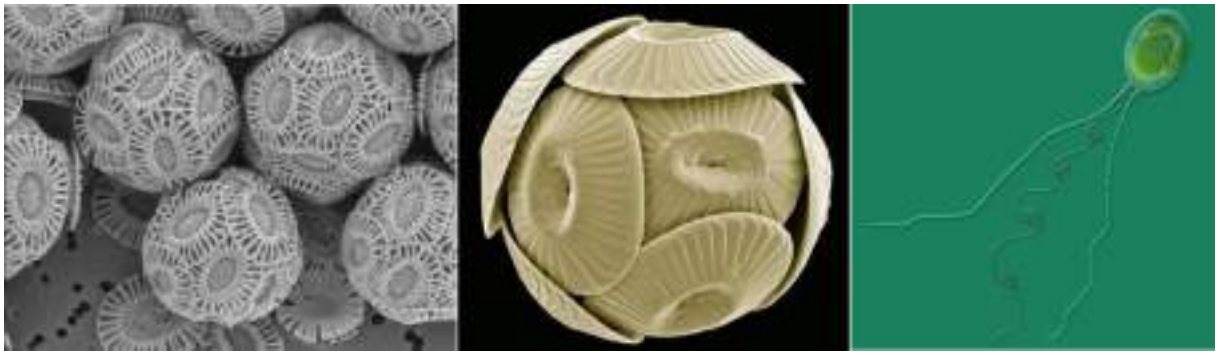


FIGURE 12 – *Emiliana huxleyi*, *Coccolithus pelagicus* et *Chrysochromulina sp.*

directement comme nourriture, principalement dans les pays d'Asie. Les eustigmatophytes sont moins connues du grand public mais très étudiées en bioindustrie. Leur importante capacité de stockage des acide gras (ces derniers pouvant atteindre 50% de leur poids sec) en font de bonnes candidates pour la production de carburant [MA et al. 2016]. Finalement, les diatomées forment le groupe d'algues photosynthétiques le plus riche en espèces avec une diversité estimée à plus de 100.000 espèces [MANN et VANORMELINGEN 2013]. Elles sont facilement reconnaissables par leur paroi cellulaire riche en silice (frustule) formant deux plaques distinctes. Elles constituent un groupe très important d'un point de vue écologique car elle pourrait représenter jusqu'à 20% de la fixation de carbone liée à la photosynthèse [Mann et al. 2017].

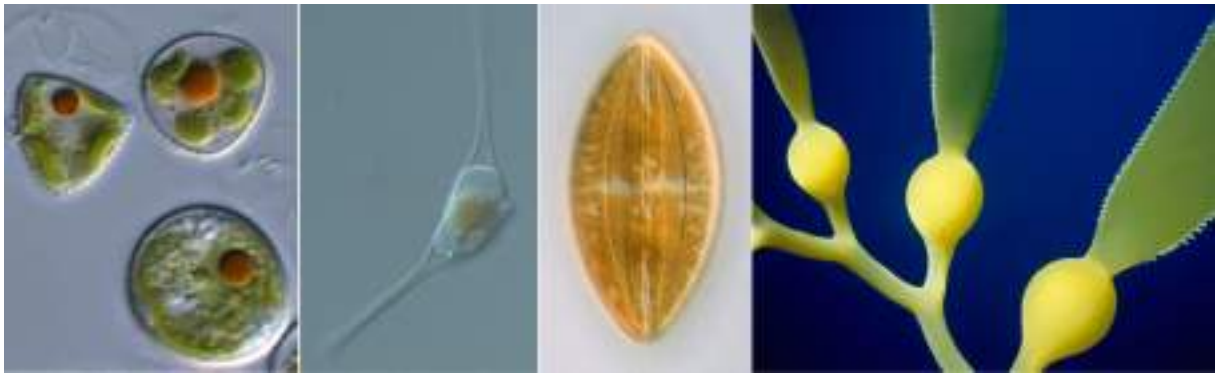


FIGURE 13 – Eustigmatophytes, Chrysophytes, *Lyrella hennedyi* (Bacillariophyceae), *Macrocystis sp.*(Phaeophyceae)

Les dinoflagellés forment un autre groupe très important du plancton marin (Figure 14). Ce groupe comprend plus de 2400 espèces décrites, la majorité vivant en milieu marin [GÓMEZ 2012]. Parmi ces espèces, environ la moitié pratique la photosynthèse. Elles sont notamment connues pour les marées rouges qu'elles forment lors de la réalisation de bloom généralement toxique [CEMBELLA 2003]. Les biotoxines qu'elles génèrent sont d'ailleurs une des principales sources d'empoisonnement lié à l'ingestion de produits marins [Lehane et Lewis]. Les dinoflagellés photosynthétiques sont à l'origine de nombreuses relations symbiotiques avec les animaux. Ces relations aident notamment à la création des

coraux, l'appauvrissement en CO_2 résultant de la photosynthèse favorisant le dépôt des carbonates de calcium [Marshall 1996]. Leurs plastes ne présentent pas systématiquement quatre membranes comme pour les autres photosynthétiques complexes rouges, leur nombre pouvant varier de 2 à 5. Cette variabilité s'explique par la grande diversité d'origine évolutive des plastes de dinoflagellés. En effet, certains genres arborent des plastes affiliés aux haptophytes [Nosenko et al. 2006], aux algues vertes [Minge et al. 2010], ou aux diatomées [Imanian et al. 2010]. Cependant, une étude phylogénomique supporte que l'ancêtre des dinoflagellés actuels était photosynthétique et que son plastide était proche de celui des chroméridés et à l'apicoplaste des apicomplexes (voir plus bas) [Janouškovec et al. 2015]. Ces plastes d'origine ancestrale contiennent un pigment de la famille des caroténoïdes propre aux dinoflagellés, la péridinine. L'existence de dinoflagellés possédant des plastes d'autres origines, ainsi que l'observation de gènes originaires d'autres organismes photosynthétiques dans leur génome, suggèrent que les dinoflagellés sont capables de remplacer leur plaste relativement facilement. Les différentes espèces ne réalisant plus la photosynthèse sont d'ailleurs capables de récupérer cette capacité de manière transitoire en conservant leur proie photosynthétique sur de longues périodes (cleptoplastidie, [Johnson 2011 ; Bodył 2018]).

Figure 14 – *Karenia brevis*, *Polykrikos lebourae*, *Gymnodinium catenatum*, *Lingulodinium polyedrum* (source : http ://www.tol-web.org)

Les chroméridés (Chromerida) forment un groupe récemment découvert d'algues photosynthétiques. Actuellement, seulement deux espèces ont été décrites, *Chromera velia* [Moore et al. 2008] et *Vitrella brassicaformis* [Oborník et al. 2012] (Figure 15). Ces deux organismes ont été isolés à partir des tissus des coraux *Plesiastrea versipora* et *Leptastrea purpurea* respectivement. Les chroméridés sont très intéressants d'un point de vue évolutif car ils sont affiliés aux apicomplexes, un groupe de parasites unicellulaires liés à des maladies connues comme la toxoplasmose ou la malaria. Ces derniers possèdent généralement un plaste non photosynthétique, l'apicoplaste, qui suggère que leur ancêtre était photosynthétique [Janouškovec et al. 2015]. Dans la phylogénie nucléaires des eucaryotes, les apicomplexes et les chroméridés forment un groupe monophylétiques, lui-même groupe-frère des dinoflagellés. On retrouve donc un clade contenant un mélange complexe

d'organismes photosynthétiques et non-photosynthétiques. Ceci suggère que l'histoire de l'acquisition de la photosynthèse est plus complexe que ce qui est raconté par les plastes ...

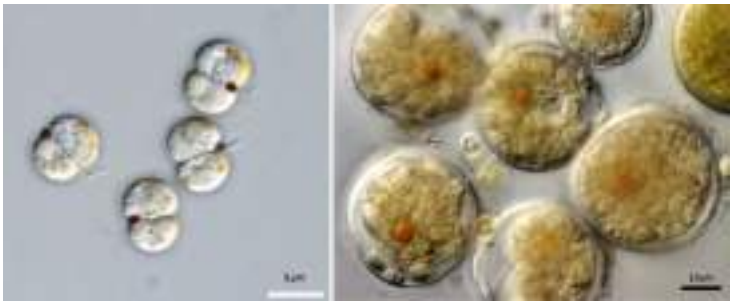


FIGURE 15 – *Chromera velia* et *Vitrella brassicaformis*

Intégration de la photosynthèse à la phylogénie des eucaryotes

La grande diversité des eucaryotes photosynthétiques a été présentée dans la précédente partie. Toutefois, comment se positionne cette diversité au sein de l'arbre du vivant et comment la photosynthèse est-elle apparue parmi ces groupes ? Pour répondre à ces questions, il faut mettre en avant la double nature du phénomène. En effet, l'histoire évolutive de la photosynthèse correspond au croisement de deux histoires distinctes. La première correspond à l'histoire des endosymbiontes et est liée à l'évolution des génomes plastidiaux. Elle m'a servi de guide pour présenter les grands sous-ensembles d'organismes photosynthétiques mais, si son début a pu paraître simple, elle se complexifia rapidement. La seconde se focalise sur l'évolution des hôtes et dépend de l'évolution de leur génome nucléaire et mitochondrial. C'est le croisement de ces deux histoires qui permet d'étudier la transmission de la photosynthèse au sein des eucaryotes. Pourtant, l'histoire des hôtes demeure plus difficile à résoudre et est souvent en contradiction avec celle des symbiontes. Ce dernier point laisse supposer un historique de transmission très complexe pour la photosynthèse.

Comme nous l'avons vu, l'histoire de l'acquisition de la photosynthèse chez les eucaryotes commence avec les photosynthétiques primaires, qui possèdent tous des plastes entourés de deux membranes issus d'une cyanobactérie. Ce trait partagé fit supposer leur origine monophylétique, et ce bien avant les premières phylogénies moléculaires. Cependant, ces dernières ne supportaient en général pas cette hypothèse, surtout quand elles étaient inférées avec des marqueurs d'origine nucléaire [BHATTACHARYA et al. 1995 ; BHATTACHARYA et WEBER 1997 ; KEELING et al. 1999]. Avec les débuts de la phylogénomique

et la concaténation de marqueurs (i.e. supermatrice), la monophylie des algues rouges, plantes vertes et glaucophytes devint plus fréquemment supportée [Baldauf et al. 2000 ; Moreira et al. 2000 ; Rodríguez-Ezpeleta et al. 2005]. Il fut estimé qu'un minimum de 15000 positions était nécessaire pour retrouver ce groupe monophylétique avec un support raisonnable [Rodríguez-Ezpeleta et al. 2005], expliquant l'échec des phylogénies à un seul gène.

Au début de la phylogénomique, la quantité de données moléculaires était encore faible, surtout chez les protistes (e.g. eucaryotes unicellulaires), ce qui limitait le nombre de clades représentés dans les arbres. Une phylogénie classique des eucaryotes (Figure 16) présentait d'un côté, un groupe constitué des champignons, des choanoflagellés et des animaux (Opisthokonta), souvent associé aux Amoebozoa, et de l'autre, un ensemble hétérogène de clades. On y retrouvait généralement les Archaeplastida ainsi que les alvéolés (ciliés+apicomplexes+dinoflagellés) et les stramenopiles, mais d'autres clades pouvaient s'y intercaler commes les euglènes [Baldauf et al. 2000] ou les haptophytes et les cryptophytes [Stibitz et al. 2000]. Cet ensemble monophylétique d'eucaryotes regroupait donc la majorité des organismes photosynthétiques connus ainsi que d'autres organismes non-photosynthétiques. La présence de la photosynthèse chez ces organismes impliquait donc de multiples gains et/ou pertes de cette capacité. La diversité des photosynthétiques complexes verts fut rapidement attribuée à plusieurs gains indépendants mais l'histoire des complexes rouges est plus difficile à résoudre.

Comme mis en avant lors de la description de la diversité des organismes photosynthétiques, les organismes complexes rouges sont plus divers que leurs homologues verts. Cependant, cela ne les empêche pas de conserver de nombreux points communs. Outre la présence commune de chlorophylle c, ils utilisent également le même système pour faire franchir la seconde membrane externe de leur plaste aux protéines encodées dans le noyau [Felsner et al. 2011]. Ce système fait intervenir un complexe protéique (SELMA) dérivé une seule fois du système de dégradation des protéines du réticulum endoplasmique [Gould et al. 2015]. Ces traits communs sont en faveur de la monophylie des secondaires rouges et ont favorisé la création du groupe Chromalveolata (chromalvéolé, Figure 17) [Cavalier-Smith 1999]. Ce clade découle de l'addition des alvéolés au groupe Chromista, regroupement des haptophytes, cryptophytes et ochrophytes [Cavalier-Smith,1981]. Il donna également naissance à l'hypothèse éponyme supposant un ancêtre commun à tous les secondaires rouges et donc un seul événement d'endosymbiose. Elle se veut être une théorie parcimonieuse basée sur la monophylie plastidiale de ces organismes [Yoon et al. 2002] et la faible probabilité d'établir plusieurs fois un système complexe d'adressage des protéines au plaste [Cavalier-Smith 2003].

Si l'hypothèse chromalvéolée explique les différents points communs aux algues complexes

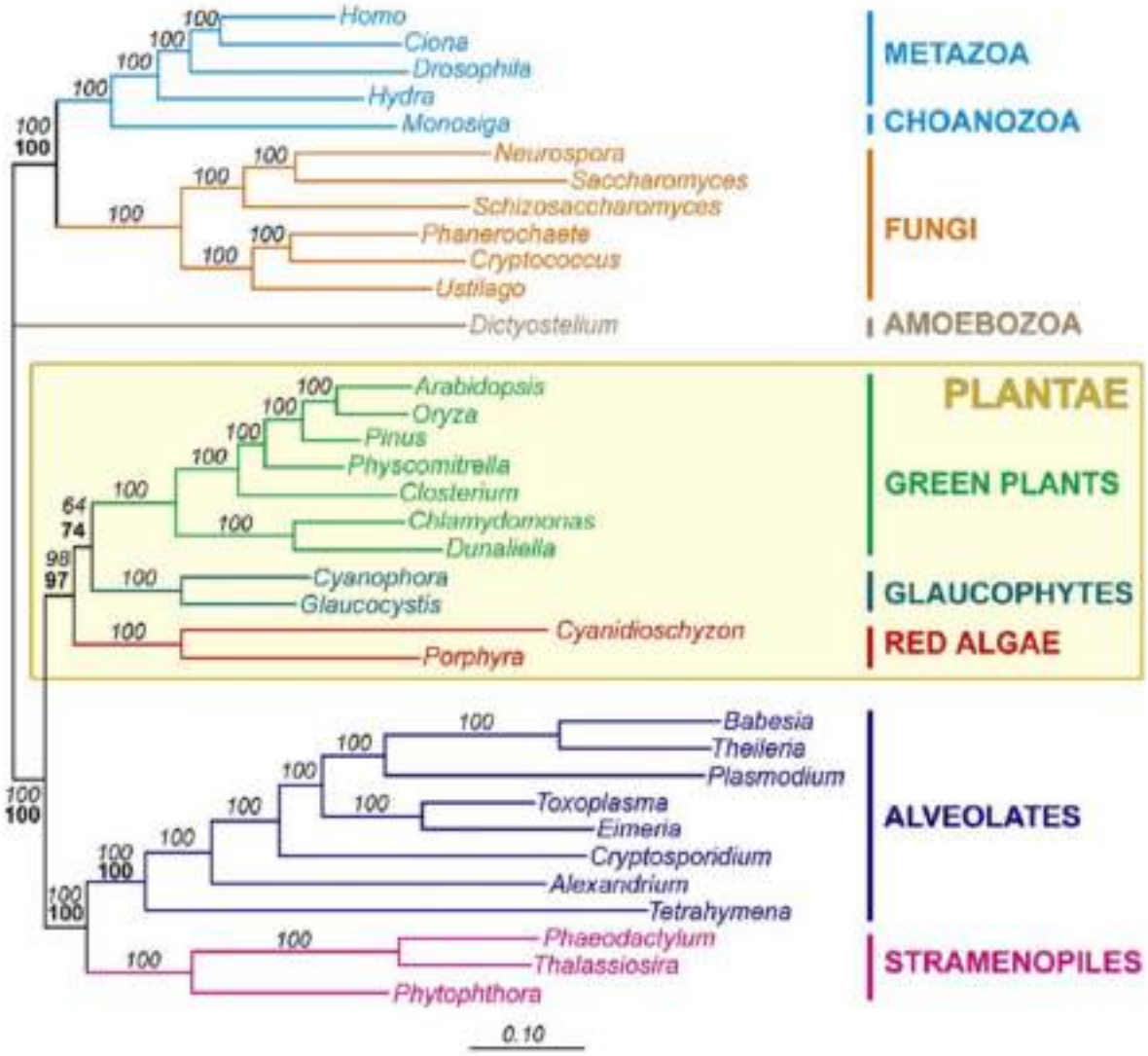


FIGURE 16 – Arbre phylogénétique des eucaryotes inféré à partir d'une concaténation de 143 proteins (30.113 sites). Les nombres en italique représentent les valeurs de support obtenus à partir de 100 réplicats de bootstrap sous PhyML avec le modèle WAG+F+Γ alors que les nombres en dessous (en gras) correspondent aux supports de l'analyse sML avec 10.000 réplicas RELL [RODRÍGUEZ-EZPELETA et al. 2005]

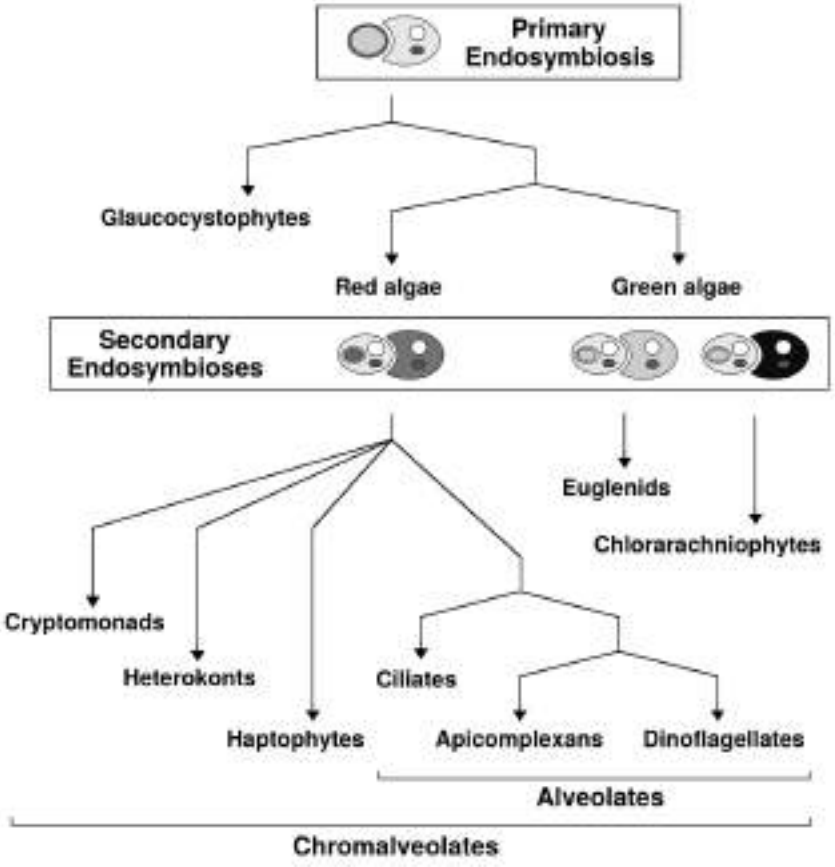


FIGURE 17 – Hypothèse Chromalvéolée)

rouges, elle possède aussi des défauts. En effet, comme nous l'avons déjà vu, le groupement des chromalvéolés contient de nombreuses espèces non-photosynthétiques. L'hypothèse se veut donc parcimonieuse vis-à-vis de l'acquisition de la photosynthèse mais pas de sa perte. Déjà lors de sa proposition, elle impliquait la perte de la photosynthèse chez de nombreux organismes parmi différents groupes, tels que les ciliés (Ciliophora), les apicomplexes (Apicomplexa), une partie des dinoflagellés (Dinophyceae) et la base des stramenopiles (Phytophthora, Blastocystis). La présence de reste de la photosynthèse chez les Apicomplexa (l'apicoplaste, plaste non photosynthétique) allait d'ailleurs dans ce sens bien que les autres clades ne montraient pas de signes aussi probants [Reyes-Prieto et al. 2008]. Malgré cela, elle perdit petit à petit en intérêt suite aux avancées opérées dans la phylogénie des eucaryotes. Celles-ci furent possibles grâce à la combinaison de l'arrivée des nouvelles technologies de séquençage, permettant l'augmentation des données génomiques disponibles (notamment l'apparition du clade Rhizaria), l'amélioration des modèles évolutifs [Lartillot et Philippe 2004 ; Le et Gascuel 2008] et la généralisation de l'usage de la phylogénomique. En effet, le nombre d'organismes non-photosynthétiques s'intercalant dans les chromalvéolés augmenta au fur et à mesure des études permises par ces progrès. Le plus important fut le positionnement du regroupement des foraminifères et des cercozaires dans un groupe nommé Rhizaria en tant que groupe frère des alvéolés et des stramenopiles [Burki et al. 2007]. L'association de ces trois clades en un super groupe nommé SAR est maintenant retrouvée dans la majorité des études phylogénétiques. Il s'ensuit la découverte d'organismes non photosynthétiques affiliés aux haptophytes (Centrohélide) et aux cryptophytes (Palpitomonas et Goniomonas) ainsi que le positionnement instable de ces derniers par rapport au Archaeplastida et au SAR. Ces changements dans la phylogénie des hôtes rendent actuellement l'hypothèse chromalvéolée peu probable et favorisent l'acquisition multiple de la photosynthèse via les algues rouges, comme pour les secondaires verts (Figure 18).

Toutefois, les théories d'acquisition multiple n'expliquent pas l'existence des différents caractères communs aux algues complexes rouges, l'apparition convergente de tous ces critères étant difficilement explicable. Pour répondre à ce problème, de nouvelle théories sont apparues [Baurain et al. 2010 ; Stiller et al. 2014]. Elles proposent des endosymbioses d'ordre supérieur (tertiaire voir quaternaire) ainsi que le transfert séquentiel des plastes. Ces théories suggèrent une facilité de l'établissement de l'endosymbiose et font écho aux divers remplacements de plastes chez les dinoflagellés [Nosenko et al. 2006 ; Minge et al. 2010 ; Imanian et al. 2010]. De plus, de nombreux organismes secondaires montrent une tendance à avoir des génomes relativement chimériques, présentant à la fois des gènes d'origines rouges et vertes [Moustafa et al. 2009 ; Ponce-Toledo et al. 2018]. Ces possibles remplacements de plastes ancestraux rendent la compréhension de l'acquisition de la photosynthèse relativement difficile.

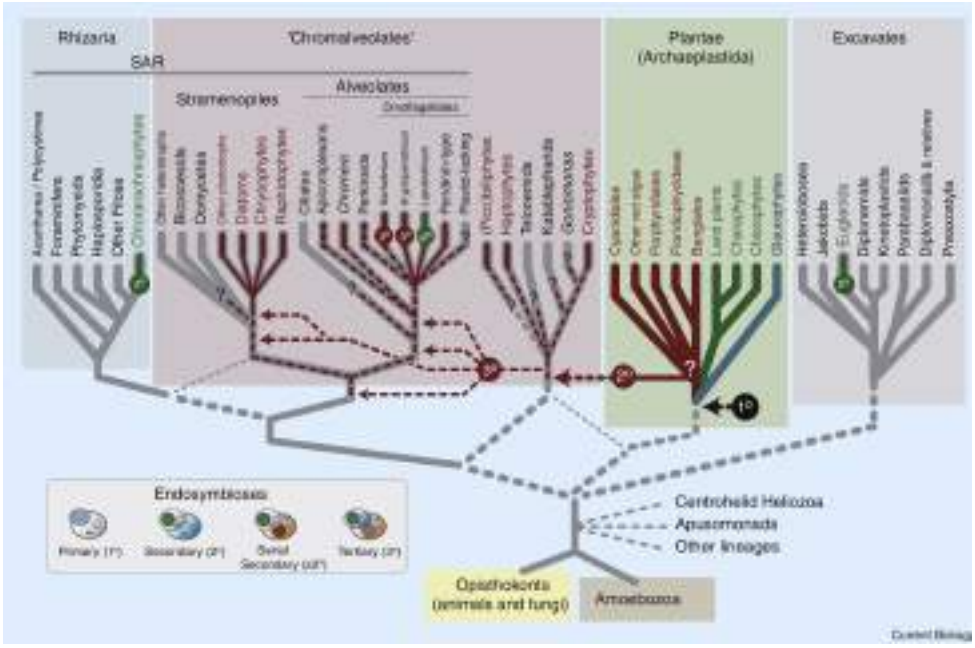


FIGURE 18 – Hypothèse pour l'origine et la propagation de la photosynthèse chez les eucaryotes. La figure met en avant les acquisitions primaires, secondaires et tertiaires possibles au sein de six super groupes eucaryotes. source(ARCHIBALD 2009)

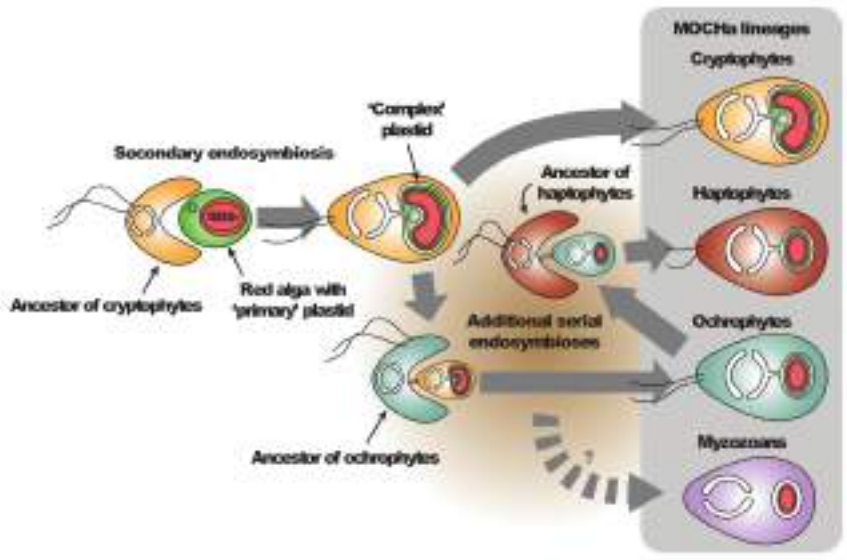


FIGURE 19 – Hypothèse de l'acquisition de la photosynthèse par endosymbioses séquentielles. source(BURKI 2017)

Pour vérifier que l'obtention de la photosynthèse s'est construite par de multiples transferts de plastes, il est nécessaire de trouver des marqueurs décrivant ces transferts horizontaux, c'est-à-dire contredisant la phylogénie des espèces. Cependant, ceci nécessite d'abord de connaître avec précision cette phylogénie. Actuellement, les phylogénies nucléaires dans ce domaine des eucaryotes présentent souvent des incongruences, notamment sur l'ordre de spéciation à l'intérieur des SAR et des Archaeplastida. De même, le positionnement des groupes Haptista et Cryptista est instable, se déplaçant entre les deux groupes susmentionnés avec des supports statistiques faibles. Ces incertitudes rendent donc difficiles, en l'état, la comparaison des phylogénies nucléaires et plastidiales.

Problèmes et limitations lors de l'inférence phylogénomique

Plusieurs facteurs peuvent expliquer l'échec de la résolution complète et non-ambiguë de l'arbre des eucaryotes. Ceux-ci concernent plusieurs étapes du processus d'inférence phylogénétique et pas uniquement le procédé final de la reconstruction de l'arbre à partir de la concaténation de multiples alignements. Ils peuvent bien évidemment être d'ordre méthodologique, c'est-à-dire purement technique, mais également d'ordre pratique. Par exemple, l'échantillonnage d'espèces disponibles actuellement, bien qu'en constante augmentation, est loin de représenter correctement la diversité existante [Sibbald et Archibald 2017]. D'autre part, même une représentation complète du vivant de notre époque n'empêcherait pas l'existence de multiples lignées perdues lors des diverses extinctions [Gouy et al. 2015]. Nous sommes donc forcés d'établir un arbre du vivant en nous basant sur une fraction très réduite de l'information et cette obligation pourrait rester le facteur principal expliquant la difficulté de l'inférence phylogénétique.

Erreur stochastique et phylogénomique

Ce manque d'information affectait particulièrement les débuts de la phylogénie moléculaire. En effet, la quantité de signal phylogénétique dépend à la fois de la précision du modèle d'évolution des séquences mais surtout de la quantité de caractères étudiés. Si le signal extrait est trop faible, on ne peut pas distinguer la meilleure topologie entre plusieurs solutions ce qui mène à des noeuds faiblement soutenus. Ce problème, très connu en statistique, se nomme l'erreur d'échantillonnage ou erreur stochastique et dépend donc principalement de la quantité de données analysable. Pour réduire son impact, la phylogénie moléculaire moderne a opéré un premier changement, le passage à la phylogénomique.

Celui-ci marque le passage de l'utilisation d'un nombre de caractères restreints (dépendant de la taille du gène utilisé) à celui d'une grande partie, voire la majorité, du génome des espèces étudiées. Dans le domaine de l'analyse des séquences moléculaires, on distingua d'abord deux manières de réaliser la phylogénomique (Figure 20)[Delsuc et al. 2005]. La première consiste à récupérer un ensemble d'alignements de séquences orthologues avant de les concaténer pour former ce qu'on appelle une supermatrice, laquelle sera analysée pour obtenir l'arbre des espèces. La seconde nécessite également la récupération des alignements mais afin d'obtenir le meilleurs arbre correspondant à chacun indépendamment. Cet ensemble d'arbres est alors combiné pour former un superarbre.

La méthode du super arbre ne résout pas directement le problème de l'erreur stochastique car chaque arbre de gène reste possiblement affecté. Ce n'est pas le cas avec la méthode de la supermatrice pour laquelle le nombre de positions analysées par le modèle est bien augmenté. En plus d'augmenter directement la quantité de signal extractible, cela permet également d'améliorer l'estimation des paramètres des modèles d'évolution, et en conséquence, d'utiliser des modèles plus complexes. Cependant, la supermatrice pose des hypothèses fortes, celle que l'évolution de chacun des gènes sélectionnés puisse être expliquée par un modèle unique et que leur phylogénie soit identique à la phylogénie des espèces. Aller à l'encontre de ces hypothèses engendre un autre problème lié à la qualité de la modélisation de l'évolution.

Erreur systématique et choix du modèle

Un modèle d'évolution de séquences ne constitue qu'une représentation simplifiée du processus réel de substitution. Il émet des hypothèses qui doivent pouvoir expliquer les données analysées. Cependant, si une partie des données ne valide pas ces hypothèses, on dit qu'elles violent le modèle, ce qui peut produire une erreur lors de l'inférence appelée erreur systématique [Delsuc et al. 2005 ; Rodríguez-Ezpeleta et al. 2007a]. Au contraire de l'erreur stochastique qui se définit par un manque de signal, l'erreur systématique produit un signal appelé signal non-phylogénétique, ce dernier rentrant en compétition avec le véritable signal phylogénétique. Si il est important, il peut mener à des résultats à la fois faux et statistiquement supportés [Philippe et al. 2011b]. En phylogénomique, l'impact de ce type d'erreur est potentiellement plus important vu que le nombre de positions analysé est grand.

Un exemple commun d'artefact de reconstruction lié à l'erreur systématique est le regroupement d'espèces évoluant rapidement, représenté par une branche longue, au sein de l'arbre phylogénétique. Ce phénomène d'attraction des longues branches (LBA, i.e. long branch attraction) est dû à l'échec du modèle à évaluer un processus évolutif présent

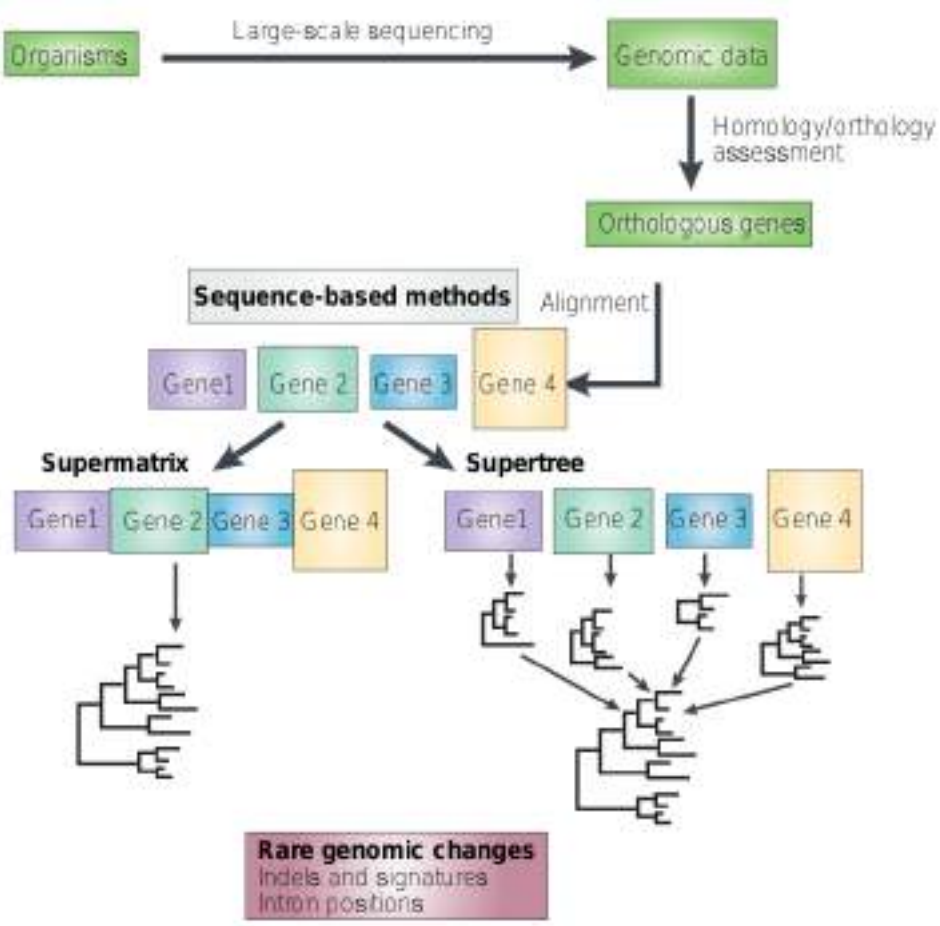


FIGURE 20 – Organigramme présentant les étapes clés d'une inférence phylogénomique classique. Elle sépare deux cas, la supermatrice et le superarbre. source(DELSUC et al. 2005)

chez les espèces évoluant plus rapidement (e.g. hétérotachie). Les multiples substitutions ayant lieu chez ces espèces ont donc plus de chances d'être mal inférées ce qui mènent des convergences à être interprétées comme des synapomorphies produisant un signal non-phylogénétique qui favorise le rapprochement erroné de ces espèces [Germot et Philippe 1999].

Une première possibilité pour réduire l'impact de l'erreur systématique consiste à utiliser le modèle le plus adapté au jeu de données analysé. Par exemple, l'erreur systématique à l'origine du phénomène de LBA peut être réduite en utilisant un modèle implémentant l'hétérogénéité du processus substitutionnel entre sites, comme le modèle CAT [Lartillot et al. 2007]. Il existe des tests statistiques pour faciliter le choix du modèle aussi bien en maximum de vraisemblance (e.g. Likelihood Ratio Test, Akaike Information criterion, Bayesian information criterion) qu'en inférence bayésienne (Cross-validation ou facteur de Bayes) [Sullivan et Joyce 2005]. Certains tests sont d'ailleurs directement incorporés dans certains programmes d'inférence phylogénétique (e.g. IQTREE et ModelFinder [Nguyen et al. 2015 ; Kalyaanamoorthy et al. 2017]). Cependant, il faut avoir conscience qu'il ne s'agit ici que d'une comparaison entre modèles implémentés et non d'une mesure concrète de l'adéquation entre le modèle et les données (voir [Feuda et al. 2017] et [A Shepherd et Klaere 2019]). Cela signifie que le meilleur modèle n'explique jamais parfaitement les données et peut donc produire de l'erreur systématique ; il sera juste le moins mauvais. De plus, il est assez rare qu'un nouveau modèle d'évolution prenant en compte une hétérogénéité du processus substitutionnel soit rapidement inclus dans les programmes d'inférence, souvent pour des raisons computationnelles. Ces tests sont donc loin de comparer l'ensemble des modèles existants.

La seconde manière d'éviter la production d'erreurs systématiques est par le retrait des sites/gènes ou espèces contredisant les hypothèses d'homogénéité du processus substitutionnel [Philippe et Roure 2011]. Par exemple, il est commun de favoriser les espèces évoluant lentement par rapport à celles évoluant rapidement pour représenter un clade. De la même manière, on tend à éviter les marqueurs évoluant trop rapidement, pour lesquels il est difficile de vérifier l'homologie position par position. Cette méthode est notamment utilisée pour remédier à la saturation mutationnelle. Celle-ci affecte les sites subissant de multiples substitutions pouvant mener à des cas d'homoplasie, c'est-à-dire le partage d'un résidu identique pour des causes indépendantes de l'homologie (e.g. la convergence ou la réversion). Du fait de leur nombre de caractères possibles réduits, les séquences nucléotidiques saturent plus rapidement que les séquences d'acides aminés. C'est notamment pour cette raison que ces dernières sont favorisées quand il s'agit de résoudre des phylogénies anciennes. Dans ces cas de saturation, la méthode consiste à retirer petit à petit les sites évoluant le plus rapidement (analyse slow-fast, [Brinkmann et Philippe 1999]). On observe alors l'évolution des supports pour les différentes hypothèses topologiques avec

la diminution des sites rapides (et donc de la potentielle erreur systématique). Un autre moyen de remédier à la saturation pour les séquences en acides aminés et en nucléotides est le recodage des séquences analysées. Le principe est de regrouper les différents résidus en un nombre de catégories réduit représentant des catégories biochimiques (recodage dayhoff en 6 catégories pour les acides aminés [Hrdy et al. 2004], ou en purine/pyrimidine pour les nucléotides [Woese et al. 1991]) ou les propriétés de l'alignement [Susko et Roger 2007]. Ce recodage permet de supprimer les substitutions observées à l'intérieur d'une même catégorie et donc de réduire la saturation [Feuda et al. 2017]. Cependant, si le signal des positions rapides devient plus simple à analyser, celui des positions plus lentes se trouve diminué voire perdu. Il peut donc être difficile d'évaluer les résultats obtenus suite à un recodage.

Incongruence entre l'histoire des gènes et des espèces

Outre la complexité du processus substitutionnel d'évolution des séquences, d'autres éléments peuvent produire de l'erreur systématique. Lors de spéciations proches dans le temps (i.e. lors d'une radiation), des cas de tris de lignée incomplets (incomplete lineage sorting = ILS) peuvent apparaître entre espèces. Dans ce cas, toutes les mutations discrétisant une espèce peuvent ne pas encore être fixées dans la population qui va présenter un polymorphisme lors de l'occurrence d'un second événement de spéciation (Figure 21). Ces mutations peuvent alors se fixer indépendamment dans les lignées subséquentes sans forcément respecter la phylogénie des espèces (Figure 21a et b) [Whitfield et Lockhart 2007]. Ce cas constitue une violation des modèles précédemment discutés car ces derniers suppose qu'une mutation est instantanément fixée dans la population, l'arbre d'un gène est donc toujours censé représenter l'arbre des espèces. L'intégration de l'ILS dans les études phylogénétiques a actuellement beaucoup de succès [Tonini et al. 2015], même si le phénomène n'est pas applicable en toutes circonstances. Il est pris en compte par les analyses de quartet visant à inférer l'arbre d'espèces à partir d'une série d'arbres de gènes [Mirarab et Warnow 2015 ; Zhang et al. 2018]. Cependant, ces méthodes ont pour limitation la présence parfois faible de signal pour les marqueurs pris individuellement (erreur stochastique) menant à un résultat final souvent biaisé [Gatesy et Springer 2014]. L'ILS est également pris en compte par certaines méthodes de réconciliation, qui infèrent simultanément les arbres de gènes et l'arbre d'espèces [Heled et Drummond 2010 ; De Oliveira Martins et al. 2016] mais les implémentations actuelles se limitent à des modèles d'évolution des séquences simples pour des raisons computationnelles.

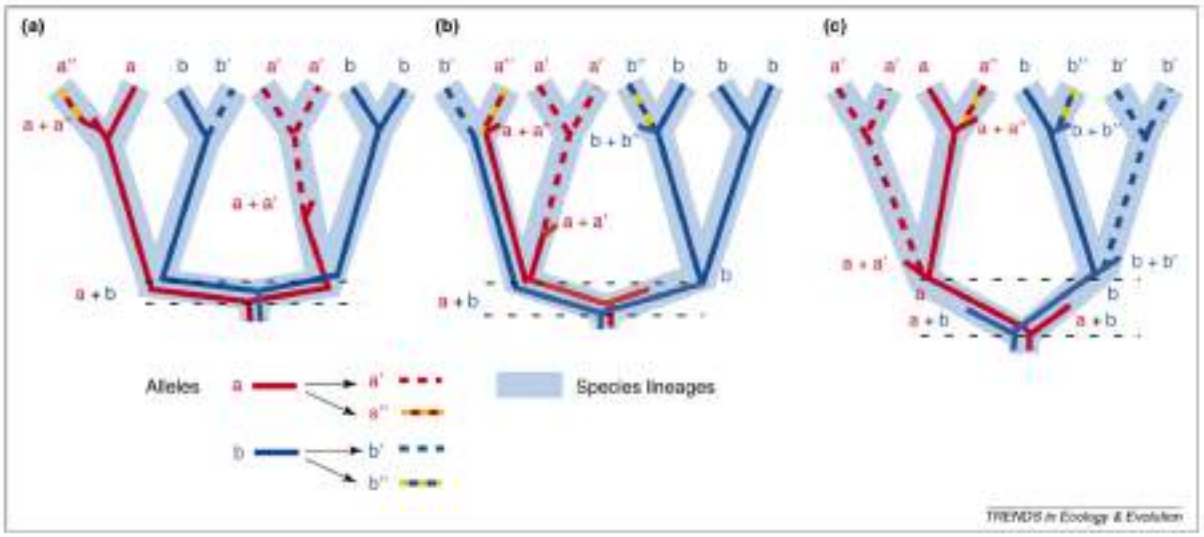


FIGURE 21 – Représentation du tri de lignée incomplet. source(WHITFIELD et LOCKHART 2007)

Qualité des données utilisées

Un dernier élément pouvant affecter l'inférence phylogénétique, et rarement pris en considération, est la qualité des données employées pour la réaliser. Des erreurs réalisées aux différentes étapes précédant l'inférence phylogénétique per se ne peuvent être considérées comme de l'erreur systématique mais peuvent conduire à des arbres incorrects [PHILIPPE et al. 2011b]. Celles-ci peuvent avoir lieu dès le traitement des échantillons avant le séquençage. D'une part, l'identification correcte de l'organisme récupéré est souvent délicate, surtout chez les organismes unicellulaires présentant peu de caractères morphologiques discriminants. D'autre part, de mauvaises pratiques sur le terrain ou en laboratoire peuvent facilement mener à la contamination des échantillons. Les sources les plus communes de contaminations étant naturelles (parasite, symbionte, nourriture) mais peuvent également être techniques (dû aux manipulations sur paillasse). La création des librairies de séquençage est notamment une source possible de contamination entre échantillons (cross-contaminations, [SIMION et al. 2018]). La contamination des séquences biologiques a un impact non négligeable sur la qualité des données utilisées en phylogénie [CORNET et al. 2018]. Ce dernier est d'autant plus important que l'on s'intéresse à l'évolution des organismes par des procédés autres que la filiation, que ce soit par transfert de matériel génétique ou par hybridation. En effet, il n'est pas toujours aisé d'identifier une contamination, surtout si l'on n'en connaît pas exactement la source. Une mauvaise interprétation de l'origine d'une séquence contaminante peut alors mener à une hypothèse de transfert erronée.

L'acquisition des données moléculaires elles-mêmes constitue une autre source d'erreurs possibles. Les étapes successives de séquençage et d'assemblage présentent leurs propres

défis et types d'erreurs, différents en fonction des technologies utilisées [El-Metwally et al. 2013]. Il s'ensuit la définition des marqueurs utilisés c'est-à-dire les séquences correspondant aux gènes codants pour des protéines. Celles-ci peuvent être obtenues directement à partir de la transcriptomique et/ou par annotation des séquences génomiques. Cette dernière fait également intervenir des modèles complexes d'identification des séquences codantes devant prendre en compte les régions non transcrites (5' et 3' UTR), la détection correcte des codons de départ et d'arrêt de la traduction, l'épissage des transcrits (retrait des introns, épissage alternatif), etc. Chacun de ces éléments peut être source d'erreurs.

La dernière étape sources d'erreurs correspond à celle de la construction des alignements. Nous avons déjà discuté notamment du fait de distinguer les séquences orthologues parmi l'ensemble des séquences homologues bien que ce point n'affecte pas toutes les méthodes d'inférence (e.g. les événements de duplication et de transfert sont pris en compte par les méthodes de réconciliation [Boussau et al. 2013 ; Szöllősi et al. 2015]). La qualité de l'alignement de séquences a également une importance pour les méthodes d'inférence nécessitant un alignement [Talavera et Castresana 2007]. Bien que le problème de l'alignement puisse sembler facile lorsqu'il concerne des domaines conservés, celui-ci peut devenir un vrai défi lorsqu'il s'agit de considérer des fragments divergents de séquences. Ce type de segment peut affecter la qualité d'un alignement. Cependant, de nombreuses méthodes de filtre ont été développées pour éliminer les possibles erreurs d'alignement afin d'éviter l'analyse de positions non homologues [Castresana 2000 ; Capella-Gutiérrez et al. 2009 ; Criscuolo et Gribaldo 2010 ; Wu et al. 2012 ; Sela et al. 2015]. Néanmoins, l'effet final de l'application de ces filtres reste débattu (surtout lors de l'inférence des arbres de gènes) car ils impactent la fragile balance entre bruit et signal [Tan et al. 2015].

Un modèle parfait pour la phylogénomique

L'inférence phylogénétique est donc impactée par un large spectre de disciplines, allant de l'évolution à la biochimie en passant par l'écologie et la génétique. De nombreux éléments sont à prendre en compte et l'on pourrait considérer, à raison, qu'il est vain de vouloir maintenir séparées les différentes étapes menant à l'inférence de l'arbre des espèces. En effet, ne serait-il pas plus simple de créer un modèle prenant en compte tous les éléments décrits jusqu'à présent, de l'extraction de l'ADN à la création de l'arbre final ? Certains modèles ayant déjà été abordés vont dans ce sens, comme les modèles de réconciliation [Heled et Drummond 2010 ; Boussau et al. 2013 ; Szöllősi et al. 2015 ; De Oliveira Martins et al. 2016]. En effet, ces derniers considèrent à la fois l'évolution des séquences et des gènes en réalisant l'inférence jointe des arbres de gènes et de l'arbre d'espèce. En

plus du phénomène d'ILS, ce type de modèle joint permet également de considérer les cas de duplication et perte (paralogie) ainsi que les transferts de gène (xénologie) dans une même inférence [Szöllősi et al. 2013b ; Szöllősi et al. 2013a ; Szöllősi et al. 2015]. Sur le même principe, on peut combiner l'inférence de l'arbre à l'alignement des séquences [Redelings et Suchard 2005 ; Lunter et al. 2005]. Si les modèles joints infèrent simultanément plusieurs étapes clés du processus phylogénétique, on peut également trouver des modèles améliorés par l'apport d'informations externes. On peut notamment citer l'alignement de séquences protéiques tenant compte de leur structure ou l'assemblage et l'annotation de séquences en croisant les données génomiques et transcriptomiques en provenance de plusieurs types de technologie de séquençage [Chevreux et al. 1999].

Bien que l'idée d'un tel modèle soit séduisante, elle n'est malheureusement pas réalisable, voire réaliste. En effet, les modèles dont je viens de parler ont pour particularité d'être extrêmement demandeur en ressources, qu'elles soient informationnelles ou computationnelles. Du point de vue informationnel, il est rare d'avoir à sa disposition des types et des sources de données multiples. Cette affirmation est surtout vraie si l'on ne travaille pas sur des organismes modèles, ce qui est commun en phylogénie. D'un point de vue computationnel, l'utilisation de modèles complexes entraîne une augmentation de la combinatoire à évaluer (des combinaisons de plusieurs problèmes, chacun NP-complet), requérant des capacités et des temps de calcul importants. L'inférence phylogénétique se trouve finalement limitée, devant faire face à la gestion de contraintes multiples pour répondre au mieux aux questions posées. Il n'est donc pas rare de devoir faire des concessions dans la création du jeu de données, en terme de nombres d'espèces ou de gènes incorporés, ou dans les modèles employés afin de rendre le problème analysable dans des temps acceptables avec les ressources disponibles. Ces choix techniques ne sont pas triviaux car ils peuvent aussi impacter le résultat final. Ils doivent donc être réalisés intelligemment et leur impact devrait être estimé.

Objectifs du projet de thèse

Le projet de cette thèse a pour objectif d'étudier l'acquisition de la capacité de photosynthèse oxygénique chez les eucaryotes par des méthodes phylogénomiques. Cet objectif se subdivise en plusieurs points principaux :

1. Tout d'abord, l'étude de la phylogénie des eucaryotes en prenant en compte les nouvelles données de séquençage obtenues lors du projet MMETSP [Keeling et al. 2014], de l'expédition Tara [Seeleuthner et al. 2018], ainsi que celles apparues dans les bases de données biologiques. Ce point se focalise sur la résolution des relations de spéciation entre les eucaryotes photosynthétiques du point de vue des

hôtes. Il vise également à évaluer en quelle proportion l'absence de résolution dans l'arbre des eucaryotes peut être attribuée à des artéfacts méthodologiques plutôt qu'à l'absence de signal phylogénétique.

2. Le recoupement de la phylogénie eucaryotes avec la phylogénie des plastes. Ce point vise à réévaluer les hypothèses d'acquisition des plastes.
3. L'étude des gènes non-photosynthétiques transférés par endosymbiose chez les eucaryotes photosynthétiques complexes, l'objectif étant de déterminer les avantages évolutifs amenés par ce flux de gène.
4. La datation de la phylogénie des eucaryotes. Ce point a pour objectif de contextualiser les flux de gènes identifiés au point 3 afin de formuler des hypothèses sur leur impact écologique et évolutif au sein des communautés de phytoplancton.

Dans le cadre de cette thèse, mon travail s'est focalisé sur les deux premiers points parmi ceux que je voulais aborder. Plus précisément, il s'est articulé autour de l'impact de l'erreur systématique sur l'inférence phylogénétique et sur l'amélioration du rapport entre signal phylogénétique et non phylogénétique. J'ai étudié ces deux éléments à plusieurs étapes du processus phylogénétique, de l'annotation des séquences protéiques à l'inférence de l'arbre des espèces, en passant par l'identification des séquences orthologues.

Au sein de ce manuscrit de thèse, j'ai organisé les éléments clés de mon travail en trois chapitres :

1. Le premier chapitre traite de l'impact relatif des erreurs présentes dans les séquences biologiques et des erreurs d'alignement sur les analyses évolutives. Il présente un nouveau programme de filtrage d'alignement de séquences visant à en retirer les segments non-homologues dans le but de réduire l'impact des séquences erronées. Ce chapitre a fait l'objet d'une publication dans BMC evolutionary biology.
2. Le second chapitre s'intéresse aux considérations méthodologiques lors de la création d'un jeu de données phylogénomique. Il se focalise sur la récupération massive de groupes d'orthologie comprenant un nombre élevé d'organismes tout en s'assurant de la validité de la relation d'orthologie entre chaque séquence. L'objectif final est d'obtenir un jeu de données le plus complet et propre possible pour inférer la phylogénie des eucaryotes. Le chapitre se termine donc par l'inférence d'un arbre des eucaryotes et par une analyse de l'impact du modèle utilisé en fonction du nombre d'espèces analysées. Il est basé sur le travail réalisé en début de thèse pour réaliser un poster présenté lors de la conférence Jacques Monod ayant eu lieu à Roscoff en mai 2016.

3. Le troisième et dernier chapitre reprend une étude détaillée de la phylogénie des stramenopiles (=hétérocontes). Elle se focalise sur la résolution de la phylogénie des ochrophytes (=stramenopiles photosynthétiques) et sur la congruence entre la phylogénie nucléaire et la phylogénie plastidiale. Autour de ces thématiques, nous avons reformulé l'intensité du signal phylogénétique inféré vis-à-vis de l'erreur systématique en introduisant le concept d'efficacité d'extraction.

Je joins également au manuscrit une étude à laquelle j'ai collaboré sur la phylogénie des animaux. Elle en constituera l'annexe.

Chapitre 1

Évaluation de l'impact des erreurs de séquence primaire sur les analyses en évolution

---

1.1 Résumé du chapitre 1

Ce premier chapitre se base sur nos résultats publiés dans BMC Evolutionary Biology. Dans cette publication, nous questionnons les pratiques actuelles de filtrage des alignements de séquences multiples. Ces dernières se focalisent principalement sur les possibles erreurs commises lors de l'alignement en déterminant et retirant les positions pour lesquelles ce dernier semble ambiguë. Cependant, les séquences employées peuvent contenir des erreurs obtenues lors des étapes de séquençage et d'assemblage ou encore lors de l'annotation des portions codantes. Ces erreurs peuvent également impacter la qualité de l'alignement et ne sont pas ciblées par les outils actuellement utilisés. De même, aucune étude ne s'est intéressé à l'impact de ces erreurs sur les analyses en évolution dépendantes de ces alignements.

Ici, nous présentons un nouveau logiciel, HmmCleaner, dont l'objectif est de retirer les segments de séquences individuelles présentant une faible similarité avec le reste des séquences de l'alignement afin de traiter les erreurs de séquence primaire. Ces performances face à des erreurs simulées sont comparées à celles d'autres logiciels de filtre, notamment ceux visant les erreurs d'alignements. Nous analysons l'impact de ces logiciels sur les résultats obtenus lors de la détection de sélection positive ou de l'inférence phylogénétique, notamment sur la topologie ou les longueurs de branches.

RESEARCH ARTICLE Open Access

Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences

Arnaud Di Franco[1], Raphaël Poujol[2], Denis Baurain[3] and Hervé Philippe[1,2*]

Abstract

Background: Multiple Sequence Alignments (MSAs) are the starting point of molecular evolutionary analyses. Errors in MSAs generate a non-historical signal that can lead to incorrect inferences. Therefore, numerous efforts have been made to reduce the impact of alignment errors, by improving alignment algorithms and by developing methods to filter out poorly aligned regions. However, MSAs do not only contain alignment errors, but also primary sequence errors. Such errors may originate from sequencing errors, from assembly errors, or from erroneous structural annotations (such as incorrect intron/exon boundaries). Even though their existence is acknowledged, the impact of primary sequence errors on evolutionary inference is poorly characterized.

Results: In a first step to fill this gap, we have developed a program called HmmCleaner, which detects and eliminates these errors from MSAs. It uses profile hidden Markov models (pHMM) to identify sequence segments that poorly fit their MSA and selectively removes them. We assessed its performances using > 700 amino-acid MSAs from prokaryotes and eukaryotes, in which we introduced several types of simulated primary sequence errors. The sensitivity of HmmCleaner towards simulated primary sequence errors was > 95%. In a second step, we compared the impact of segment filtering software (HmmCleaner and PREQUAL) relative to commonly used block-filtering software (BMGE and TrimAl) on evolutionary analyses. Using real data from vertebrates, we observed that segment-filtering methods improve the quality of evolutionary inference more than the currently used block-filtering methods. The formers were especially effective at improving branch length inferences, and at reducing false positive rate during detection of positive selection.

Conclusions: Segment filtering methods such as HmmCleaner accurately detect simulated primary sequence errors. Our results suggest that these errors are more detrimental than alignment errors. However, they also show that stochastic (sampling) error is predominant in single-gene evolutionary inferences. Therefore, we argue that MSA filtering should focus on segment instead of block removal and that more studies are required to find the optimal balance between accuracy improvement and stochastic error increase brought by data removal.

Keywords: Multiple sequence alignment, Profile hidden Markov models, Low similarity segments, Primary sequence error, Phylogeny, Positive selection

Background

Evolutionary studies require the identification of homologous characters. Except for highly divergent proteins, the recognition of homologous protein-coding genes is generally straightforward because the availability of many positions provides enough statistical power, at least some protein regions being well conserved (i.e., show a high similarity). In contrast, the identification of homology at the residue level, through multiple sequence alignment (MSA), is more difficult. Limited statistical power and low similarity may generate ambiguously aligned regions (AARs). Due to the high combinatorics of sequence alignment, some parts of AARs are expected to be aligned wrongly more often than correctly. Despite efforts in improving alignment methods [1], errors still affect MSAs and may negatively impact subsequent analyses. During phylogenetic inference,

* Correspondence: herve.philippe@sete.cnrs.fr
[1]Station d'Ecologie Théorique et Expérimentale de Moulis, CNRS, Moulis, France
[2]Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, Montréal, Québec, Canada
Full list of author information is available at the end of the article

they generate a non-phylogenetic signal, conflicting with the genuine (historical) phylogenetic signal in the data [2, 3]. Their presence also inflates estimates of positive selection [4, 5].

A common approach to reduce the impact of alignment errors is a posteriori filtering of MSAs. The rationale of the block-oriented strategy is that alignment errors are in excess in AARs, the variable regions of the proteins (in particular those with a high rate of insertion/deletion). Several software packages [6–12] were designed to identify AARs based on various criteria, such as the stability of the MSA to the guide tree [12] or the validation of a set of rules dependent on the conservation pattern [6, 8, 9]. AARs are expected to contain non-homologous residues in most sequences, but also genuine homologous residues in the remaining sequences. Removal of AARs is therefore expected to simultaneously decrease non-phylogenetic and phylogenetic signal, but the first more than the second. Some studies suggest that block-filtering software improves evolutionary inference [13–16], whereas other authors find support for the opposite [17, 18].

Another source of noise in MSAs are primary sequence errors. These stem from sequencing errors, assembly errors or, in the case of amino-acid MSAs, structural annotation errors (such as incorrect intron/exon boundaries). Fundamentally different from alignment errors, primary sequence errors (especially those affecting only one or a few sequences) are unlikely to be removed by block-filtering programs, except if they are included within AARs. To properly handle such errors, filtering software should be designed to remove amino-acid segments sequence by sequence, instead of block by block.

Besides, primary sequence errors provide a strong non-historical signal that is more likely to bias evolutionary estimates (e.g., by lengthening the corresponding terminal branches in a phylogeny). Accordingly, a few studies have shown that they can be a source of erroneous signal [3] or even drive alignment errors [5]. Yet, this aspect is generally not taken into account while analyzing MSAs. In fact, nothing is known about the relative importance of primary sequence errors versus alignment errors in evolutionary analysis of real MSAs.

Here, we present HmmCleaner, a program dedicated to the detection and removal of primary sequence errors in multiple alignments of protein sequences. It implements an approach looking for low similarity segments specific to one sequence using a profile hidden Markov model (pHMM) built from the whole alignment with HMMER [19]. In the following sections, we first introduce the HmmCleaner principle. Then we explain the optimization of its parameters, characterize its performance by simulating primary sequence errors and compare it to PREQUAL performance [20], a recently released software package with a similar approach based on pairHMM. Then, we address the effect of filtering software on evolutionary analysis. First, we determine whether the use of HmmCleaner avoids the erroneous detection of positive selection when frameshift errors have been voluntarily introduced. Second, using empirical datasets, we compare the effect of segment- and block-filtering methods on evolutionary inferences (single--gene phylogenetic reconstruction and branch length estimation) as a first insight into the relative impact of alignment errors and primary sequence errors.

Results and discussion

Overview of HmmCleaner and parameter optimization

HmmCleaner identifies primary sequence errors by detecting low similarity segments in an MSA (Fig. 1). In our framework, low similarity segments are stretches of residues that are highly divergent with respect to the full alignment (in terms of sequence, length or both). They are identified through four steps. First, a pHMM is built from the MSA using HMMER (Fig. 1a); it will be used as the reference, i.e., the underlying model having generated each sequence of the MSA. It can either be built upon (i) all sequences of the MSA (complete strategy) or (ii) all sequences except the currently analyzed one (leave-one-out strategy). Second, each sequence of the MSA is evaluated with the pHMM (Fig. 1b), which yields one profile-sequence alignment per sequence through the heuristic of HMMER [19]. Third, each profile-sequence alignment is analyzed by considering the four categories of match between the sequence and the pHMM consensus using a four-parameters matrix that increases a cumulative similarity score when the residue is expected by the pHMM or decreases it otherwise (Fig. 1c). The evolution of this similarity score depicts the variation of the corresponding sequence fit to the pHMM along its whole length. Fourth, low similarity segments are defined as continuous segments where the similarity score was lower than the maximal value and among which at least one residue reaches a null score (Fig. 1d, see Materials and Methods for details).

To optimize the four parameters of the scoring matrix, we developed a simulator that introduces primary sequence errors into existing MSAs. The principle is to take a genuine protein-coding alignment of nucleotide (nt) sequences and to randomly introduce a unique error in a specified number of sequences. The resulting sequences are then translated into amino acids (aa) before realignment. Here, we chose to generate frameshift errors, each one followed by the opposite (compensatory) mutation after a predefined number of out-of-frame codons. This approach allowed us to use multiple alignments of true protein sequences resulting from real evolutionary processes whereas primary sequence errors are simulated, contrary to Whelan et al. [20], who started from simulated sequences.

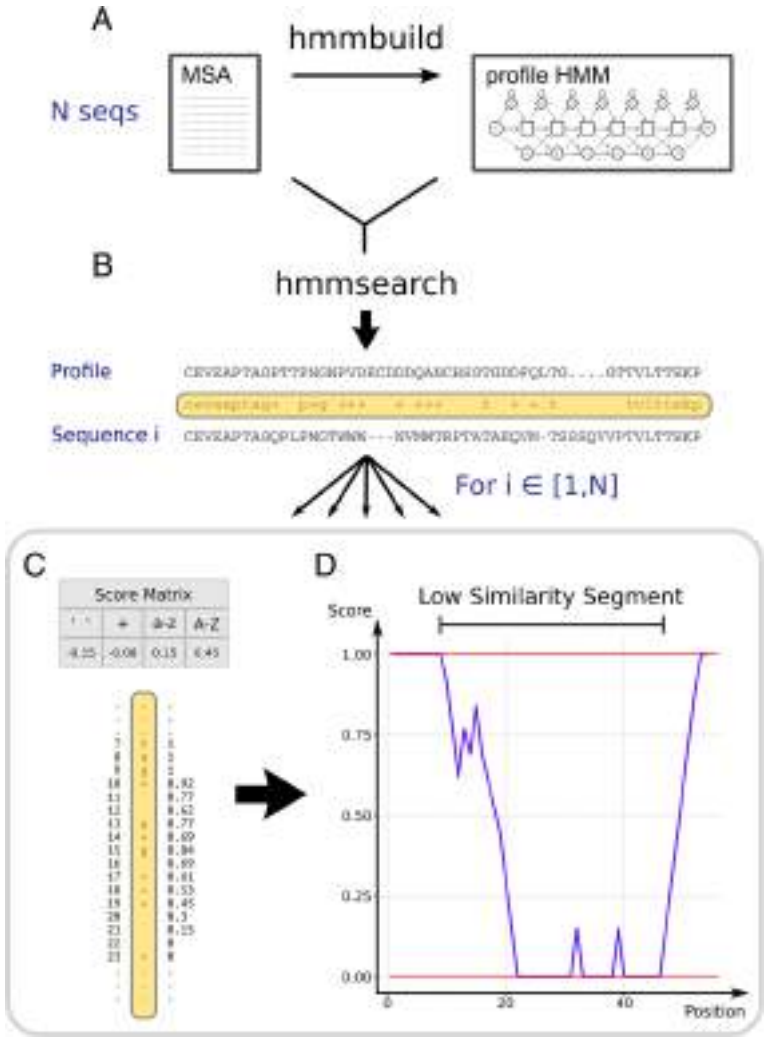


Fig. 1 Overview of the four steps of HmmCleaner. **a**. Creation of a pHMM from the MSA. In the diagram representing the pHMM, squares correspond to main states of the model whereas diamonds are insertion states and circles deletion states. **b**. Alignment of one envelope of a given sequence of the MSA. **c**. Computation of the cumulative similarity score based on the analysis of the string of match/mismatch (underlined in yellow) using the four-parameter scoring matrix. **d**. Determination of low similarity segments as continuous sequence stretches with similarity score < 1 in which at least one position reached a null score

Two empirical datasets of 100 MSAs composed of prokaryotic genes (either Euryarchaeota or Cyanobacteria) were submitted to our simulator in four variants of species sampling (5, 10, 25 and 50 species), so as to generate 100 replicates of these 800 combinations containing 1 to 5 primary sequence errors of length 10 to 100 aa. The random sampling of species allowed us to test a large variety of tree shape, given that HMMER implicitly assumes that a star tree topology has generated the alignment. The 80,000 simulated MSAs were then used to explore the effect of the four parameter values on both the sensitivity (detection of truly non-homologous segments) and the specificity (non-detection of genuinely homologous segments) of HmmCleaner. We chose to work only on unambiguously aligned regions (UARs), reasoning that it is more difficult to differentiate non-homologous segments from homologous but highly divergent ones in ambiguously aligned regions (AARs). Indeed, over a grid of 2835 quartets of parameter values (9*c1*, 7*c2*, 9*c3*, 5*c4*, see Materials and Methods), HmmCleaner only reached a limited specificity (93%) in AARs (Additional file 1 Figure S1), and this came at the expense of a low sensitivity (22%). In contrast, in UARs (Fig. 2), numerous quartets of parameter values led to bothhigh sensitivity and specificity (> 90%), showing that HmmCleaner reliably detects simulated primary sequence errors. Moreover, these results held true when focusing on either Euryarchaeota or Cyanobacteria, and when varying the operational definition of UARs/AARs (Additional file 1 Figure S2A-D). In contrast, sensitivity was reduced with smaller

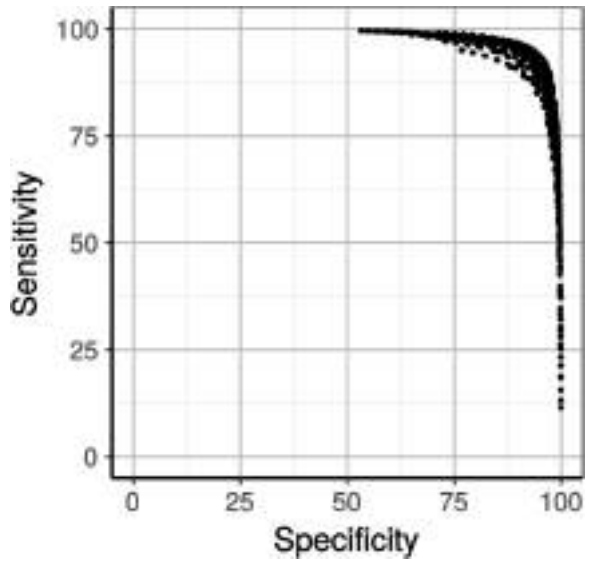


Fig. 2 Mean sensitivity and specificity of HmmCleaner towards detection of primary sequence errors introduced in unambiguously aligned regions (UARs). Each dot corresponds to the two means of the values obtained across 80,000 simulations and 3 operational definitions of UARs for one of the 2835 combinations of the 4 parameters of the scoring matrix

species samples (Additional file 1 Figure S2E), as expected since pHMMs are more powerful when built from a larger number of sequences.

In the end, we selected as the default matrix the parameter set that maximized the mean sensitivity and the mean specificity across all variations of the conditions of simulation. This set of parameters yields both a global sensitivity and specificity of 94% in UARs (Table 1). It improves specificity compared to the empirically determined parameters present in a previous, unpublished, implementation of HmmCleaner, referred to as version 1.8. However, the usual trade-off between specificity and sensitivity, and the negative effect of smaller numbers of species on sensitivity, led us to define three new additional scoring matrices. First, we created a parameter set to use when the number of sequences in the MSA is large, optimized only from simulations performed with 50 species. This scoring matrix achieved a global sensitivity of 97% and a global specificity of 96% in UARs of species-rich MSAs. Second, for users wishing a low false positive rate, we built a high-specificity matrix reaching a global specificity of 97% while keeping a sensitivity of 88%. Finally, we also generated a matrix that simultaneously addresses these two requirements (see Table 1).

Since the default scoring matrix had been optimized using MSAs aligned with MAFFT using the L-INS-i algorithm and by building a single pHMM per MSA (complete strategy), we checked that parameter optimization was robust to a change of the aligner software and the HmmCleaner strategy. While both sensitivity and specificity revealed virtually insensitive to the aligner software (MAFFT [21] with two different algorithms, MUSCLE [22] and Clustal Omega [23], Additional file 1 Figure S3), the leave-one-out strategy showed a slightly higher sensitivity (98% versus 97%), but a lower specificity (79% versus 86%), with respect to the complete strategy (Additional file 1 Figure S4). Given these results and the computational burden implied by the leave-one-out strategy, we decided to stick to the complete strategy in the remaining of this article.

Impact of error length, number and conservation context on HmmCleaner performance

To investigate the impact of the length and number of primary sequence errors on the sensitivity and specificity of HmmCleaner, 640,000 simulations introducing a total of 4,960,000 individual errors were run on MSAs from 4 different prokaryotic lineages (Alphaproteobacteria and Crenarchaeota in addition to the Cyanobacteria and Euryarchaeota used so far). Introduced errors were 10 to 100 aa in length and 1 to 15 in number per MSA of 25 randomly selected sequences. Neither error length or number, nor MSA lineage substantially impacted specificity and sensitivity of HmmCleaner, except in two cases (Fig. 3). First, specificity decreased with evolutionary depth and diversity of the clade (Cyanobacteria < Alphaproteobacteria < Crenarchaeota < Euryarchaeota, Fig. 3f), which is in agreement with the idea that HmmCleaner wrongly detects some homologous low similarity segments (see below). Second, sensitivity was severely

Table 1 Sensitivity and specificity of HmmCleaner

matrix name	c1	c2	c3	c4	global sensitivity	global specificity	sensitivity for 50 seqs	specificity for 50 seqs
default	−0.150	−0.080	0.150	0.450	94.26%	94.42%	98.42%	93.97%
species-rich	−0.175	− 0.175	0.150	0.400	85.70%	97.02%	97.36%	96.48%
high-specificity	−0.125	−0.125	0.175	0.400	88.89%	97.34%	96.64%	97.08%
species-rich and high- specificity	−0.125	− 0.125	0.150	0.400	74.03%	98.78%	94.11%	98.56%
HmmCleaner 1.8	−0.300	−0.100	0.200	0.500	94.56%	91.57%	98.74%	90.73%

The four new scoring matrices provided with HmmCleaner v2 and the scoring matrix equivalent to HmmCleaner v1.8 for comparison. c1-c4: values of the elemental scores for the four levels of residue conservation provided by HMMER. Global sensitivity and specificity were computed across all conditions of simulation, whereas the last two columns only used the most species-rich MSAs

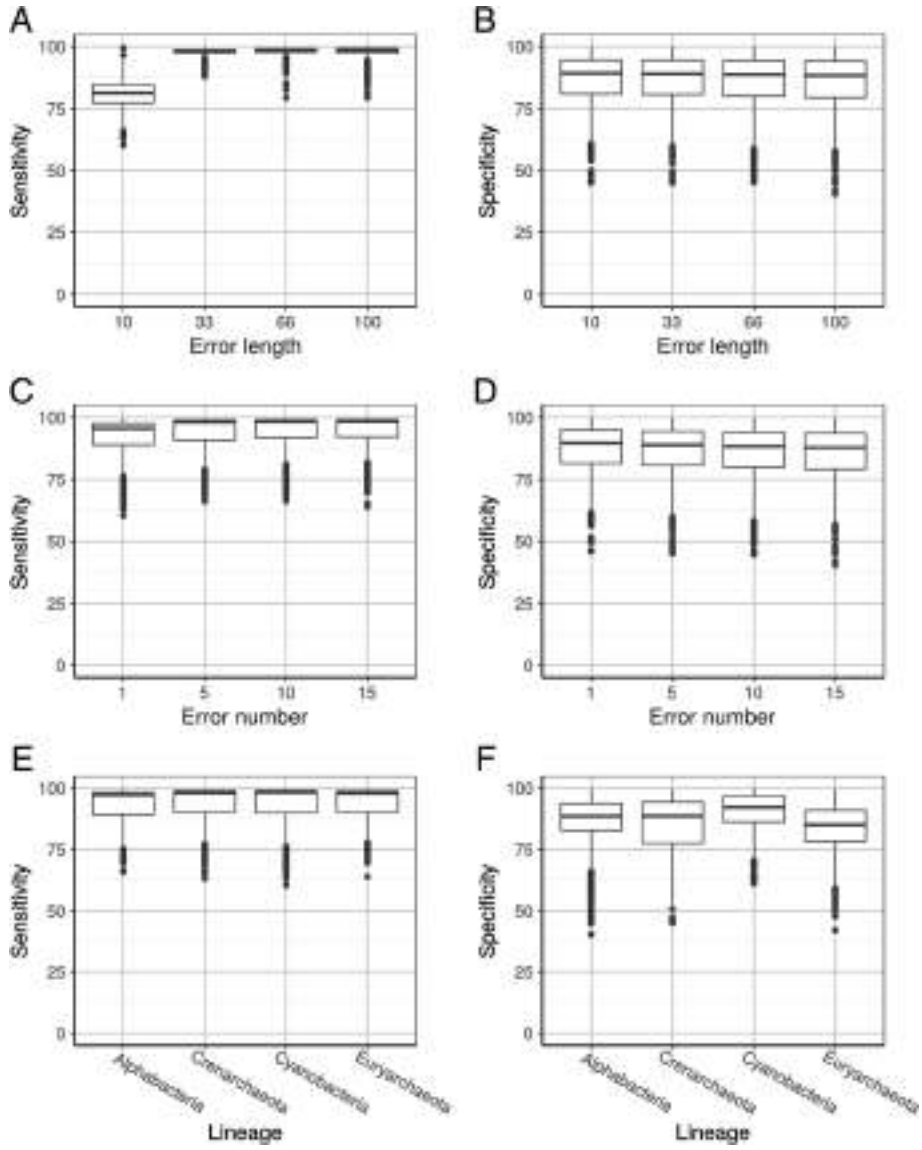


Fig. 3 Impact of the length and number of primary sequence errors, and of the prokaryotic lineage, on sensitivity (**a**,**c**,**e**) and specificity (**b**,**d**,**f**) of HmmCleaner used with the default scoring matrix. **a**,**b**. Effect of primary sequence error length. **c**,**d**. Effect of the number of primary sequence errors. **e**,**f**. Effect of the prokaryotic lineage. Box-plots were computed across all considered MSAs and values are means averaged over the different conditions of simulation

impacted by short error lengths (Fig. 3a), owing to the limited statistical power provided by such short primary sequence errors.

The same kind of simulations were used to compare HmmCleaner and its different parameter sets to PREQUAL (Additional file 1 Table S1, Additional file 1 Figure S5). Overall, PREQUAL showed higher specificity (92.4% vs 86.7%) and lower sensitivity (83.3% vs 93.3%) compared to the new HmmCleaner default scoring matrix. As expected from the known sensitivity/specificity tradeoff, using the specificity-oriented parameter set reduced the difference in specificity (92.4% vs 91.8%) while keeping sensitivity slightly higher than PREQUAL (83.3% vs 86.1%). Only the "large specificity" parameter set surpassed PREQUAL specificity but at the cost of lower sensitivity. The behavior of HmmCleaner and PREQUAL towards error length and taxonomic diversity is similar. However, PREQUAL sensitivity diminished with increasing error numbers, whereas we observed the opposite for HmmCleaner. Our hypothesis is that more errors increase the probability of having overlapping identical errors. As PREQUAL considers the best posterior probability per residue across a series of closely related sequences, only one identical residue is enough to consider it as correct. In summary, the behavior

of HmmCleaner and PREQUAL is similar, which could be due to their common reliance on HMM, and HmmCleaner appears more sensitive yet less specific than PREQUAL.

To accurately analyze how sensitivity was impacted by error length, more simulations containing a single error 1 to 33 aa in length were carried out with the default scoring matrix of HmmCleaner. As expected, sensitivity increased with error length (Fig. 4), achieving a mean sensitivity < 90% for error lengths < 13 aa. Theoretically, when using the default scoring matrix, an error < 7 aa cannot be detected by HmmCleaner in UARs. This is because this length corresponds to the minimal number of increments needed to decrease the cumulative score from 1 to 0 (– 0.15*7, see Fig. 1). Yet, it is possible to detect shorter errors (e.g., ~ 35% of 6 aa errors) when they are included in divergent regions (AARs) where the score is already < 1. Conversely, the fact that the score is often < 1 increases the probability of reaching 0 by chance, and thus of creating false positives. In other words, HmmCleaner retains some sensitivity for short errors at the expense of its specificity. Importantly, mean sensitivity is > 95% for error lengths > 17 aa, which indicates that only short primary sequence errors will remain in the cleaned MSAs.

HmmCleaner sensitivity proved to be robust to the conservation context of the regions in which primary sequence errors were introduced, except for error lengths < 10 aa (Additional file 1 Figure S6). In particular, short errors were more easily detected in gap-rich regions (Additional file 1 Figure S6A) than in fast-evolving regions (Additional file 1 Figure S6B). In gappy regions, a possible explanation for the good sensitivity could be that there are only few sequences to locally define the pHMM. Consequently, HMMER expects the presence of a highly specific segment of amino acids and is thus more severe when the observed segment does not correspond. Regarding the worse sensitivity in fast-evolving regions, our interpretation is that the pHMM is less specific (flat profile) and can more easily accommodate any divergent segment, including primary sequence errors. In contrast, the level of alignment ambiguity (AAR versus UAR) did not affect the detection of simulated primary sequence errors, whatever the error length (Additional file 1 Figure S6C). HmmCleaner thus accurately detects all simulated errors but shorter ones in all types of regions.

Overall efficiency of filtering algorithms on primary sequence errors

Having studied the efficiency of segment-filtering methods (HmmCleaner and PREQUAL) only on frameshift primary sequence errors in prokaryotic sequences, it was important to test them on other types of error and other types of sequences. To this end, we considered eukaryotic sequences (112 alignments from mammals and 170 from vertebrates) and two new error types: (i) scrambled amino acid segments generated by shuffling the corresponding underlying individual nucleotides and (ii) arbitrary insertions obtained by inserting shuffled nucleotide segment (see Materials and Methods for details). The sensitivity of HmmCleaner (Table 2) was virtually identical for prokaryotes and eukaryotes, despite very different evolutionary depths (from mammals to euryarchaeotes). Similarly, scrambled segments were detected as efficiently as frameshifts (~ 96%) while arbitrary insertions were more easily recognized (99%). PREQUAL yielded similar results (Table 2), except that its sensitivity was lower for frameshifts in eukaryotes (only 85.64%). Regarding specificity, error type had no effect, except for

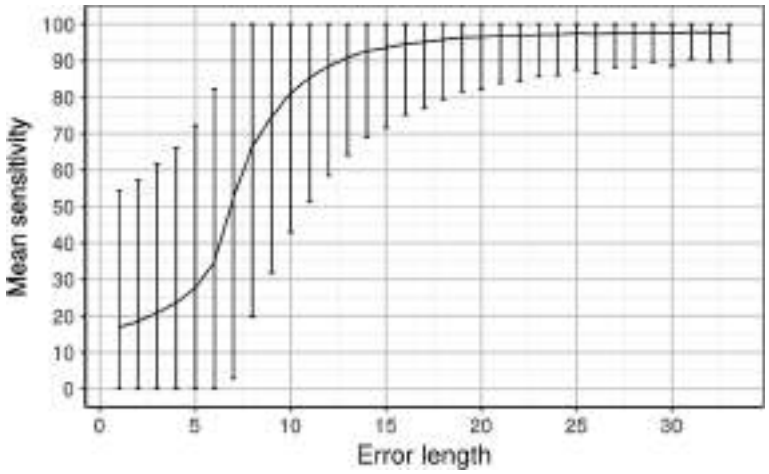


Fig. 4 High-resolution analysis of the impact of the length of primary sequence errors on the sensitivity of HmmCleaner used with the default scoring matrix. The plain line represents the improvement in mean sensitivity with increasing error length, while error bars show the variability across 400,000 simulated primary sequence errors in 400 MSAs

Table 2 Sensitivity and specificity of filtering software over different error types

sensitivity	prokaryotic MSAs			eukaryotic MSAs		
	frameshifts	scrambled	insertions	frameshifts	scrambled	insertions
HmmCleaner	96.95%	96.24%	99.24%	95.99%	96.26%	99.34%
PREQUAL	93.13%	96.77%	99.83%	85.64%	97.34%	99.95%
BMGE	16.00%	19.23%	97.55%	8.57%	12.57%	96.16%
OD-seq	20.11%	25.09%	59.89%	8.08%	9.34%	59.80%
GUIDANCE2	4.90%	5.37%	0.25%	2.33%	3.25%	1.69%
specificity	prokaryotic MSAs			eukaryotic MSAs		
	frameshifts	scrambled	insertions	frameshifts	scrambled	insertions
HmmCleaner	87.28%	87.24%	85.64%	94.70%	94.70%	94.63%
PREQUAL	92.80%	92.78%	92.76%	94.77%	94.65%	94.59%
BMGE	91.01%	90.88%	91.10%	97.05%	96.93%	96.94%
OD-seq	94.48%	94.46%	92.55%	95.20%	95.25%	94.90%
GUIDANCE2	99.55%	99.48%	99.85%	98.47%	98.35%	98.51%

arbitrary insertions in prokaryotes, which slightly decreased HmmCleaner specificity (85.64% versus ~ 87%), possibly because insertions disturbed the alignment of correct sequences [5]. In contrast, HmmCleaner was more specific for eukaryotic than prokaryotic sequencs (~ 94.5% versus ~ 87%), our eukaryotic genes being less divergent. The same was true for PREQUAL, but to a smaller extent (~ 94.5% versus ~ 92.8%). In summary, sensitivity and specificity of both segment-filtering methods were relatively unaffected by error type and data type. These results were especially welcome for HmmCleaner, which had been trained on prokaryotic frameshifts alone.

Segment-filtering methods were developed based on the hypothesis that block-filtering methods are not adapted to detect primary sequence errors. To formally test this assumption, we confronted PREQUAL and HmmCleaner to the block-filtering software BMGE and to OD-seq [24] and GUIDANCE2 [12], two methods designed to filter outlier sequences from MSAs. Our expectations were that block- and outlier-filtering methods would display a limited specificity, even when accurately detecting the simulated errors, due to the former removing the "culprit" segment in all sequences and the latter removing entire sequences. As shown in Table 2, sensitivity of these filtering methods was < 25%, except for arbitrary insertions, for which BMGE was very efficient (~ 97%), and to a lesser extent OD-seq (~ 59%). The performance of BMGE was not surprising, since the insertion of a random segment typically constitutes a divergent block. In contrast, the specificity of block- and outlier-filtering methods (Table 2) was generally higher than the specificity of segment-filtering methods. This was especially true for Guidance (~ 99%), and probably attributable to these methods removing less data than segment-filtering methods. Indeed, as expected from their rationale, methods that filter outlier blocks or outlier sequences appear by design far less sensitive to primary sequence errors than segment filtering methods.

Sources of HmmCleaner false positives

As shown in Fig. 3, specificity of HmmCleaner is lower than its sensitivity and is also more variable across MSAs. When genuinely homologous segments are highly divergent, i.e., display a weak similarity to other sequences of the MSA, false positives are unavoidable. Accordingly, specificity was higher in UARs (Fig. 2) than in AARs (Additional file 1 Figure S1). A refined analysis shows that HmmCleaner specificity decreases with the gap frequency (Additional file 1 Figure S7A), the evolutionary rate (Additional file 1 Figure S7B) and the fraction of AARs (Additional file 1 Figure S7C). This confirms that its low specificity is due to evolutionary divergence.

Such a negative correlation between sequence divergence and HmmCleaner specificity can be due to: (i) the presence of overlooked primary sequence errors in our datasets, (ii) the presence of alignment errors that would result in detection errors, (iii) the detection of segments corresponding to insertion events, or (iv) the detection of homologous but divergent segments that look like primary sequence errors (see above). For hypothesis one to be true, we should observe approximately the same false positive rate in both UARs and AARs. Yet, this was not the case (Fig. 2 and Additional file 1 Figure S1). Similarly, for hypothesis two to be true, we would expect an important impact of the aligner software on the false positive rate. This was not the case either (Additional file 1 Figure S3). Therefore, the last two hypotheses should explain most of the observed false positives.

To confirm this interpretation, we ran HmmCleaner on raw MSAs, i.e., without introducing errors, and characterized the segments detected. We considered segments detected in regions with ≥70% of gaps as linked to insertion events. Those segments accounted for 14% of all detected segments. The mean pairwise identity of the remaining segments was mainly distributed between 10 and 30% (Fig. 5), with an average of 19%, indicating HmmCleaner false positives consisted almost exclusively of low similarity segments. Interestingly, the identity window of 10 to 30% is known as the "twilight zone" in structural biology, a zone in which defining homology based on sequence identity alone is hazardous at best. Using known protein structures to define homology, Rost [25] concluded that "above a cut-off roughly corresponding to 30% sequence identity, 90% of the pairs were homologous; below 25% less than 10% were". Accordingly, the low similarity segments detected by HmmCleaner, even if they do not correspond to true primary sequence errors, are extremely difficult to align and likely contain alignment errors. In contrast, only a small fraction of the segments detected by HmmCleaner (1.8%) had a mean pairwise identity ≥40%. This tiny minority could be considered as the "real" HmmCleaner false positives.

Detection of positive selection in the presence of primary sequence errors

Having addressed the efficiency of HmmCleaner at dealing with simulated primary sequence errors, we can now assess its usefulness on empirical evolutionary inferences, such as detection of positive selection, topological accuracy and inference of branch lengths. The presence of a primary sequence error in a MSA (i.e., a highly divergent segment) is expected to severely increase the number of non-synonymous substitutions, thereby creating a strong signal for positive selection in the branch leading to the corresponding sequence. To test this idea, we simulated an out-of-frame segment of length 10 to 50 aa into 116 MSAs from the OrthoMAM database, before realigning the MSA (MAMMALIA dataset, see Materials and Methods). Over 10 replicates of the simulation, branch-specific positive selection detected by the standard likelihood ratio test method of Nielsen and Yang [26, 27] increased from 8.28 to 95.69% (Table 3). The few cases with non-significant likelihood ratio test results corresponded to shorter erroneous segments (~ 20 versus ~ 30 aa), which could have been introduced into divergent regions. As expected, a primary sequence error generates a strong erroneous signal of positive selection.

The use of HmmCleaner on the 116 raw MSAs decreased branch-specific positive selection from 8.28 to 3.88% (Table 3). A detailed manual analysis of the MSAs in which the signal for positive selection had disappeared generally found the presence of structural annotation errors that were correctly detected and removed by HmmCleaner. The remaining 3.88% of significant likelihood ratio test results may correspond to real positive selection or to structural annotation errors not detected by HmmCleaner. Importantly, the use of HmmCleaner on the MSAs in which we had introduced primary sequence errors drastically reduced the detection of positive selection (7.76%). This value was slightly higher than the control (3.88%) and likely due to incomplete removal of the errors by HmmCleaner, in agreement with the sensitivity estimated above (Figs. 3-4). PREQUAL performed similarly to HmmCleaner but was less efficient at reducing the detection of positive selection after error insertion (13.62%), in agreement with its

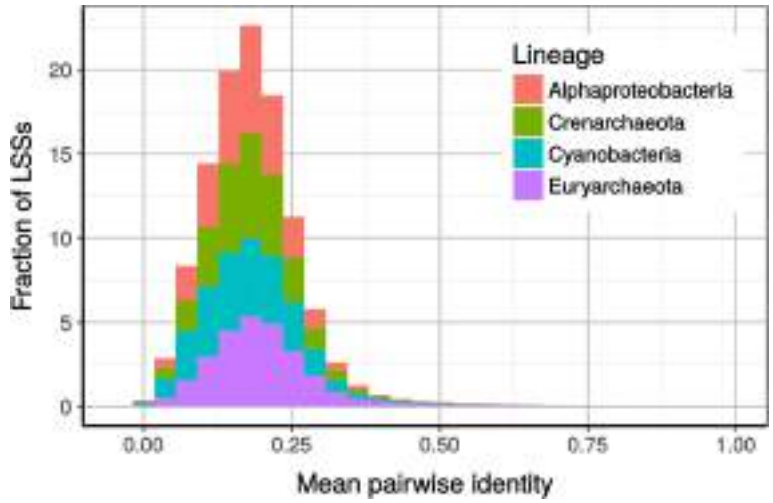


Fig. 5 Mean pairwise identity of the low similarity segments (LSSs) detected on 400 raw MSAs (no simulation) by HmmCleaner used with the default scoring matrix

Table 3 Detection of positive selection in the presence of primary sequence errors

	Raw	HmmCleaner	PREQUAL	BMGE	TrimAl
MSAs with positive selection in a targeted sequence	8.28%	3.88%	2.67%	7.24%	7.15%
MSAs with positive selection in a targeted sequence with an error	95.69%	7.76%	13.62%	95.00%	94.74%

lower sensitivity. In contrast, the use of block-filtering methods (BMGE and TrimAl) had a negligible effect on MSA containing simulated errors (95 and 94.74% respectively, versus 95.69%), which confirms that such methods are not adapted to the removal of primary sequence errors (Table 2).

Nonetheless, our simulations did not allow us to verify that HmmCleaner behaves correctly in real cases of positive selection. To this end, we selected MSAs with well-established presence of sites showing positive selection [27, 28]. Those cases were primate lysozyme c gene, primate cancer gene BRCA1, MHC from human and angiosperm phytochrome genes. Interestingly, neither HmmCleaner nor PREQUAL did remove any residues from primate or human alignments. In contrast, they did on phytochrome genes but subsequent analyses were as significant on the filtered MSAs than on the raw MSAs. In conclusion, both software did not appear to negatively impact detection of true positive selection, at least at a small evolutionary scale.

Relative effect of primary sequence and alignment errors on phylogenetic inference

To evaluate the relative importance of primary sequence errors and alignment errors on real data, we compared the effect of HmmCleaner, PREQUAL and two block-filtering software (BMGE and TrimAl) on the accuracy of phylogenetic inference. We also examined filtering methods that reduce the stochastic (sampling) error (removal of partial sequences and selection of the longest genes), as this type of error might be critical for single-gene inferences. Two aspects of phylogenetic inference were considered: tree topology and branch lengths. Two datasets of orthologous genes for which the correct species phylogeny is reasonably well established were used: (1) 14,261 genes from mammals, obtained exclusively from genomic data [29], named MAMMALIA, (2) 4593 genes from vertebrates, obtained mainly from transcriptomic data [30], named VERTEB RATA. Our expectation was that primary sequence errors would be more frequent in MAMMALIA than in VER TEBRATA due to incorrect structural annotations of genomic data [31].

Phylogenetic accuracy as measured by the frequency of correctly recovered clades was computed for various conditions (Table 4). Major improvements were observed in three cases: (i) inferences based on nt MSAs over those based on aa MSAs with MAMMALIA (63.19% versus 45.29%), (ii) removal of partial sequences (< 100 aa) in VERTEBRATA (68.70% versus 65.64%) and (iii) use of the longest genes (+ 6.58%, + 8.30% and + 3.21% for nt and aa MAMMALIA and aa VERTEBRATA, respectively). All these cases actually correspond to a reduction of the stochastic error, either due to (i) the increased amount of information present in nt, (ii) the removal of sequences without enough signal to be accurately positioned, or (iii) the larger number of positions, respectively. Stochastic error is therefore the predominant limiting factor for the accuracy of single-gene phylogenies. This is in agreement with Tan et al. [18], who observed that phylogenetic accuracy decreases with the amount of data removed by block-filtering software. This suggests that a filtering method should be highly specific in its removal of erroneous data, otherwise the potential improvement in phylogenetic accuracy could be overthrown by the increase in stochastic error.

Generally speaking, the effect of various filtering methods, including HmmCleaner, on phylogenetic accuracy was limited (Table 4). For VERTEBRATA, BMGE, HmmCleaner and TrimAl all slightly decreased accuracy (64.83%, 65.23% and 65.28% versus 65.64%). The performance of HmmCleaner is interesting, because it removes more residues than BMGE and TrimAl (5.6% versus 4.8% and 1.8%, respectively). HmmCleaner thus appears to discard almost exclusively segments that are poorly informative for inferring phylogeny, which is expected because it removes low similarity segments. Accordingly, a random removal of the same amount of data than HmmCleaner decreased accuracy more severely (1.08% versus 0.41%). Moreover, studying the effect of HmmCleaner on each clade of the vertebrate phylogeny reveals that it slightly improved accuracy within clades mainly represented by species for which genomic data had been used (mammals and birds). The use of the large parameter set, which is justified by the presence of > 50 species in VERTEBRATA MSAs, slightly improved accuracy (65.41% versus 65.23% for default parameters), likely because less data were removed (4.22% versus 5.61%). Similarly, the better specificity of PREQUAL (Table 2 and Additional file 1 Table S1) could explain its slightly higher accuracy (65.73%) and the reduced amount of data removal (3.06%). For the VER TEBRATA dataset, which likely contains few primary sequence errors, the removal of data is slightly deleterious (except for PREQUAL, with an improvement of 0.09%), illustrating how precise data filtering should be.

In contrast, the MAMMALIA dataset demonstrated the positive effect of using HmmCleaner on genomic-based

Table 4 Topological accuracy of single-gene phylogenies

VERTEBRATA		
version	mean	loss (%)
RAW	65.64%	NA
RAW (long)	68.85%	NA
HMM	65.23%	5.61
HMM-L	65.41%	4.22
HMM-LS	65.58%	2.68
PREQUAL	65.73%	3.06
BMGE	64.83%	4.83
TrimAl	65.28%	1.8
HMM Random	64.56%	5.61
HMM + BMGE	63.94%	13.38
HMM + TrimAl	62.90%	13.76
MIN	68.71%	0.71
MIN + HMM	68.67%	6.43
MAMMALIA		
version	mean	loss (%)
RAW (NT)	63.19%	NA
RAW (NT long)	69.77%	NA
HMM (NT)	66.29%	2.92
HMM-L (NT)	66.13%	2.57
HMM-LS (NT)	65.75%	2.12
PREQUAL (NT)	64.77%	2.92
BMGE (NT)	63.59%	3.16
TrimAl (NT)	63.77%	3.76
HMM Random (NT)	62.59%	2.92
HMM + BMGE (NT)	66.56%	4.7
HMM + TrimAl (NT)	66.45%	5.16
RAW (AA)	45.29%	NA
RAW (AA long)	53.59%	NA
HMM (AA)	46.98%	2.92
HMM-L (AA)	46.65%	2.57
HMM-LS (AA)	46.48%	2.12
PREQUAL (AA)	45.63%	2.92
BMGE (AA)	45.45%	3.16
TrimAl (AA)	45.36%	3.76
HMM Random (AA)	44.56%	2.92
HMM + BMGE (AA)	46.52%	4.7
HMM + TrimAl (AA)	46.64%	5.16

[1]mean: average frequency of correctly recovered clades, [2]loss: fraction of residues removed from the raw MSAs. [3]long: Values for the half longest MSAs. See the legend of Fig. 6 for the complete set of abbreviations

datasets, which are more likely to contain annotation errors: accuracy improved from 63.19% to 66.30% for nt MSAs. BMGE and TrimAl also increased accuracy, but less than HmmCleaner (63.59% and 63.77%, respectively), while the random removal of characters expectedly decreased accuracy (62.59%). PREQUAL is in between (64.77%), probably because of its reduced sensitivity (all the more so that structural annotation errors are often shared by unrelated taxa and PREQUAL performed poorly when multiple errors are present in a given alignment (Additional file 1 Table S1). The same pattern was observed for aa sequences (Table 3).

Finally, since segment- and block-filtering methods have different targets (primary sequence and alignment errors, respectively), it could be of interest to combine them, as already done in practice for recent large phylogenomic matrices [30, 32, 33]. To test this, we applied BMGE and TrimAl on the MAMMALIA and VERTEB RATA alignments already cleaned by HmmCleaner. Data loss was important, especially for VERTEBRATA (~ 13.5%), potentially increasing the impact of stochastic error. Accordingly, for VERTEBRATA, the combination of the two types of filters show the lowest accuracy among all our analyses. In contrast, for MAMMALIA, the accuracy increased when both filters were applied versus when a single one was used: from ~ 64.7% to ~ 66.5% (nt) and ~ 45.5% to ~ 46.5% (aa). The comparison of segment+block-filtering with segment-filtering was more ambiguous: the accuracy increased for nt MSAs (from 66.29% for HmmCleaner to 66.56% for HmmCleaner+BMGE, the best accuracy observed for MAMMALIA) but decreased for aa MSAs (from 46.98% for HmmCleaner to 46.52% for HmmCleaner+BMGE). These contrasted results illustrate the difficulty of data filtering, data loss increasing stochastic error while decreasing reconstruction errors.

In conclusion of this section, HmmCleaner is more efficient than BMGE and TrimAl, and to a lesser extent than PREQUAL, at improving topological accuracy for genome-based MSAs, whereas filtering methods slightly decrease accuracy for transcriptome-based MSAs. When primary sequence errors are not negligible, the increase of stochastic error due to data filtering is overcome by the reduction of non-phylogenetic signal. More generally, the better performance of segment-filtering methods (HmmCleaner and PREQUAL) versus block-filtering methods (BMGE and TrimAl) suggests that primary sequence errors (especially annotation errors) are more detrimental to phylogenetic inference than alignment errors.

Since primary sequence errors had only limited (yet detectable) impact on topological accuracy, we wondered if errors could not be "buffered" by the lengthening of the terminal branches leading to the erroneous sequences. To study this possibility, we computed the

correlation coefficient between the branch lengths of each single gene and the branch lengths of the concatenated tree, by constraining single-gene trees to the topology of the concatenation. As the genes under study were orthologous, the correlation coefficients were expected to be high [32], fluctuating only because of stochastic sampling noise and heterotachy [34]. For both MAMMALIA and VERTEBRATA aa datasets (Fig. 6), the average correlation coefficients were 0.662 and 0.710, respectively, but 0.778 for MAMMALIA nt MSAs. Again, stochastic error due to the loss of information generated by translation appears as a key factor. In contrast to topological accuracy, the improvement in correlation provided by HmmCleaner was similar for the three datasets (aa VERTEBRATA: 0.066, aa MAMMALIA: 0.089, nt MAMMALIA: 0.079). Interestingly, for MAMMALIA, the average correlation coefficient for cleaned aa MSAs was similar to the one of raw nt MSAs (0.749 and 0.778). PREQUAL performed similarly to HmmCleaner, but was slightly less efficient, even for VERTEBRATA (Fig. 6), likely because of its lowest sensitivity. In sharp contrast, block-filtering methods (BMGE and TrimAl) had virtually no impact, even when applied after HmmCleaner. Segment-filtering methods seem thus to be more efficient than block-filtering methods to remove primary sequence errors affecting branch-length estimates. This is in agreement with our hypothesis that primary sequence errors, which are the target of HmmCleaner and PREQUAL, are more detrimental to evolutionary inferences than alignment errors, which are the target of BMGE and TrimAl.

Finally, we examined the effect of filtering software on the branch lengths of the concatenated trees. All pairwise comparisons (e.g., BMGE versus HmmCleaner) yielded correlation coefficients > 0.98. Interestingly, for

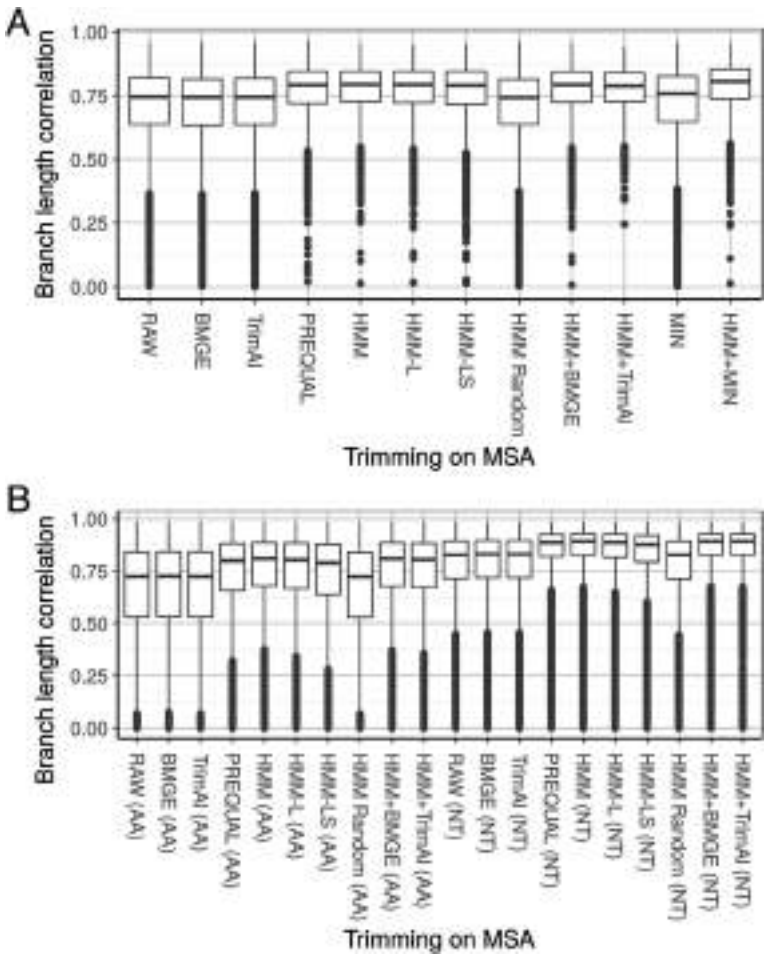


Fig. 6 Distribution of correlation coefficients of branch lengths between single gene-tree and the corresponding concatenated tree in different configurations on the VERTEBRATA dataset (**a**) and the MAMMALIA dataset (**b**). RAW: raw MSAs, BMGE: after BMGE with loose settings, TriamAl: after TrimAl in gappy-out configuration, PREQUAL: after PREQUAL, HMM: after HmmCleaner with default preset, HMM-L: after HmmCleaner with large preset, HMM Random: after removing the same number of residues per sequence as HmmCleaner would do but at random, HMM + BMGE: running BMGE after HmmCleaner, HMM + TrimAl: running TrimAl after HmmCleaner, MIN: after removing sequences with < 100 aa, HMM + MIN: combination of HMM then MIN. See Additional file 1: Table S2 for mean values

VERTEBRATA, a few outliers were identified, corresponding to shortened branch lengths in HmmCleaner-based phylogenies for species from which genomic data had been used (e.g., *Ornithorhynchus*, *Takifugu* or *Latimeria*). We interpret these differences as the result of the removal of structural annotation errors. The negligible impact of filtering software on correlation coefficients in the case of concatenation is likely due to the law of large numbers. In contrast, the tree length (or total branch length) of the concatenated trees was severely modified by all filtering software. For aa supermatrices, the tree length without filtering was 4.46 (14.71) for MAMMALIA (VERTEBRATA). It decreased to 3.85 (11.08) with BGME and to 3.38 (9.07) with HmmCleaner. In agreement with their objectives, this suggests that filtering methods are efficient at removing the more divergent residues that increase tree length. Interestingly, HmmCleaner reduced tree length more than BMGE. As HmmCleaner and BMGE both removed similar numbers of residues (3.1% and 5.6% versus 3.2% and 4.8%, see Table 3), HmmCleaner appears to target divergent residues (due to either primary sequence error or fast evolutionary rate) more efficiently than BMGE.

Conclusion

In this article, we presented a new version of HmmCleaner, a software package that automatically identifies and removes low similarity segments in MSAs with the purpose of limiting the negative effect of primary sequence errors on evolutionary inferences. The performance of our method was investigated through analyses of both simulated and empirical data. HmmCleaner shows an excellent sensitivity to primary sequence errors ≥12 aa in length in simulations. Its specificity to simulated errors is also high, with its false positives mostly corresponding to insertions or low similarity segments that would be difficult to handle in subsequent steps of analysis.

We showed that segment-filtering software (HmmCleaner and PREQUAL) have more positive effects on evolutionary inferences (detection of branch-specific positive selection, topological accuracy and branch-length estimation) than the commonly used block-filtering software (BMGE and TrimAl). This suggests that primary sequence errors are more detrimental to evolutionary analyses than alignment errors. Therefore, we argue that the efforts of the research community should address both alignment and primary sequence errors, in other words that more energy should be devoted on structural annotations. In this respect, HmmCleaner proved to be efficient at pointing them out, being slightly more sensitive than the recently developed PREQUAL [20].

Given the pervasiveness of primary sequence errors, we recommend the use of segment-filtering methods in high-throughput analyses of eukaryotic genomic data. On the long run, it would be interesting to evaluate whether HmmCleaner (or other equivalent toolssuch as PREQUAL) could replace block-filtering software. For now, HmmCleaner targets low similarity segments that are by essence difficult to align and therefore may decrease the frequency of alignment errors, possibly to the extent of making them negligible. In this respect, the advantage of specifically removing erroneous segments instead of entire blocks is to reduce the amount of data lost for the subsequent analyses, hence limiting the rise in stochastic error, which we have shown to be the major limiting factor for the accuracy of single-gene phylogenies.

Methods

HmmCleaner algorithm

HmmCleaner detects low similarity segments in four steps (Fig. 1). First, we create a profile HMM (pHMM) based on the observed data (the MSA) (Fig. 1a). The pHMM is a model of the ancestral sequence that can generate all the observed sequences. It is built with the HMMER function hmmbuild using default options with two exceptions. We change the fragtresh option, giving equal weight to each sequence (--fragtresh = 0) and we apply a Laplace + 1 prior instead of the default mixture Dirichlet prior (option --plaplace). In our method, the pHMM can either be built upon (i) all sequences of the MSA (complete strategy) or (ii) all sequences except for the one being analyzed (leave-one-out strategy).

Second, we estimate the probability that the pHMM generates each amino acid of a given sequence of the MSA, with the hypothesis that a primary sequence error will have a very low probability. To do so, each sequence of the MSA is evaluated with the pHMM using hmmsearch with default options, which yields profile-sequence alignments (Fig. 1b). HMMER performs this step following a heuristic of homology search at the end of which it defines a set of subsequences (envelopes) estimated to fit a part of the pHMM. Each envelope is then precisely fitted to the profile using the full Forward/Backward algorithm and a maximum expected accuracy alignment is returned [19]. Those alignments allow us to identify which segments of each sequence of the MSA are expected to have been generated by the pHMM, and within each segment, they provide the posterior probability that a specific amino acid has been generated by the pHMM as well as the level of match of each amino acid to the consensus of the pHMM. We used the four discrete categories of match/mismatch defined by HMMER (highlighted in yellow in Fig. 1b), instead of the posterior probability, because preliminary analyses showed that this strategy was more efficient to detect primary sequence errors. This is probably because posterior probability depends both on the quality of the match and on the quality of the alignment around the site while our method

focuses solely on the match quality with the assumption that the alignment is correct.

The first two categories defined by HMMER represent residues that do not match the pHMM consensus: blank character (log-odds negative score based on emission probability compared to background frequency) and '+' character (log-odds positive score based on emission probability compared to background frequency, which could be considered a conservative substitution). The last two categories represent residues that match the pHMM consensus: amino acid characters in lowercase (emission probability < 50%) and uppercase (emission probability > 50%). These characters are used to create a string of match/mismatch for HmmCleaner purposes. Segments of this string corresponding to subsequences that do not fit the pHMM (and thus are missing from HMMER output) are filled with blank characters, so as to have a full-length representation of each sequence.

Third, a cumulative similarity score is calculated for each sequence, based on scoring of the four categories of the match/mismatch string. Since we expect that a primary sequence error will mainly consist of mismatches, scoring parameters c1 (blank) and c2 ('+') are negative whereas parameters c3 (lower case residue) and c4 (upper case residue) are positive (Fig. 1c). The cumulative similarity score increases when the residue is expected by the pHMM and decreases otherwise (Fig. 1d), representing the evolution of the sequence fit to the pHMM along the sequence. It is computed from left to right, starting at a maximal value of 1, representing a perfect fit to the pHMM, and it is strictly comprised between 0 and 1 included.

Fourth, a low similarity segment is defined wherever the cumulative score reaches zero. Its start is set after the last position where the score was 1, while its end is defined by the last position of the segment where the score was null or by the end of the sequence (Fig. 1d).

Dataset creation

To optimize the parameters and study the performance of HmmCleaner, we created four datasets by assembling MSAs of protein-coding genes sampled from four different prokaryotic lineages (Alphaproteobacteria, Cyanobacteria, Euryarchaeota, and Crenarchaeota). We chose prokaryotes to minimize the presence of annotation errors, as these lineages are mostly devoid of introns, simplifying the structural annotation of their genes. Yet, a few structural annotation errors will likely subsist, in particular due to sequencing errors, incorrect start codon predictions, programmed ribosomal frameshifts and programmed transcriptional realignments [35]. For each lineage, we retrieved annotated RefSeq genomes from NCBI FTP (61 Alphaproteobacteria, 195 Cyanobacteria, 42 Crenarchaeota and 179 Euryarchaeota) and used the corresponding proteomes to define orthogroups with OrthoFinder (E-value = 10e-5; inflation parameter = 1.5) [36]. In order to maximize the proportion of true orthologous sequences, only orthogroups with at least 75% of our taxon sampling and < 10% of multiple copies were selected. To minimize the sequence length heterogeneity in the orthogroups, we studied the length distribution of the clusters and kept only those having a mean sequence length ≥ 100 aa and with < 5% outliers in the sequence length distribution. Outliers were defined as sequences having a length shorter than the mean length minus 1.96 times the standard deviation. In addition, we removed these outlier sequences from the retained orthogroups. To assemble the four final datasets of 100 MSAs each, we selected at random 100 orthogroups for each of the four lineages, and aligned their sequences with MAFFT 7.309 [21] (L-INS-i algorithm, 5000 iterations). Since our simulations introduce frameshifts in nt sequences (see below), we transferred the alignment gaps from protein sequences to the corresponding nt sequences.

To study the impact of HmmCleaner on evolutionary inferences, we used two additional datasets assembled from animal sequences. The first dataset (MAMMALIA) corresponded to the 14,261 orthologous genes with ≥50% of 43 species present from OrthoMAM v9 [29]. As both nt and amino-acid alignments were available for download, we used both types of sequences. In contrast, only amino-acid sequences were available for the second dataset (VERTEBRATA). The latter corresponded to the 4593 orthologous genes from Irisarri et al. [30]. Because these authors had used filtering softwares during their dataset construction, we had to re-apply their last step (selection of a single sequence per organism and construction of chimeric sequences when necessary) using SCaFoS [37] on a pre-filtering version of the corresponding MSAs.

Simulator

To study the properties of HmmCleaner, we developed a simulator designed to create primary sequence errors in protein MSAs. In a first step, it takes an existing protein-coding alignment of nt sequences and randomly introduces a primary sequence error in a specified number of sequences. Primary sequence errors can be of three types, (i) a frameshift followed by the opposite (compensatory) mutation after a predefined number of out-of-frame codons, (ii) a scrambled segment resulting from the shuffling of individual nucleotides over a predefined number of codons or (iii) the arbitrary insertion of a segment shuffled as in (ii). Then, it translates all sequences to proteins (ignoring STOP codons, mapped to x characters) and realigns them using MAFFT 7.309 [21] (L-INS-i algorithm, 5000 iterations). In a second step, HmmCleaner is run on the resulting MSA and the detected low similarity segments are compared to the locations of the simulated errors to quantify the number of true positives, false positives, false negatives and true

negatives. To allow a fine-grained analysis of the behavior of HmmCleaner, our simulator further characterizes the context of each position of the original MSA by its gap frequency, substitution rate and conservation level, as determined by block-filtering software. More precisely, we used BMGE [9] at three different stringency settings (strict, entropy cutoff of 0.4 and gap cutoff of 0.05; medium, 0.5 and 0.2 corresponding to default parameters; loose, 0.6 and 0.4). As Gblocks [6] yields similar results, only BMGE is considered in this article (data not shown).

Parameter optimization

To optimize the four parameters of the scoring matrix of HmmCleaner, we simulated frameshift errors on the two large datasets (Cyanobacteria and Euryarchaeota). For each nt MSA, 4 subsets of sequences of different sizes (5, 10, 25 and 50 sequences) were drawn 100 times at random. On each of these samples, 1 to 5 sequences were randomly affected by a primary sequence error of length 10 to 100 aa. HmmCleaner was then run on each resulting amino-acid MSA (complete strategy) under 2835 different combinations of its four parameters. There were 9, 7, 9 and 5 possible values, respectively, for c1 (– 0.05 to – 0.25 by step of 0.025), c2 (– 0.02 to – 0.08 by step of 0.01), c3 (0.05 to 0.25 by step of 0.025) and c4 (0.4 to 0.6 by step of 0.05). These ranges were defined based on preliminary simulations aimed at thoroughly exploring the zone of high specificity and high sensitivity. Moreover, BMGE was run on each simulated MSA to allow a partitioned analysis of the results between ambiguously aligned regions (AARs) and unambiguously aligned regions (UARs).

To ensure that our parameter optimization was robust, we studied the impact of introducing variations in our simulation protocol. First, we focused on a single lineage at a time (either Euryarchaeota or Cyanobacteria) and observed that both sensitivity and specificity computed across the 2835 quartets of parameter values were highly correlated between these two lineages (Pearson correlation coefficient > 0.99, Additional file 1 Figure S2A-B). Second, we compared the results obtained with different operational definitions of UARs. The strictest and the most relaxed configurations of BMGE settings were also highly correlated (> 0.99, Additional file 1 Figure S2C-D). Third, we considered the potential impact of the number of sequences in the MSAs (5, 10, 25 or 50). In this case, we expected a larger effect owing to the dependence of pHMM statistical power on the amount of observations available to build the models. Specificity showed a high correlation (> 0.95) while comparing MSAs with 5 sequences to those with 50 sequences (Additional file 1 Figure S2F). In contrast, sensitivity was more affected, and the correlation coefficient dropped to 0.78 (Additional file 1 Figure S2E). In agreement with our intuition, sensitivity was always better with 50 than with 5 sequences. However, the parameter quartets leading to high sensitivity for 5 sequences were the same to yield high sensitivity for 50 sequences, indicating that the number of species does not much impact the optimization of HmmCleaner parameters.

Characterization of HmmCleaner performance

To characterize the impact of the length and number of frameshift errors on HmmCleaner sensitivity and specificity, we ran simulations at 4 predetermined error lengths (10, 33, 66 and 100 aa) and with 4 different numbers of sequences affected (1, 5, 10 or 15). For each of these 16 combinations, 100 simulations were performed on each of the 100 MSAs of the 4 lineages using subsets of 25 randomly drawn sequences. HmmCleaner was run on these 160,000 MSAs per prokaryotic lineage (640,00 MSAs in total) using the default scoring matrix and the complete strategy. Ten additional simulations were carried out in the same conditions (16 combinations of error characteristics on subsets of 25 sequences) to compare HmmCleaner with the four scoring matrix and PREQUAL. PREQUAL was run with default parameters and without the removal of repeated regions. For the high-resolution analysis of the impact of the error length on sensitivity, the same type of simulation was run with only one sequence affected by a primary sequence error of length ranging from 1 to 33 aa.

Additional simulations were carried out to expand the observations obtained on frameshift errors to the other types of primary sequence errors that our simulator can generate. At the same time, we extended our dataset to MSAs of eukaryotic species by selecting 112 alignments from the MAMMALIA dataset (MSA with > 25 sequences out of the 116 genes used for positive selection, see below) and 170 alignments from VERTEBRATA for which we retrieved nucleotide sequences (MSAs with less than 3 missing species). For each type of errors (frameshifts, scrambled segments and arbitrary insertions), subsets of 25 sequences were drawn out of the complete MSA and 1 to 5 sequences were affected by an error of length between 10 to 100 aa. This simulation was run 100 times per MSA for HmmCleaner (default), BMGE (loose) and PREQUAL, and 10 times for OD-SEQ (default parameters) and GUIDANCE2 (default threshold to determine outlier sequences).

To study the characteristics of the low similarity segments detected by HmmCleaner, so as to characterize the sources of its false positives, it was run with the default scoring matrix and the complete strategy on the raw MSAs from the four prokaryotic lineages, as MSAs of eukaryotic lineages are more likely to contain real primary sequence errors (mainly incorrect structural annotation). For each detected segment, we computed

the gap frequency in the corresponding region of the MSA and its mean pairwise identity. Pairwise identity itself was considered as computable when ≥10% of the low similarity segment residues were facing a residue (and not a gap) in the opposite sequence. Likewise, mean pairwise identity was computed only when ≥10% of the pairs were computable.

Effect of HmmCleaner on evolutionary inferences

Analyses of positive selection were performed on a subset of the MAMMALIA dataset. To reduce structural annotation errors, we selected the 446 nt MSAs with branch length R^2 above 0.95 (computed as in Simion et al. 2017 [32], see below). To limit the computational burden and to introduce errors of a sizable length representing only a few percent of the sequences, we selected the 116 MSAs between the first quartile and the median on the MSA width distribution. For each MSA, we simulated one primary sequence error of random length (10 to 50 aa) at a randomly chosen position of a randomly chosen sequence 10 times and aligned with MAFFT. For each simulation, we tested for positive selection in the affected branch for 10 versions of the corresponding MSA: (1) the original MSA, (2) the original MSA cleaned by 5 filtering software configurations (HmmCleaner, PREQUAL, BMGE and TrimAl), (3) the erroneous MSA, and (4) the erroneous MSA cleaned by 5 filtering software configurations (HmmCleaner, PREQUAL, BMGE and TrimAl). Detection of positive selection was performed using a likelihood ratio test between two models [26, 27]: model A, in which ω estimation is free (i.e., allowing positive selection), and model B, in which ω is fixed to 1 (i.e., no selection). Likelihood values for both models were obtained using codeml (both models: runmode = 0, method = 0, clock = 0, model = 2, CodonFreq = 2, NSsites = 2, fix_kappa = 0, kappa = 2; model A: fix_omega = 0, omega = 0.2; model B: fix_omega = 1, omega = 1). Positive selection was considered present when the likelihood ratio test between models A and B returned a value > 13.82 (Chi-square critical value for alpha = 0.001 and 2 degrees of freedom).

To test the effect of filtering methods on the accuracy of single-gene phylogenies, we used the orthologous genes of the datasets MAMMALIA and VERTEBRATA. Eleven different filtering setups were considered: (1) RAW: without any alteration the MSA, (2) HMM: HmmCleaner with the default scoring matrix (for nt MSA, low similarity segments were detected on the corresponding protein MSAs and then reported), (3) HMM-L: HmmCleaner with the "large" scoring matrix, (4) PREQUAL: PREQUAL with default parameters, (5) BMGE: BMGE in loose settings, (6) TrimAl: TrimAl with gappy-out option, (7) HMM Random: removal of the same number of residues per sequence as HmmCleaner would have done but at random, (8) HMM + BMGE: running BMGE as in 5 after HmmCleaner, (9) HMM + TrimAl: running TrimAl as in 6 after HmmCleaner, (10) MIN: removal of the sequences with < 100 residues and (11) HMM + MIN: a combination of running HmmCleaner then removing sequences as in MIN.

Single-gene trees were inferred with RAxML v8 [38] with the PROTGAMMALGF model for protein MSAs and the GTRGAMMA model for nt MSAs. Frequencies of correctly recovered clades were computed with a custom script comparing the single-gene trees to the topology of Irisarri et al. [30] for the VERTEBRATA dataset and to a concatenated tree of the 137 most complete aa MSAs inferred with PhyloBayes-MPI [39] using the CAT+G model [40] for the MAMMALIA dataset. These two topologies are in agreement with existing knowledge of vertebrate relationships, even if ambiguities persist for a few nodes (e.g., the relative position of Xenarthra and Afrotheria).

For each MSA, branch lengths were computed with RAxML using the same model as previously while constraining the topology to the respective reference species tree. Single-gene branch lengths were then compared to the branch lengths of the reference tree, after pruning the species missing in the MSA under study. Finally, the correlation coefficient of the two sets of branch lengths was computed with a custom script.

Additional file

Additional file 1: Table S1. Comparison of mean sensitivity and mean specificity between HmmCleaner presets and PREQUAL. **Table S2. Mean R2 value for branch length**. **Figure S1.** Mean sensitivity and specificity of HmmCleaner towards detection of primary sequence errors introduced in ambiguously aligned regions (AARs). **Figure S2.** Impact of the conditions of simulation on sensitivity (A,C,E) and specificity (B,D,F) of HmmCleaner. **Figure S3.** Impact of the multiple alignment software on sensitivity (A) and specificity (B) of HmmCleaner used with the default scoring matrix. **Figure S4.** Impact of the HmmCleaner algorithm on its sensitivity (A) and specificity (B) when used with the default scoring matrix. **Figure S5.** Impact of the length and number of primary sequence errors, and of the prokaryotic lineage, on sensitivity (A,C,E) and specificity (B,D,F) of PREQUAL. A,B. **Figure S6.** Impact of the conservation context of introduced primary sequence errors on sensitivity of HmmCleaner used with the default scoring matrix for different error lengths. **Figure S7.** Impact of the conservation context of introduced primary sequence errors on specificity of HmmCleaner used with the default scoring matrix for different numbers of errors. (PDF 1324 kb)

Abbreviations

aa: Amino acids; AAR: Ambiguously aligned region; LRT: Likelihood ratio test; MSA: Multiple sequences alignment; nt: Nucleotides; pHMM: Profile hidden Markov model; UAR: Unambiguously aligned region

Acknowledgments

We thank Nicolas Lartillot for his numerous insights throughout the method development, as well as for comments on the manuscript, Henner Brinkmann for his help in method validation, Rik Verdonck for critical reading, and two anonymous reviewers for helpful comments. Computations were made on the supercomputers Mp2 and Ms2 from the Université de Sherbrooke, managed by Calcul Québec and Compute Canada. The operation of this supercomputer is funded by the Canada Foundation for

Innovation (CFI), the ministère de l'Économie, de la science et de l'innovation du Québec (MESI), and the Fonds de recherche du Québec - Nature et technologies (FRQ-NT).

Funding
This work was supported by Canadian Research Chair program and Natural Sciences and Engineering Research Council and by the TULIP Laboratory of Excellence (ANR-10-LABX-41).

Availability of data and materials
HmmCleaner is written in Perl 5 and is based on Bio::MUST modules (https://metacpan.org/author/DBAURAIN), aimed at providing seamless integration with the MUST ecosystem [41]. It can be downloaded from CPAN at the following address (https://metacpan.org/release/Bio-MUST-Apps-HmmCleaner) or can be directly installed through the various CPAN clients using the module name Bio::MUST::Apps::HmmCleaner. Datasets used in this study are available at the following address: https://doi.org/10.6084/m9.figshare.6004250.v1

Authors' contributions
ADF implemented the new version of the software package, carried out the simulations and positive selection analyses and wrote the manuscript; RP implemented the first version of the software package; DB helped with the implementation of the new version of the software package and wrote the manuscript; HP helped with the implementation of both version of the software package, carried out the phylogenetic inferences and wrote the manuscript. All authors read and approved the final manuscript

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Station d'Ecologie Théorique et Expérimentale de Moulis, CNRS, Moulis, France. [2]Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, Montréal, Québec, Canada. [3]InBioS–PhytoSYSTEMS, Unité de Phylogénomique des Eucaryotes, Université de Liège, Liège, Belgium.

Received: 10 August 2018 Accepted: 2 January 2019
Published online: 11 January 2019

References

1. Chatzou M, Magis C, Chang J-M, Kemena C, Bussotti G, Erb I, et al. Multiple sequence alignment modeling: methods and applications. Brief Bioinform. 2016;17:1009–23.
2. Wong KM, Suchard MA, Huelsenbeck JP. Alignment uncertainty and genomic analysis. Science (80-.). 2008;319:473–476.
3. Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, et al. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 2011;9.
4. Schneider A, Souvorov A, Sabath N, Landan G, Gonnet GH, Graur D. Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. Genome Biol Evol. 2009;1:114–8.
5. Markova-Raina P, Petrov D. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. Genome Res. 2011;21:863–74.
6. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000;17:540–52.
7. Dress AW, Flamm C, Fritzsch G, Grünewald S, Kruspe M, Prohaska SJ, et al. Noisy: identification of problematic columns in multiple sequence alignments. Algorithms Mol Biol. 2008;3:7.
8. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–3.
9. Criscuolo A, Gribaldo S. BMGE (block mapping and gathering with entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol. 2010;10:210.
10. Kück P, Meusemann K, Dambach J, Thormann B, von Reumont BM, Wägele JW, et al. Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. Front Zool. 2010;7:1–12.
11. Wu M, Chatterji S, Eisen JA. Accounting for alignment uncertainty in phylogenomics. PLoS One. 2012;7:1–10.
12. Sela I, Ashkenazy H, Katoh K, Pupko T. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. Nucleic Acids Res. Oxford University Press. 2015;43:W7–14.
13. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. 2007;56:564–77.
14. Jordan G, Goldman N. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. Mol Biol Evol. 2012;29:1125–39.
15. Privman E, Penn O, Pupko T. Improving the performance of positive selection inference by filtering unreliable alignment regions. Mol. Biol. Evol. Oxford University Press. 2012;29:1–5.
16. Karin EL, Susko E, Pupko T. Alignment errors strongly impact likelihood-based tests for comparing topologies. Mol Biol Evol. 2014;31:3057–67.
17. Spielman SJ, Dawson ET, Wilke CO. Limited utility of residue masking for positive-selection inference. Mol Biol Evol. 2014;31:2496–500.
18. Tan G, Muffato M, Ledergerber C, Herrero J, Goldman N, Gil M, et al. Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. Syst Biol. 2015;64:778–91.
19. Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol. 2011;7.
20. Whelan S, Irisarri I, Burki FPREQUAL. Detecting non-homologous characters in sets of unaligned homologous sequences. Bioinformatics. 2018:1–2.
21. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
22. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.
23. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal omega. Mol Syst Biol. 2011;7:539.
24. Jehl P, Sievers F, Higgins DG. OD-seq: outlier detection in multiple sequence alignments. BMC bioinformatics. BioMed Central. 2015;16:269.
25. Rost B. Twilight zone of protein sequence alignments. Protein Eng Des Sel. 1999;12:85–94.
26. Nielsen R, Yang Z. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics. 1998;148:929–36.
27. Zhang J, Nielsen R, Yang Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 2005;22:2472–9.
28. Yang Z, Swanson WJ. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol Biol Evol. 2002;19:49–57.
29. Ranwez V, Delsuc F, Ranwez S, Belkhir K, Tilak MK, Douzery EJP. OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. BMC Evol Biol. 2007;7:1–12.
30. Irisarri I, Baurain D, Brinkmann H, Delsuc F, Sire JY, Kupfer A, et al. Phylotranscriptomic consolidation of the jawed vertebrate timetree. Nat Ecol Evol. 2017;1:1370–8.
31. Sharma V, Hiller M. Increased alignment sensitivity improves the usage of genome alignments for comparative gene annotation. Nucleic Acids Res Oxford University Press. 2017;45:8369–77.
32. Simion P, Philippe H, Baurain D, Jager M, Richter DJDJ, Di Franco A, et al. A large and consistent Phylogenomic dataset supports sponges as the sister group to all other animals. Curr Biol. 2017;27:958–67.
33. Amemiya CT, Alföldi J, Lee AP, Fan S, Philippe H, MacCallum I, et al. The African coelacanth genome provides insights into tetrapod evolution. Nature Nature Publishing Group. 2013;496:311–6.
34. Lopez P, Casane D, Philippe H. Heterotachy, an important process of protein evolution. Mol Biol Evol Oxford University Press. 2002;19:1–7.
35. Sharma V, Firth AE, Antonov I, Fayet O, Atkins JF, Borodovsky M, et al. A pilot study of bacterial genes with disrupted ORFs reveals a surprising profusion of

protein sequence recoding mediated by ribosomal frameshifting and transcriptional realignment. Mol Biol Evol. 2011;28:3195–211.
36. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16:1–14.
37. Roure B, Rodriguez-Ezpeleta N, Philippe H. SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. BMC Evol Biol. 2007;7:1–12.
38. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.
39. Lartillot N, Rodrigue N, Stubbs D, Richer J. PhyloBayes MPI : phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst Biol. 2013;62:611–5.
40. Lartillot N, Philippe HA. Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol. 2004;21:1095–109.
41. Philippe H. MUST, a computer package of management utilities for sequences and trees. Nucleic Acids Res. 1993;21:5264–72.

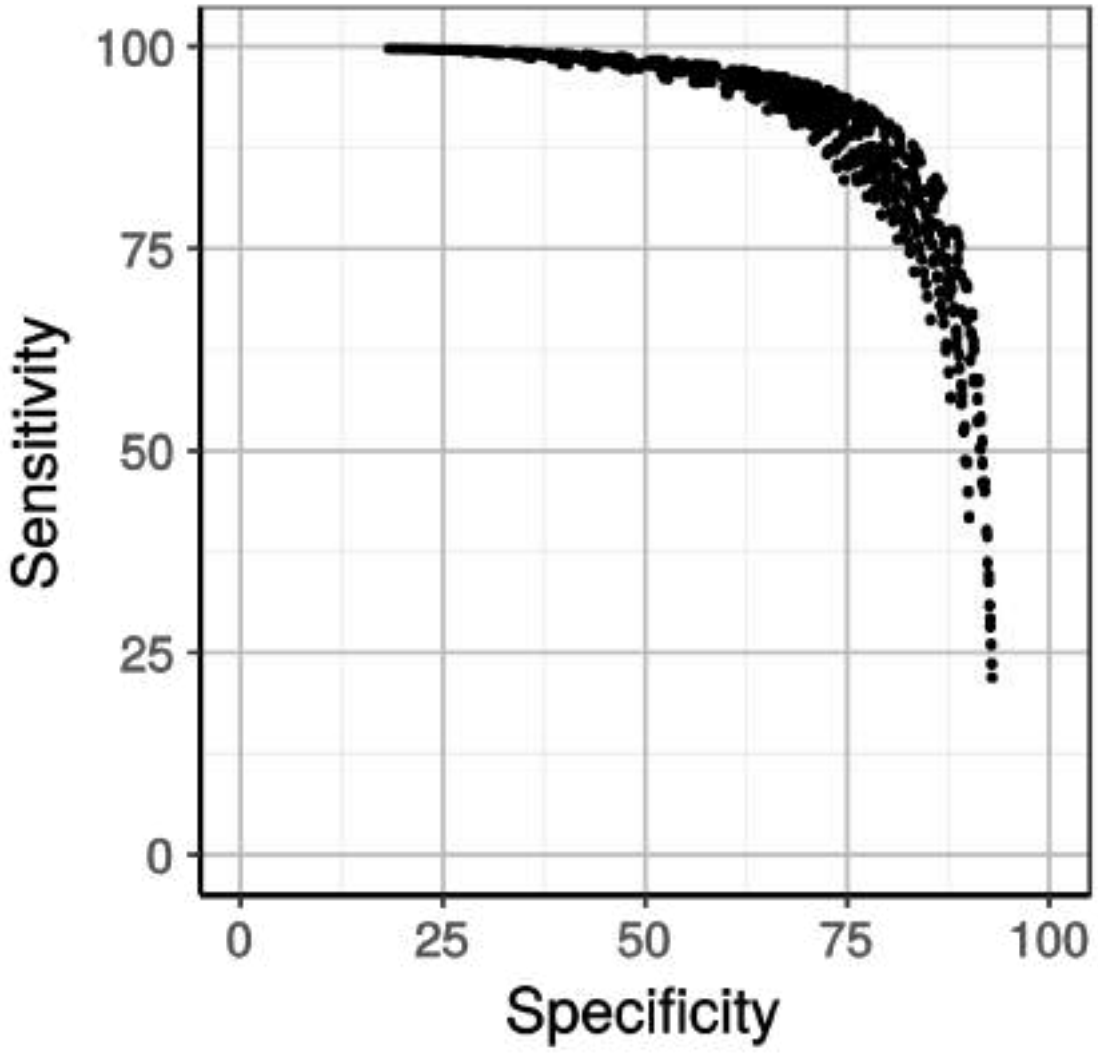
Supplementary Table 1. Comparison of mean sensitivity and mean specificity between HmmCleaner presets and PREQUAL

Sensibility	Overall	Error length				Error number				Lineage			
		10	33	66	100	1	5	10	15	Alpha-proteobacteria	Crenarcheota	Cyanobacteria	Euryarcheota
PREQUAL	83.33% (15.92)	60.72% (10.36)	91.34% (6.23)	91.70% (8.26)	89.56% (10.73)	87.68% (17.44)	85.63% (15.05)	81.90% (14.50)	78.11% (14.84)	84.76% (14.77)	83.82% (16.69)	79.72% (16.96)	85.02% (14.56)
Default	93.30% (9.11)	80.35% (8.86)	97.36% (3.04)	97.80% (3.06)	97.67% (3.36)	91.7% (11.04)	93.75% (8.38)	93.89% (8.23)	93.84% (8.30)	92.90% (8.78)	93.52% (8.84)	92.69% (10.64)	94.07% (7.91)
Large	90.16% (11.07)	74.16% (9.28)	95.81% (3.82)	95.85% (4.34)	94.80% (5.36)	88.32% (12.58)	90.28% (10.59)	90.96% (10.32)	91.06% (10.42)	89.05% (10.62)	90.66% (10.93)	90.24% (12.36)	90.67% (10.16)
Specificity	86.10% (17.41)	58.49% (11.76)	94.93% (3.99)	95.69% (4.27)	95.27% (4.93)	84.65% (17.87)	86.36% (17.26)	86.68% (17.20)	86.69% (17.25)	84.89% (16.51)	86.52% (17.55)	85.59% (19.55)	87.38% (15.71)
Large_ Specificity	76.20% (26.59)	32.96% (12.47)	90.65% (6.76)	91.64% (7.22)	89.55% (9.00)	73.77% (26.37)	76.23% (26.08)	77.19% (26.80)	77.63% (26.98)	73.94% (24.39)	77.15% (27.31)	76.61% (29.31)	77.12% (24.97)
Specificity	Overall	Error length				Error number				Lineage			
		10	33	66	100	1	5	10	15	Alpha-proteobacteria	Crenarcheota	Cyanobacteria	Euryarcheota
PREQUAL	92.42% (6.34)	93.83% (6.09)	92.44% (6.31)	92.26% (6.44)	92.16% (6.51)	92.99% (6.00)	92.67% (6.18)	92.24% (6.41)	91.80% (6.71)	90.28% (8.25)	92.61% (6.04)	94.99% (4.15)	91.82% (5.26)
Default	86.70% (10.28)	87.13% (9.93)	86.97% (10.04)	86.62% (10.33)	86.07% (10.80)	87.64% (9.81)	87.01% (10.10)	86.36% (10.41)	85.79% (10.70)	86.23% (11.08)	85.39% (11.41)	90.88% (7.46)	84.29% (9.46)
Large	90.98% (7.37)	91.40% (7.02)	91.23% (7.12)	90.87% (7.43)	90.41% (7.84)	91.78% (6.92)	91.23% (7.22)	90.69% (7.45)	90.21% (7.77)	90.20% (7.95)	90.46% (8.32)	93.72% (5.29)	89.53% (6.79)
Specificity	91.80% (6.99)	92.13% (6.73)	91.96% (6.82)	91.71% (7.02)	91.38% (7.34)	92.48% (6.63)	92.03% (6.85)	91.56% (7.05)	91.12% (7.33)	91.39% (7.29)	91.23% (8.07)	94.14% (5.14)	90.41% (6.53)
Large_ Specificity	94.65% (4.88)	95.00% (4.61)	94.72% (4.77)	94.55% (4.93)	94.33% (5.19)	95.18% (4.57)	94.85% (4.75)	94.47% (4.94)	94.09% (4.75)	94.11% (5.01)	94.51% (5.78)	96.13% (3.61)	93.86% (4.56)

Supplementary Table 2. Mean R2 value for branch length

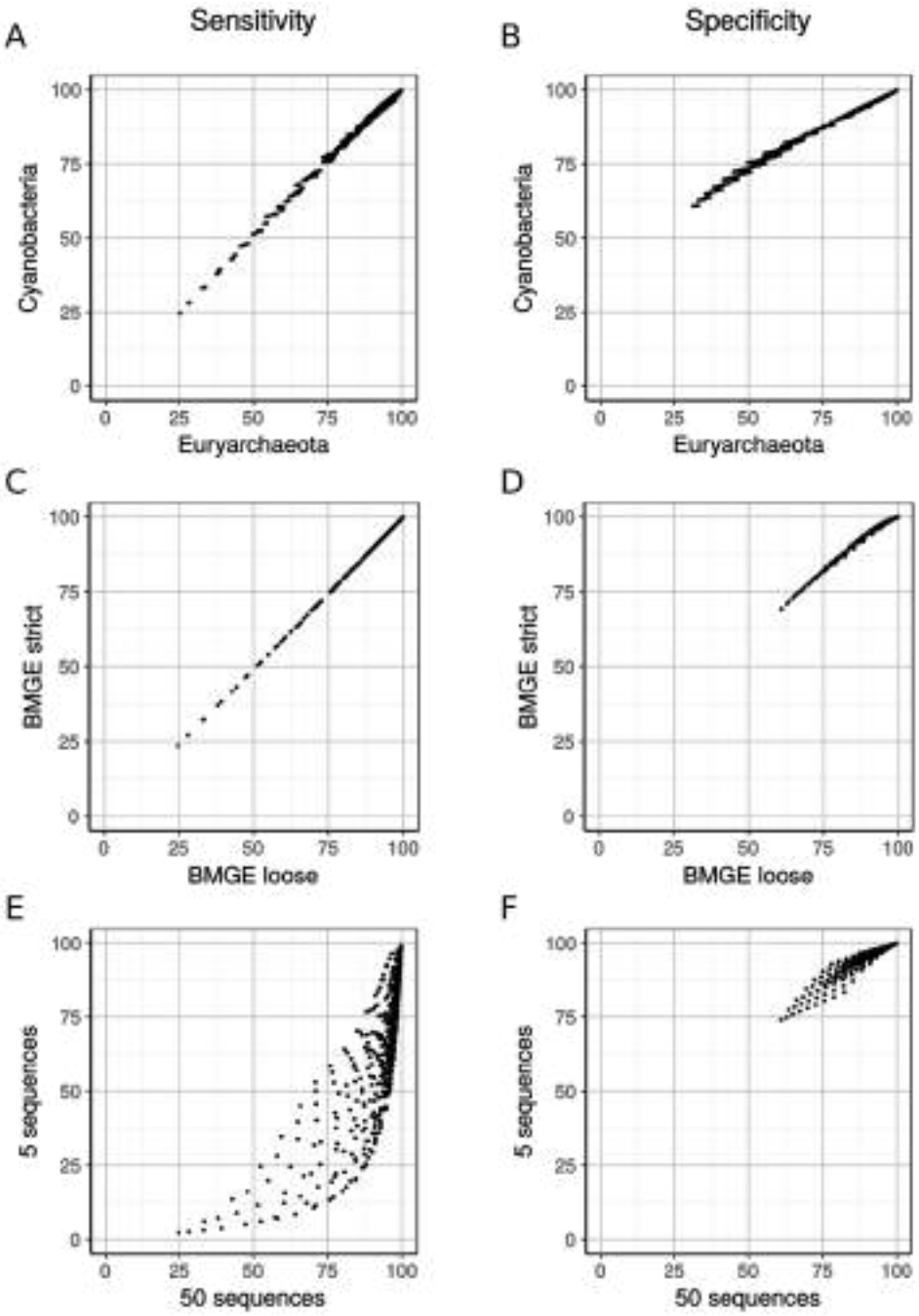
VERTEBRATA	
version	mean R2 BL
RAW	0.709
BMGE	0.705
TrimAl	0.707
PREQUAL	0.770
HMM	0.775
HMM-L	0.773
RANDOM	0.705
HMM+BMGE	0.773
HMM+TrimAl	0.771
MIN	0.716
MIN+HMM	0.787

MAMMALIA	
version	mean R2 BL
RAW (AA)	0.660
BMGE (AA)	0.662
TrimAl (AA)	0.660
PREQUAL (AA)	0.736
HMM (AA)	0.749
HMM-L (AA)	0.740
HMM Random (AA)	0.659
HMM+BMGE (AA)	0.748
HMM+TrimAl (AA)	0.743
RAW (NT)	0.776
BMGE (NT)	0.778
TrimAl (NT)	0.777
PREQUAL (NT)	0.849
HMM (NT)	0.855
HMM-L (NT)	0.844
HMM Random (NT)	0.775
HMM+BMGE (NT)	0.855
HMM+TrimAl (NT)	0.856



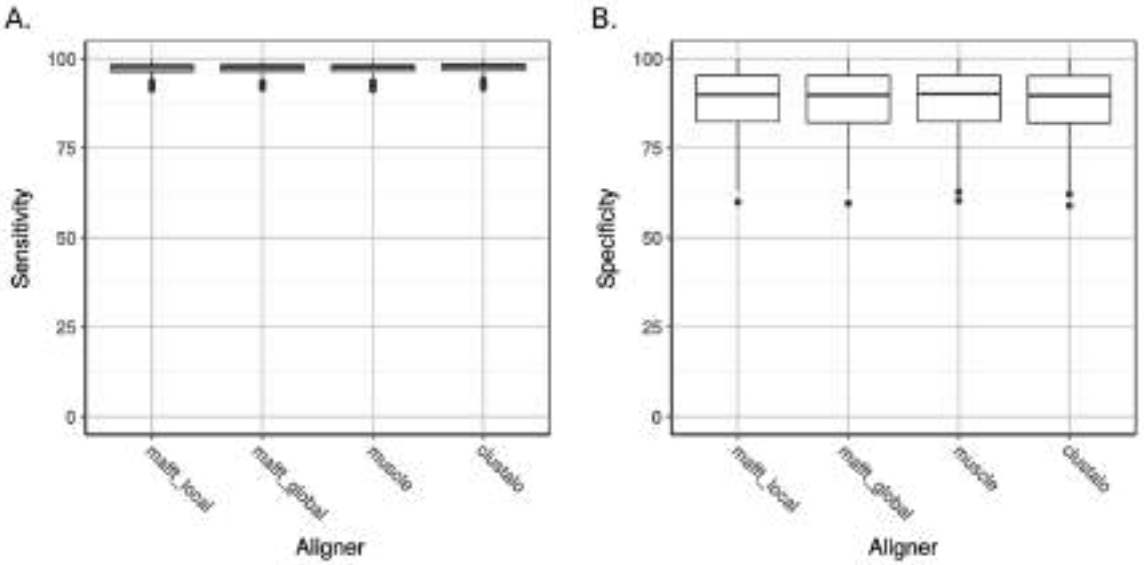
Suppl. Figure 1

Mean sensitivity and specificity of HmmCleaner towards detection of primary sequence errors introduced in ambiguously aligned regions (AARs). Each dot corresponds to the two means of the values obtained across 80,000 simulations and 3 operational definitions of AARs for one of the 2835 combinations of the 4 parameters of the scoring matrix.



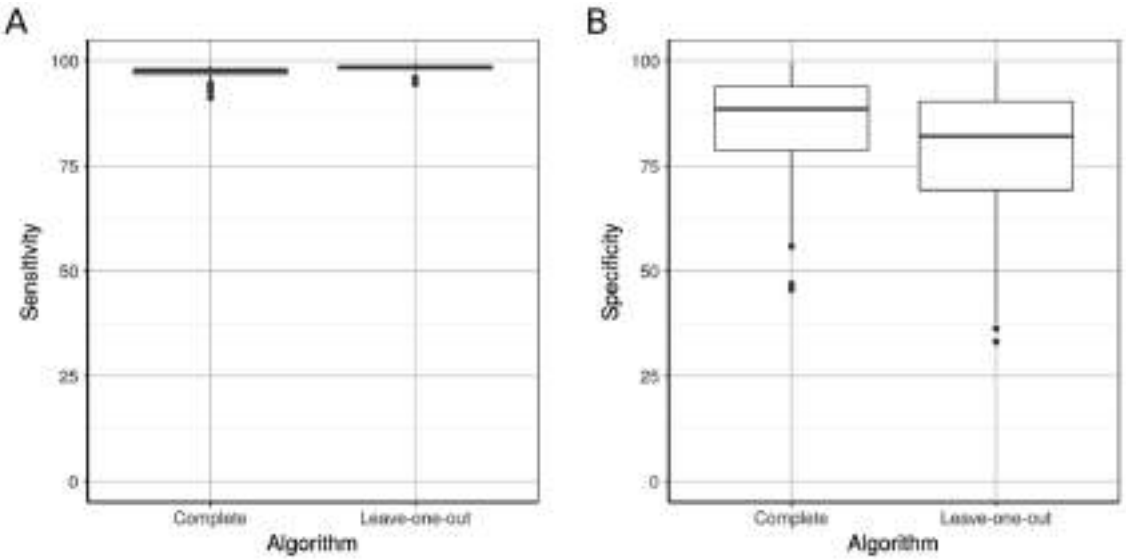
Suppl. Figure 2

Impact of the conditions of simulation on sensitivity (A,C,E) and specificity (B,D,F) of HmmCleaner. A,B. Comparison between simulations using MSAs from Euryarchaeota and Cyanobacteria. C,D. Comparison between determination of UARs with BMGE using loose and strict settings. E,F. Comparison between simulations using MSAs of 5 and 50 sequences.



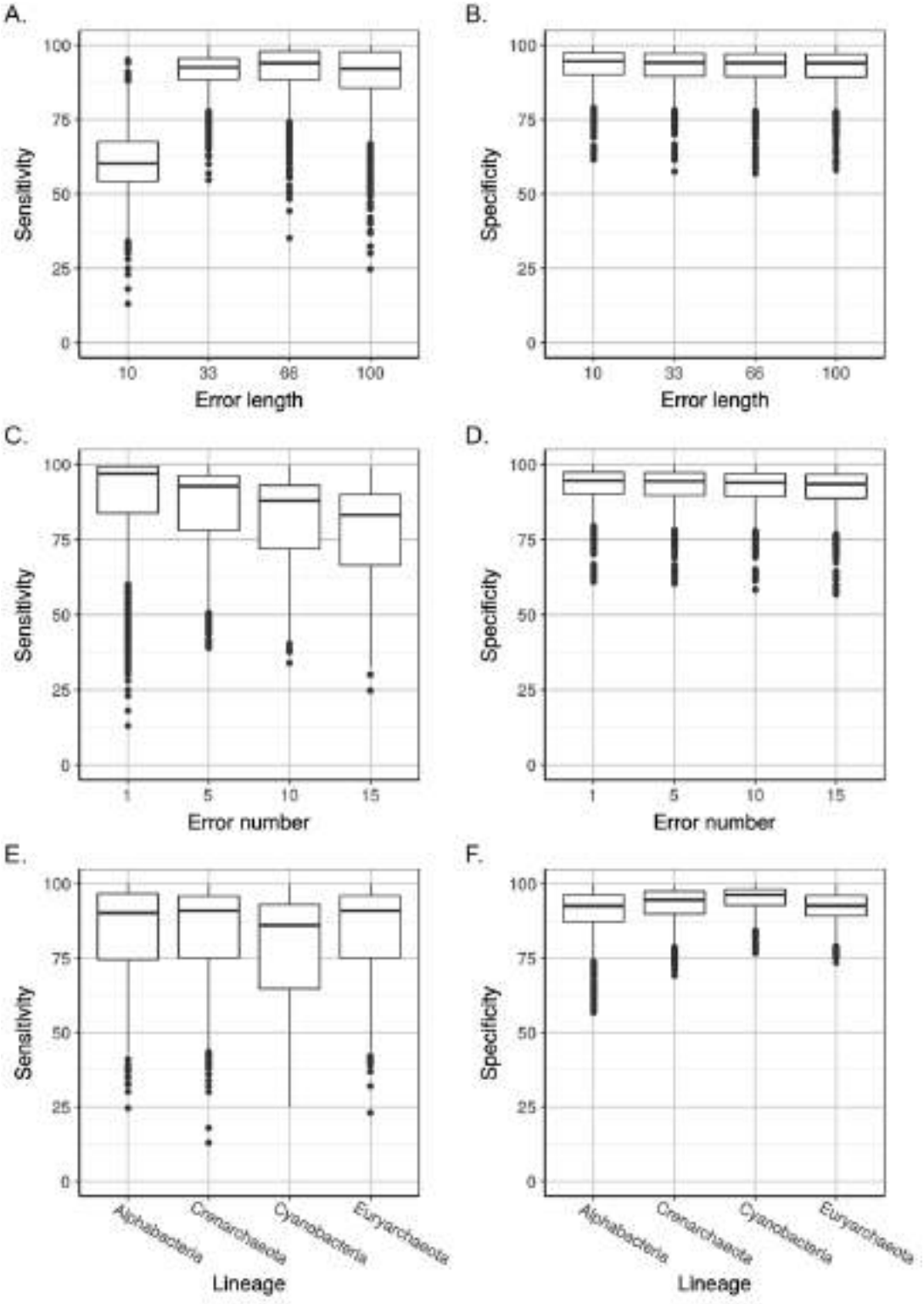
Suppl. Figure 3

Impact of the multiple alignment software on sensitivity (A) and specificity (B) of HmmCleaner used with the default scoring matrix. The compared aligners were MAFFT with L-INS-i algorithm (mafft_local, default aligner), MAFFT with G-INS-i algorithm (mafft_global), MUSCLE and Clustal Omega. Simulations were run on MSAs with 25 species (either Alpha-proteobacteria or Crenarchaeota), by introducing 1 to 5 primary sequence errors of length 10 to 100 aa, and both AARs and UARs were considered when computing the statistics. Box-plots were computed across all considered MSAs.



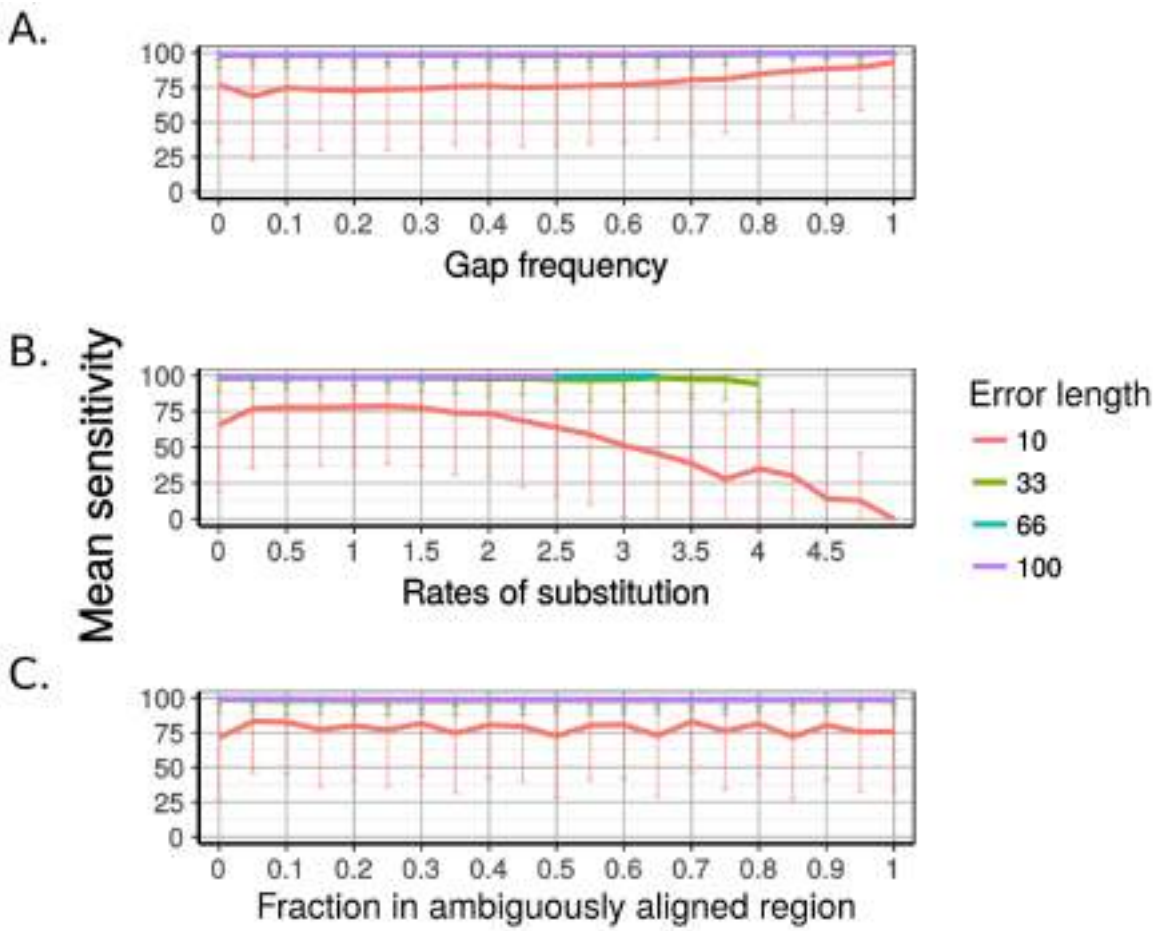
Suppl. Figure 4

Impact of the HmmCleaner algorithm on its sensitivity (A) and specificity (B) when used with the default scoring matrix. The compared algorithms were the “complete strategy” (using all sequences to build the pHMM) and the “leave-one-out strategy” (using all sequences but the analyzed one). Simulations were run on MSAs with 25 species (either Alpha-proteobacteria or Crenarchaeota), by introducing 4 predetermined numbers of primary sequence errors (1, 5, 10 and 15) of 4 specific lengths (10, 33, 66 and 100 aa), and both AARs and UARs were considered when computing the statistics. Box-plots were computed across all considered MSAs.



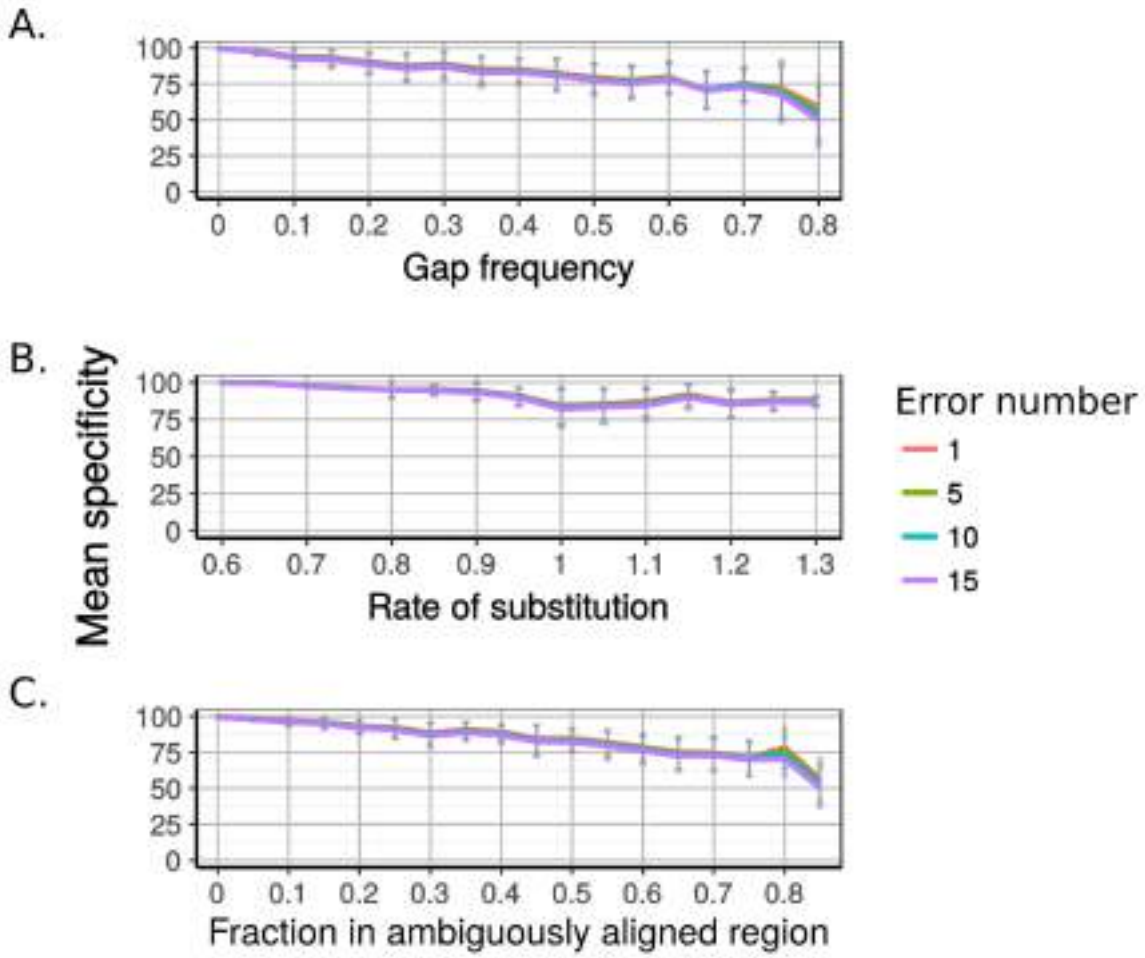
Suppl. Figure 5

Impact of the length and number of primary sequence errors, and of the prokaryotic lineage, on sensitivity (A,C,E) and specificity (B,D,F) of PREQUAL. A,B. Effect of primary sequence error length. C,D. Effect of the number of primary sequence errors. E,F. Effect of the prokaryotic lineage. Box-plots were computed across all considered MSAs and values are means averaged over the different conditions of simulation.



Suppl. Figure 6

Impact of the conservation context of introduced primary sequence errors on sensitivity of HmmCleaner used with the default scoring matrix for different error lengths. A. Sensitivity relative to the mean gap frequency in the region of insertion. B. Sensitivity relative to the mean rate of substitution in the region of insertion. C. Sensitivity relative to the fraction of the region of insertion defined as AAR by BMGE (loose settings).



Suppl. Figure 7

Impact of the conservation context of introduced primary sequence errors on specificity of HmmCleaner used with the default scoring matrix for different numbers of errors. A. Specificity relative to the mean gap frequency in the MSA. B. Specificity relative to the mean rate of substitution in the MSA. C. Specificity relative to fraction of the MSA defined as AAR by BMGE (loose settings).

Chapitre 2

Phylogénomique des Eucaryotes

2.1 Résumé du chapitre 2

Ce seconde chapitre se base sur des résultats présentés lors de la conférence Jacques Monod ayant eu lieu à Roscoff en mai 2016. J'y présente les différentes méthodologies appliquées lors de la création d'un jeu de données de gènes orthologues dans le but d'inférer la phylogénie des eucaryotes. J'y considère les méthodes automatiques de définition d'orthogroupes basées sur la recherche de similarité entre séquences, mais également la contamination des données de séquençage, la séparation entre groupes de séquences paralogues et l'élimination des séquences divergentes. Je profite de la création de ce jeu de données pour inférer la phylogénie des eucaryotes et comparer les phylogénies obtenues sous différents modèles d'évolution, ainsi que les supports statistiques chez les différents groupes d'eucaryotes photosynthétiques.

2.2 Introduction

La photosynthèse oxydative est un processus bioénergétique présent chez de nombreux groupes d'eucaryotes. Les organismes la pratiquant représentent la majorité des Diaphoretickes, l'un des deux domaines eucaryotes [Burki et al. 2008 ; Adl et al. 2018], et un sous-ensemble du super-groupe des Excavata [Hampl et al. 2009]. L'étude de l'acquisition de ce processus est donc fortement liée à l'étude de la phylogénie des eucaryotes. Cette dernière a connu de nombreuses réévaluations au cours des dernières années, portées par la démocratisation des nouvelles techniques de séquençage et par l'usage de la phylogénomique [Bapteste et al. 2002 ; Rodríguez-Ezpeleta et al. 2005 ; Burki et al. 2008 ; Hampl et al. 2009 ; Katz et al. 2011 ; Burki et al. 2012 ; Brown et al. 2013 ; Cavalier-Smith et al. 2014 ; He et al. 2014 ; Cavalier-Smith et al. 2015a ; Cavalier-Smith et al. 2015b]. Cependant, plusieurs noeuds de l'arbre des Diaphoretickes restent non résolus, comme le positionnement des Haptista et Cryptista ou l'ordre de spéciation chez les Archaeplastida (Rhodophyta, Viridiplantae et Glaucophyta) et les SAR (Stramenopiles, Alveolata et Rhizaria) [Archibald 2009].

Les incongruences observées entre les études et la faible résolution de ces noeuds pourraient être attribuées à des considérations méthodologiques [Jeffroy et al. 2006 ; Philippe et al. 2011b ; Galtier et Daubin 2008]. En effet, le signal non phylogénétique introduit par les violations du modèle d'évolution des séquences employé peut mener à des incongruences fortement soutenues [Jeffroy et al. 2006]. Ces violations sont d'autant plus vraisemblables que les noeuds d'intérêt font référence à des événements de spéciation relativement anciens (>1 milliard d'années, [Parfrey et al. 2011]) pour lesquels la saturation du processus de substitution est probable [Whitfield et Lockhart 2007].

Outre les modèles considérés, l'échec de l'intégration de marqueurs strictement orthologues au jeu de données analysé affecte également l'inférence phylogénétique [Roy 2009 ; Philippe et al. 2011b ; Struck 2013]. L'identification de telles séquences parmi l'ensemble de séquences homologues est un exercice compliqué, également affecté par l'ancienneté des noeuds considérés [Kuzniar et al. 2008]. De plus, les organismes photosynthétiques étant riches en gènes latéralement transférés, ces derniers pouvant provenir de l'endosymbionte plastidial ou d'autres organismes [Qiu et al. 2013b ; Qiu et al. 2013a], on peut supposer que ce problème pourrait être d'autant plus ardu lorsque l'on considère les phylums photosynthétiques.

Dans ce chapitre, je me focalise sur le second point abordé, la récupération de gènes orthologues. Ainsi, je présente la construction d'un jeu de données phylogénomique à l'échelle des eucaryotes. Je décris et discute les choix que j'ai effectués à différentes étapes clés de la création du jeu de données, et cela dans le but de récupérer le plus grand nombre

de gènes orthologues avec une large représentation taxonomique. Notamment, j'aborde l'identification de groupes d'orthologie à partir de données génomiques, l'ajout de taxa via des données transcriptomiques (i.e., potentiellement partielles) et la vérification de la relation d'orthologie entre séquences. Au final, j'obtiens une matrice de 594 gènes et 370 espèces, sur laquelle je teste l'impact du choix de modèle en fonction de la quantité d'espèces analysées.

2.3 Définition de groupes de séquences orthologues

Le passage de la phylogénétique à la phylogénomique s'est fait par l'intégration d'un nombre de marqueurs à une échelle génomique dans les analyses de phylogénie moléculaire [Eisen et Fraser 2003]. Dans les faits, les analyses basées sur un ou quelques gènes se sont étendues à l'utilisation de plus d'une centaine de gènes [Delsuc et al. 2005]. On distingue deux manières de procéder à la récupération des gènes orthologues : top-down et bottom-up [Struck 2013]. La première (top-down) se concentre sur la création de graines obtenues à partir de données génomiques pour lesquelles on vérifie la relation d'orthologie entre séquences avant d'en améliorer la couverture taxonomique a posteriori. Ce rajout est généralement réalisé à partir de données transcriptomiques par des méthodes bioinformatiques de recherche de similarité vérifiant un critère de réciprocité (la séquence trouvée à partir d'une graine, i.e. une autre séquence, doit être préférentiellement similaire à cette dernière) [Altschul et al. 1997 ; Ebersberger et al. 2009]. La seconde méthode (bottom-up) consiste à définir l'orthologie entre séquences en considérant directement toutes les séquences protéiques disponibles. Les liens de similarité entre séquences sont déterminés par BLAST en considérant chaque couple de séquences, avant d'être regroupés dans une matrice de similarité. Cette matrice est alors analysée par un algorithme de clustering produisant les groupes d'orthologies supposés [Enright et al. 2002 ; Li et al. 2003 ; Kim et al. 2011].

Les études en phylogénomique des eucaryotes semblent préférer la première approche basée sur l'utilisation d'alignement de séquences pour lesquels l'orthologie a déjà été vérifiée [Rodríguez-Ezpeleta et al. 2005 ; Burki et al. 2007 ; Burki et al. 2008 ; Hampl et al. 2009 ; Burki et al. 2013 ; Brown et al. 2013 ; Cavalier-Smith et al. 2015b]. Ces études réutilisent généralement les alignements d'études précédentes, ces derniers correspondant surtout aux alignements définis par Bapteste et collaborateurs en 2002 et Philippe et collaborateurs en 2004 [Bapteste et al. 2002 ; Philippe et al. 2004]. Ces auteurs ont construit deux jeu de données de 174 alignements à l'échelle des eucaryotes en partant des transcrits obtenus chez Mastigamoeba et Pyropia yezoensis [Nikaido et al. 2000] ou Monosiga ovata pour rechercher les séquences orthologues avec TBLASTN sur

la base de données non-redondante (nr) du National Center for Biotechnology Information (NCBI : `www.ncbi.nlm.nih.gov`). Ces alignements ont subi d'intenses vérifications manuelles (concernant l'orthologie, l'alignement des séquences et les possibles contaminations) et vérifiaient des critères stricts de représentation taxonomique [Philippe et al. 2004]. Les études plus récentes ont commencé à élargir petit à petit cet échantillonnage de marqueurs (202 gènes Burki et al. 2010 et 263 gènes [Burki et al. 2013]), cependant les méthodes employées pour valider ces alignements supplémentaires sont peu détaillées (e.g. voir l'information supplémentaire des articles correspondants). Comme les marqueurs sélectionnés furent utilisés dans la majorité des études, il est difficile d'estimer les possibles biais qui leur sont liés (mais voir [Nosenko et al. 2013]).

Dernièrement, une étude a été publiée mettant l'accent sur la taille importante du jeu de données construit [Katz et Grant 2014], celui-ci ayant été obtenu à partir d'un pipeline entièrement informatisé [Grant et Katz 2014]. La matrice finale se compose en effet de 1554 gènes et 802 taxa dont 232 appartenant aux eucaryotes. Elle ne fut cependant pas analysée dans son entièreté pour l'inférence principale de l'article se concentrant sur la phylogénie des eucaryotes. En effet, cette dernière fut évaluée en se limitant aux 150 gènes présentant l'échantillonnage en eucaryotes le plus complet (ce qui correspond tout de même à 36.346 sites), l'analyse ayant été réalisée avec RAxML (modèle LG+F+Γ4) en effectuant 100 tirages de bootstrap rapide. La matrice complète (entre 55.000 et 65.000 sites en fonction des taxons étudiés) n'eut pas droit à ce traitement et ne fut inférée qu'avec ExaML [Stamatakis et Aberer 2013], une implémentation plus rapide de l'algorithme de RAxML ne permettant pas l'obtention de supports statistiques. Ce défaut fut contourné par les auteures en réalisant des analyses basées sur 20 sous-échantillonnages de sites et de gènes allant jusqu'à 45.000 sites. Malgré tout, ces résultats ne permirent pas de résoudre l'arbre des eucaryotes. Certains noeuds restèrent faiblement supportés (i.e., monophylie des Excavata ou des Plantae) et quelques organismes non-affiliés formèrent des clades dénommés "orphelins". Néanmoins, cette étude a remis en avant d'une part, la récupération automatisée de groupes de séquences orthologues comme source possible de nouveaux marqueurs et d'autre part, la variabilité des supports obtenus en fonction des gènes/sites choisis.

Pour leur pipeline, Grant et Katz (2014) se sont servi des graines obtenues par OrthoMCL [Li et al. 2003] à partir de 98 génomes eucaryotes, 44 bactériens et 16 archéens, contenues dans la base de données OrthoMCL-DB v5 [Chen et al. 2006]. OrthoMCL est un programme de clustering de séquences orthologues basé sur l'algorithme MCL [Enright et al. 2002]. Il a pour particularité de normaliser les scores de similarité entre séquences par espèce afin d'identifier les paralogues anciens. Bien que considéré par Struck (2013) comme faisant partie des programmes liés à une construction de jeu de données de type bottom-up, il est ici employé pour définir des groupes d'orthologies a priori de l'addition

de données transcriptomiques. Ma propre expérience d'utilisation d'OrthoMCL (données non publiées) me fait considérer qu'il reste difficile de s'assurer de la complétude des clusters inférés par OrthoMCL, ainsi que de la validité des relations d'orthologies entre séquences. Ceci est surement dû à la combinaison des différences de vitesse de divergence entre gènes et du caractère très arbitraire du paramètre d'inflation (I), ce dernier servant à déterminer la stringence de la découpe du réseau de similarité (Figure 2.1). En effet, pour une même valeur de I, deux groupes orthologues peuvent être correctement séparés l'un de l'autre alors que deux autres resteront ensemble, formant un groupe constitué de deux paralogues. La récupération de nombreux groupes valides de séquences orthologues peut donc nécessiter la réalisation de multiples inférences avec plusieurs valeurs de I, suivies d'un regroupement des différents résultats afin d'utiliser la valeur optimale de I pour chaque gène.

FIGURE 2.1 – Exemple de l'impact du paramètre d'inflation sur l'algorithm MCL

OrthoFinder [EMMS et KELLY 2015] est un second programme de clustering fondé sur les mêmes principes qu'OrthoMCL. Ce premier vise à corriger deux biais du second. Premièrement, OrthoFinder prend en compte un biais de BLAST ciblant les séquences de petites tailles. En effet, l'algorithme BLAST limite la valeur de score pouvant être atteint par de petites séquences, deux séquences courtes hautement similaires pouvant obtenir un score plus faible que deux séquences longues présentant une plus faible similarité. Ce biais peut empêcher l'identification des clusters de séquences courtes ou encore l'intégration d'un fragment de séquence au bon cluster quand l'entièreté des séquences de deux organismes sont comparés. Deuxièmement, il considère l'impact de la distance phylogénétique entre organismes sur la quantité de similarités pouvant exister entre deux séquences. Effectivement, on s'attend à ce que des organismes distants présentent moins de similarité entre séquences orthologues que des organismes proches, ce qui n'est pas pris en compte par une matrice de similarité. Pour contrôler ces deux biais, OrthoFinder normalise les scores de similarité de chaque paire d'espèces pour différentes catégories de tailles de séquences afin de les rendre moins sensibles aux variations de taille et de distance phylogénétique.

Afin de mieux définir les avantages d'OrthoFinder sur OrthoMCL, j'ai inféré les orthogroupes (OGs, nom donné aux clusters d'orthologies supposés [Emms et Kelly 2015]) à partir de 25 protéomes eucaryotes obtenus au NCBI (voir Annexe 3.1) en utilisant les deux méthodes. Ensuite, j'ai utilisé Forty-Two (`https://bitbucket.org/phylogeno/bio-must-apps-fortytwo/`) dans le but de vérifier la qualité des OGs inférés. Forty-Two est un logiciel visant à enrichir les alignements de séquences préexistants à partir de différents types de données biologiques (génomique, protéomique et transcriptomique) tout en conservant la relation d'orthologie entre séquences. Il fonctionne sur base d'une forme particulière de réciprocité de similarité et sera expliqué plus en détails plus loin (voir partie Forty-Two). J'ai utilisé ce programme pour ajouter, à partir des 25 protéomes précédemment utilisés, les possibles séquences orthologues manquantes à ces OGs. Finalement, j'ai quantifié le nombre de séquences se retrouvant présentes dans plusieurs OGs ainsi que le nombre d'OGs affectés par ces cas de redondance. Si OrthoFinder infère des OGs de meilleures qualités qu'OrthoMCL, je m'attends à ce qu'ils soient plus nombreux et plus riches en séquences. Je m'attends également à minimiser le nombre de séquences ajoutées dans plusieurs OGs.

OrthoMCL et OrthoFinder récupèrent des nombres comparables d'OGs (2241 et 2163 respectivement, Table 2.3) présentant un échantillonnage taxonomique correct (i.e. 60% des organismes représentés) et ne comprenant pas un nombre de séquences trop élevé (moins de 200 séquences). Cependant, pour un nombre légèrement inférieur d'OGs, OrthoFinder comptabilise une quantité bien supérieure de séquences comparée à OrthoMCL (89.401 contre 64.134, soit 40% de plus). Malgré cette différence, Forty-Two ajoute encore un nombre de séquences non négligeables (4081 séquences pour OrthoFinder et 8440 pour OrthoMCL soit 4,6% et 13,2% respectivement). L'observation la plus intéressante est faite sur la quantité d'éléments redondants. Les OGs d'OrthoMCL complétés par Forty-Two contiennent 1041 séquences présentes dans plus d'un OG. Ces séquences redondantes affectent un total de 161 clusters. Cela suggère fortement que les OGs produits par OrthoMCL sont chevauchant et donc que la découpe des OGs est mal réalisée. Cependant, on n'observe aucun élément redondant à travers les 2163 OGs inférés par OrthoFinder. Ce dernier montre donc une bonne complémentarité avec le programme d'enrichissement de séquences utilisé et semble inférer des OGs plus complets par rapport à OrthoMCL.

Statistique des orthogroupes (OGs)	OrthoMCL	OrthoFinder
# OGs comprenant >15 OTUs et <200 séquences	2241	2163
# séquences	64134	89401
Complétude des OGs avec Forty-Two		
# séquences ajoutées sur les OGs considérés	8440	4081
Éléments redondants après rajout avec Forty-Two		
# séquences	1042	0
# OGs	161	0

Afin de mieux visualiser les différences entre OrthoMCL et OrthoFinder, j'ai identifié

les OGs correspondants entre les deux méthodes (obtenus avec les paramètres par défaut, seuil BLAST de 1*10-3). J'ai utilisé une approche simple en réalisant un BLAST entre chaque séquence des 2163 OGs d'OrthoFinder considérés et celles de l'ensemble des OGs d'OrthoMCL. Deux clusters ont été considérés comme correspondants si ils étaient connectés par une paire de séquences présentant une similarité supérieure à une e-value de $1*10^{-10}$. Au total, je retrouve 6518 correspondances connectant les 2163 OGs considérés d'OrthoFinder à 6457 OGs d'OrthoMCL. 1276 OGs d'OrthoFinder sont connectés à plus d'un OG d'OrthoMCL, un seul cluster d'OrthoFinder pouvant être connecté à jusqu'à 36 OGs. À l'inverse, seulement 61 OGs d'OrthoMCL sont connectés à plus de deux OGs d'OrthoFinder, le nombre maximum de connections étant deux. Ces chiffres montrent que les OGs d'OrthoMCL sont très partiels comparés aux OGs d'OrthoFinder.

Parmi les correspondances entre OGs d'OrthoFinder et d'OrthoMCL, on peut noter des cas caractérisés par la présence d'un cluster d'OrthoMCL présentant la majorité des séquences du cluster d'OrthoFinder et une succession de clusters de très petite taille (Figure 2.2A). Ces derniers incluent soit des copies divergentes ou fragments de séquences déjà présentes dans le cluster plus large (OM4 et OM5, clairement inclus dans OF2, sont des paralogues récents de séquences d'*Arabidopsis* et *Physcomitrella* de OM1) ou des séquences appartenant à des espèces évoluant rapidement (OM3 dans OF2 et OM6 dans OF1 contiennent uniquement des séquences de l'algue rouge unicellulaire *Cyanidioschyzon*). La séparation de ces petits clusters par rapport aux deux clusters principaux inférés par OrthoMCL montre l'impact des biais liés à la taille des séquences et à la distance phylogénétique pris en compte par OrthoFinder. De plus, même si deux paralogues sont séparés par une longue branche, OrthoMCL peut malgré tout attribuer une séquence paralogue à un orthogroupe : une séquence unique du cluster OM1 est incorporé au cluster OF1, alors qu'OM1 est à peu près équivalent à OF2 (Figure 2.2A). OrthoFinder récupère donc des clusters plus complets et évite l'intégration de séquence paralogue isolée. Cependant, il a également tendance à trop regrouper, comme le montre la Figure 2.2B avec trois OGs correctement identifiés par OrthoMCL qui sont regroupés en un seul OG par OrthoFinder. Si les différents paralogues n'ont pas assez divergé l'un de l'autre, OrthoFinder va préférer les regrouper en un seul et même OG. Les OGs d'OrthoFinder se rapprochent donc plutôt de groupes complets de séquences homologues séparant potentiellement des paralogues anciens éloignés, ce qui correspond bien à la définition d'orthogroupe des auteurs [Emms et Kelly 2015].

Mes résultats m'ont poussé à préférer l'utilisation d'OrthoFinder à celle d'OrthoMCL pour définir des graines menant à la création d'alignements de séquences orthologues. En effet, des OGs contenant un maximum de séquences sont plus à même d'éviter l'intégration de paralogues isolés, plus difficilement identifiables, qui constitue un comportement fréquent de OrthoMCL. Cependant, il sera vraisemblablement nécessaire de séparer les potentiels

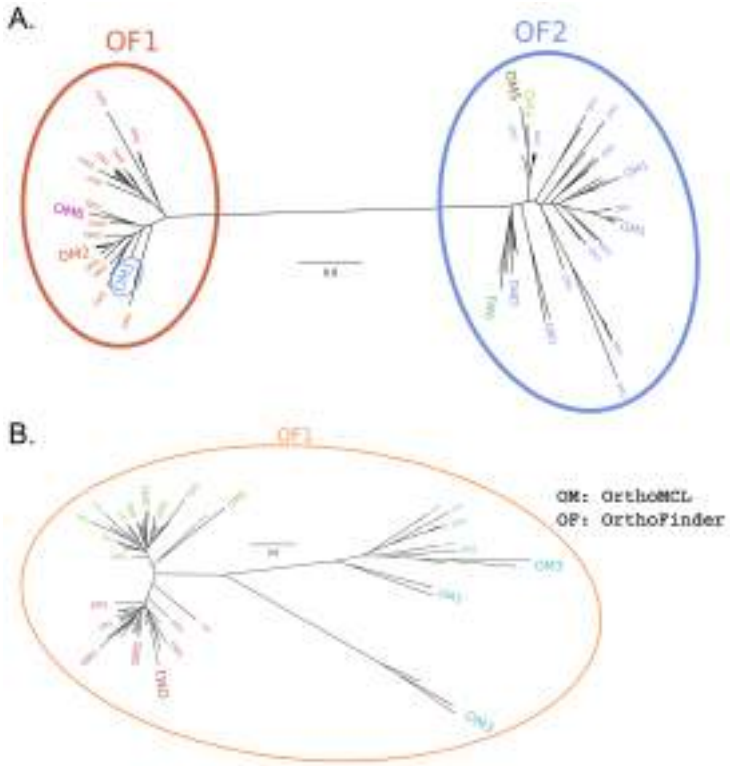


FIGURE 2.2 – Différences entre OrthoMCL et OrthoFinder

paralogues, voire d'éliminer certaines graines si leur division n'était pas évidente. J'ai donc démarré la création de mon jeu de données par l'inférence d'orthogroupes avec OrthoFinder. Je suis parti de 33 protéomes eucaryotes, 28 bactériens et 30 d'archées (voir Annexe 3.2). J'ai introduit des protéomes procaryotes afin de fournir un groupe extérieur pouvant faciliter la détermination des paralogues chez les eucaryotes et pour déterminer de possibles gènes transférés et leur provenance. J'ai réalisé la matrice de similarité en utilisant l'heuristique de BLAST implémentée dans le programme USEARCH [Edgar 2010] avec un seuil d'e-value de $1*10^{-3}$ avant de la fournir à OrthoFinder avec les paramètres par défaut. Le programme a inféré un total de 203.246 OGs que j'ai rapidement réduit aux 3983 OGs respectant les deux conditions suivantes : contenir au moins dix séquences d'eucaryotes différents et avoir au maximum 400 séquences.

J'ai considéré un dernier point avant de finaliser la création des graines. Lors de la découpe du réseau de similarité par l'algorithme MCL, certains clusters d'orthologues peuvent être connectés entre eux par un petit nombre de séquences. Ces dernières présentent un nombre restreint de liens de similarité (Figure 2.3 en bleu clair) et créent des ponts entre clusters orthologues. Elles peuvent soit être exclues des orthogroupes majeurs (elles forment alors un groupe à elles seules) ou bien faciliter le rapprochement de deux clusters d'orthologues. Ces séquences sont des séquences divergentes difficilement attribuables à l'un ou l'autre des clusters d'orthologies et pouvant empêcher la séparation nette des paralogues. Afin de les éliminer, j'ai réalisé un BLAST avec un seuil d'e-value de $1*10^{-5}$ entre chaque séquence

d'un OG pour chaque OG. J'ai éliminé successivement les séquences ne présentant pas de similarité pour 10% puis 30% de l'ensembles des séquences de leur propre OG, ce qui retira 6064 et 6735 séquences respectivement. Après vérification de la présence de 10 eucaryotes différents dans chaque OG, le nombre de graines définitives se porta à 3863.

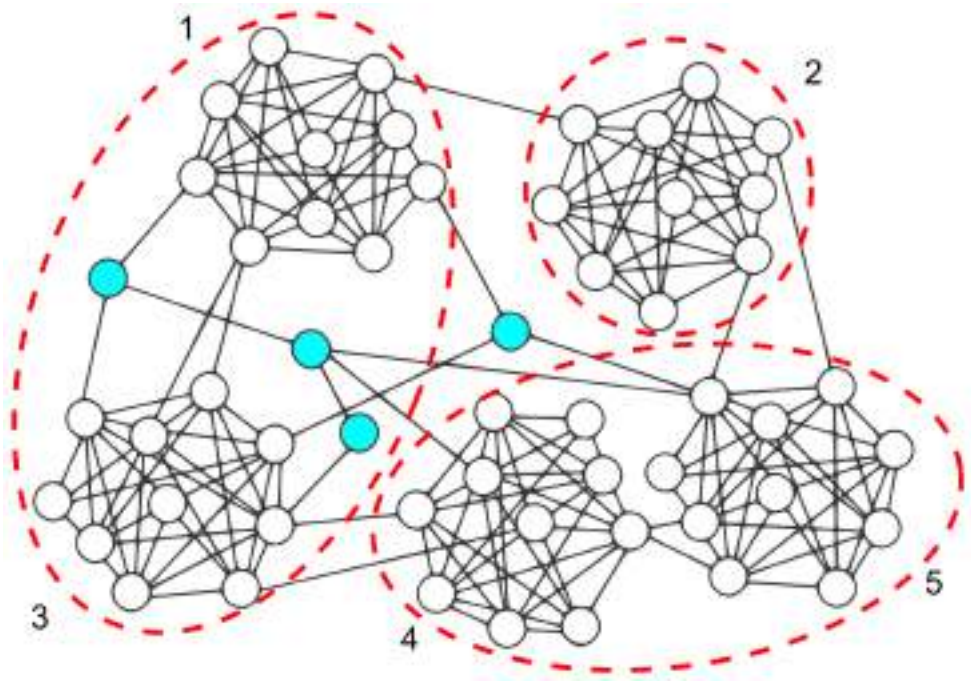


FIGURE 2.3 – Représentation du réseau de similarité connectant une famille de gènes. Les noeuds en bleu représentent des séquences divergentes connectant les groupes de séquences orthologues entre eux. Les traits pointillés représentent une coupure possible réalisée par l'algorithme MCL

2.4 Forty-Two

L'acquisition d'un ensemble d'alignements de gènes orthologues sur base de données génomiques sert de point de départ à la construction d'un jeu de données phylogénomique [STRUCK 2013]. C'est à partir des séquences constituant ces alignements que sont recherchés les orthologues correspondants chez d'autres espèces. Chez les eucaryotes, cette recherche s'opère généralement dans des données transcriptomiques à cause de la faible diversité de génomes disponibles chez les eucaryotes unicellulaires et du coût réduit du RNAseq [SIBBALD et ARCHIBALD 2017]. On utilise alors différentes méthodes bioinformatiques de recherche de similarité entre séquences. On distingue deux principaux types d'algorithmes employés : ceux basés sur l'alignement local entre séquences (exemple : BLAST [ALTSCHUL et al. 1990]), et ceux basés sur l'utilisation d'un modèle de Markov caché (i.e. HMM : Hidden Markov Model, exemple HMMER [EDDY 1998 ; EDDY 2011]).

Une fois les séquences homologues candidates récupérées, il est nécessaire de vérifier leur relation d'orthologie. Cette étape peut être réalisée par leur positionnement dans l'arbre du gène [Burki et al. 2008] ou par leur comparaison à une base de données d'orthologues [Brown et al. 2013]. Ce dernier point se rapproche du principe de réciprocité des meilleurs résultats de similarités [Rivera et al. 1998]. Si, en partant d'une séquence A du protéome de l'espèce X, on obtient comme meilleur résultat de similarité, i.e. best hit, une séquence B dans le protéome de l'espèce Y, celle-ci est considérée comme meilleur hit réciproque (BRH : Best Reciprocal Hit, voir aussi reciprocal best hit ou bidirectional best hit) si la séquence A est le "best hit" dans le protéome de l'espèce X en partant de la séquence B. Cette condition de BRH est largement employée pour définir les relations d'orthologie [Wall et al. 2003 ; Wolf et Koonin 2012]. Elle est notamment implémentée dans la plupart des programmes de recherche d'orthologues [Li et al. 2003 ; Ebersberger et al. 2009 ; Emms et Kelly 2015]. Cependant, elle n'est généralement pas un critère suffisant à cause des limitations des méthodes de recherche de similarité [Koski et Golding 2001] et tend à empêcher la détection des orthologues chez les organismes présentant un haut taux de duplication de gène [Dalquen et Dessimoz 2013].

Forty-Two est un programme ayant pour premier objectif l'identification et l'ajout de séquences orthologues à des alignements multiples de séquences (MSA : multiple sequence alignment) existants. Pour réaliser cet objectif, il utilise une heuristique dénommée BRH multiple [Figure 2.4]. L'utilisateur sélectionne d'une part une liste d'organismes présents dans ses MSAs et d'autre part un ensemble de protéomes qui serviront de référence pour valider l'orthologie. Les séquences des organismes de la première liste (i.e., "queries") sont utilisées pour rechercher avec BLAST les homologues correspondants dans deux ensembles : (i) la base de données de recherche pouvant être constituée de génomes, protéomes ou transcriptomes et (ii) l'ensemble des protéomes de référence. Il s'ensuit une vérification de l'orthologie entre les meilleurs hits obtenus dans les différents protéomes à partir des "queries" sur le principe du BRH. Cette étape permet de consolider une liste de séquences vérifiant au mieux la relation d'orthologie entre les organismes de référence. D'un autre côté, les séquences homologues récupérées au point (i) sont également comparées aux protéomes de référence. Si les meilleurs hits obtenus à partir d'un homologue font partie de la liste consolidée de séquences, ce dernier est alors considéré comme orthologue. Il est à noter que la définition de l'orthologie n'est donc pas limitée par la "query" de départ, une séquence définie comme homologue par une "query" pouvant être validée orthologue par des séquences de référence déterminées par une autre "query".

En plus de cette heuristique de vérification de l'orthologie, Forty-Two implémente aussi une vérification taxonomique des séquences ajoutées visant à identifier ou éviter les contaminations (Figure 2.4). Suite à la détermination des séquences orthologues, chacune de ces dernières est comparée aux séquences de l'alignement pour déterminer la séquence

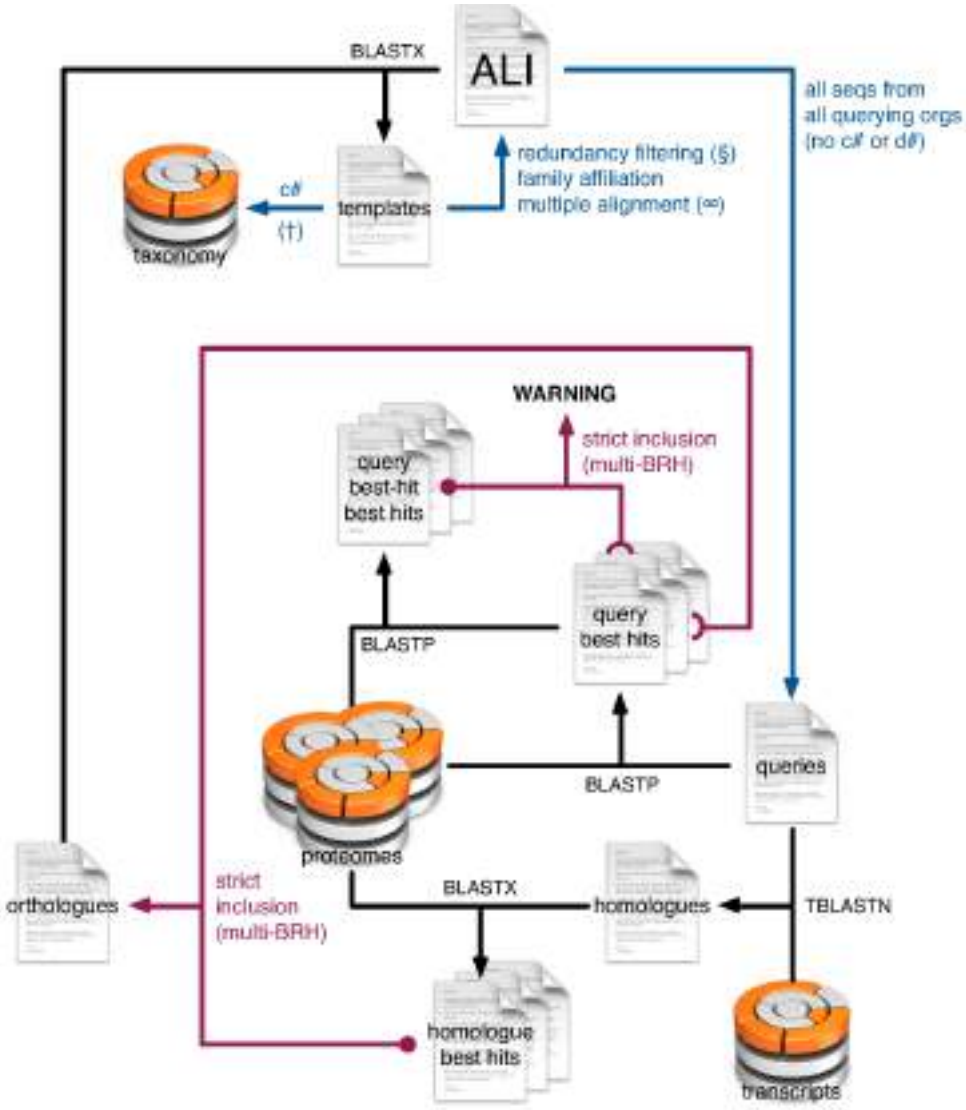


FIGURE 2.4 – Schéma du fonctionnement de Forty-Two

lui étant la plus similaire. Cette dernière sert de modèle pour réaliser l'alignement de la nouvelle séquence et permettre son insertion au MSA, mais également à lui affilier une taxonomie. Cette taxonomie est récupérée automatiquement à partir du nom de l'organisme de l'alignement de départ et correspond à celle implémentée dans la base de données du NCBI. Pour chaque nouvel organisme à ajouter, un filtre taxonomique peut être appliqué à ses séquences orthologues. La taxonomie affiliée est alors comparée à ce filtre pouvant être inclusif (la séquence doit appartenir au clade mentionné), exclusif (ne peut pas appartenir au clade mentionné) ou les deux. Cette dernière étape permet le retrait ou le marquage des séquences ne respectant pas le filtre taxonomique demandé par l'utilisateur. Elle vise à éviter des contaminations d'origines connues ou potentiellement inconnues ainsi qu'à limiter l'ajout de séquences transférées récemment.

Mes résultats précédents ont montré que la définition de séquences orthologues par le

BRH multiple de Forty-Two était compatible avec celle des orthogroupes inférés par OrthoFinder. Dans la suite de ce travail, je vais utiliser Forty-Two pour ajouter des séquences orthologues d'organismes eucaryotes aux graines définies avec OrthoFinder. Cependant, avant de réaliser cette étape, je vais également m'en servir pour deux autres points. Premièrement, il va me permettre de vérifier les divisions opérées sur les OGs présentant des signes de paralogie. Deuxièmement, je vais utiliser son filtre taxonomique pour déterminer la qualité (i.e., le taux de contamination) des transcriptomes utilisés pour améliorer l'échantillonnage de mes alignements.

2.5 Robustesse de la séparation de groupes de séquences paralogues

La définition d'un orthogroupe fournie par les auteurs d'OrthoFinder stipule que celui-ci contient à la fois les orthologues et les paralogues [Emms et Kelly 2015] et sous-entend que les graines obtenues pourraient contenir des séquences paralogues anciennes (out-paralogs). Si plusieurs groupes de séquences paralogues sont facilement distinguables au sein d'un OG, ces derniers peuvent être séparés l'un de l'autre afin de former autant de groupes orthologues. Ces derniers peuvent alors être à nouveau considérés comme utilisables en tant que graines. Cette dernière affirmation n'est vraie que si la séparation opérée reste robuste à l'ajout subséquent de nouvelles séquences. Pour vérifier ce dernier point, j'ai mis au point un protocole permettant de vérifier la robustesse de la découpe automatisée d'arbre de gène dans le but de séparer des groupes de séquences paralogues (Figure 5). Après chaque découpe d'un arbre de gène (voir [Simion et al. 2017]), je récupère les alignements correspondant aux groupes de séquences des deux sous-arbres définis par la découpe. Ensuite, j'utilise Forty-Two pour leur ajouter des séquences à partir des protéomes utilisés pour définir ces alignements. Finalement, je vérifie si les deux alignements obtenus contiennent des séquences identiques (i.e. ajoutées indépendamment dans les deux alignements). Si c'est le cas, je considère la découpe comme non-fiable et élimine les deux alignements. Ce protocole est répété jusqu'à ne plus trouver de point de coupe dans les OGs résultants.

J'ai testé ce protocole dans le cadre de la création du jeu de données phylogénomique des eucaryotes en l'appliquant sur les OGs montrant des signes de multiples cas de paralogie. Pour identifier ces derniers, j'ai inféré l'arbre de gène pour tous les OGs contenant a priori suffisamment de signal phylogénétique. Les séquences furent alignées avec MAFFT (algorithme L-INS-I, 5000 itérations, [Katoh et al. 2002 ; Katoh et Standley 2013]) avant d'en retirer les positions avec plus de 50% de gaps. Les alignements possédant moins de 50 positions informatives pour la parcimonie furent éliminés, car ne contenant

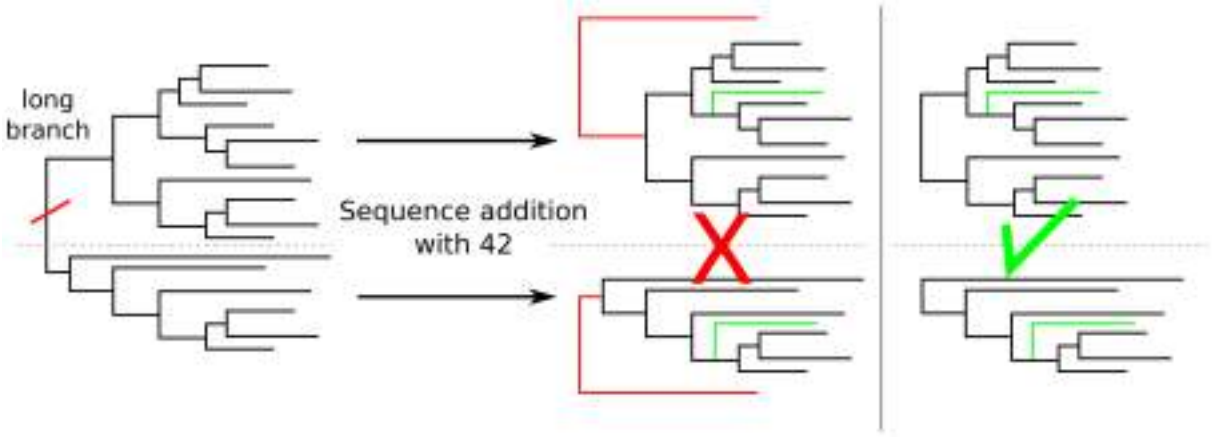


FIGURE 2.5 – Protocole de validation des coupes d'OG paralogue

probablement pas assez d'informations. Les alignements restants furent considérées pour réaliser l'inférence avec RAxML [STAMATAKIS 2014] et le modèle LG+F+Γ4. Ensuite, pour quantifier le taux de paralogie dans les arbres obtenus, j'ai défini 18 clades, dont 16 eucaryotes, pour lesquels la monophylie devait être aisément retrouvable car séparés par des branches relativement longues. Comme dans SIMION et al. (2017), pour chaque clade, j'ai quantifié le nombre de clans ne contenant que des espèces de ce clade et j'ai considéré un OG comme montrant de grands signes de paralogie lorsque plus de 70% des 18 cladea présentaient plus d'un clan (i.e. moins de 30% des clades formaient un groupe monophylétique unique). Le nombre d'OGs répondant à ce critère s'éleva à 761 qui furent considérés pour une découpe automatisée.

Pour appliquer le protocole, il faut des critères pour couper l'arbre. J'ai choisi de n'en considérer qu'un seul mais d'étudier l'impact des séquences divergentes et/ou partielles sur la validation de cette condition. Ce dernier fut de couper l'arbre si la branche séparant l'échantillonnage taxonomique en les deux ensembles maximisant la diversité, était parmi les 30% de branches les plus longues, et que les sous-arbres correspondants contenaient au minimum cinq OTUs. Concernant le retrait des séquences problématiques, celui-ci fut réalisé en considérant le pHMM (profile HMM), obtenu avec HMMER, de chaque alignement. Une séquence était retirée si elle ne s'alignait pas sur un minimum de 65% du pHMM. Le critère de découpe fut testé sur les arbres obtenus avant et après retrait de séquences et valida au final 220 et 264 sous-arbres, respectivement. Ce résultat valide donc que la découpe est facilitée par le retrait de séquences supposées problématiques. J'ai donc appliqué le protocole sur les arbres obtenus de cette manière.

Après chaque découpe, de nouvelles séquences furent possiblement ajoutées par Forty-Two lors de la réalisation du protocole (i.e., vérification de l'ajout de séquences à partir des mêmes données de départ). Ces dernières furent considérées pour valider les découpes mais pas conservées dans les alignements pour la réalisation des découpes subséquentes. Après

une première validation, les arbres ne pouvant plus être coupés par la condition de coupe furent mis de côté. Après cinq itérations de l'ensemble du protocole, 201 arbres avaient été mis de côté et 89 étaient encore considérés comme découpables. Ces derniers furent éliminés et la représentation taxonomique des autres fut analysée. Au final, j'ai conservé 159 alignements contenant au minimum 10 eucaryotes. Le protocole se révéla donc très strict en n'épargnant qu'un nombre restreint d'alignements comparé à la quantité d'OGs de départ. Vu le temps de calcul nécessaire à l'application de ce protocole, il est probablement plus raisonnable d'éviter dans un premier temps les alignements présentant des signes de paralogie plutôt que d'essayer d'en extraire les groupes orthologues individuels.

2.6 Détermination de la qualité des transcriptomes utilisés

L'utilisation de données moléculaires de haute qualité est une nécessité majeure lors de la réalisation d'études en évolution moléculaire. Un des critères affectant cette qualité est l'acquisition de séquences n'appartenant pas à l'organisme étudié. Ces contaminations, dotées d'une affiliation taxonomique erronée, sont une possible source d'erreurs importantes dans les études phylogénétiques [Philippe et al. 2011b]. Elles peuvent apparaître par exemple lors de la manipulation des échantillons avant séquençage [Merchant et al. 2014] ou à cause d'une purification non suffisante des échantillons [Simion et al. 2017]. Le séquençage en parallèle de plusieurs échantillons peut notamment aboutir à leur cross-contamination [Ballenghien et al. 2017 ; Simion et al. 2018]. En l'absence de traitements adéquats à leur encontre, les contaminations se retrouvent alors dans les bases de données biologiques [Longo et al. 2011 ; Cornet et al. 2018].

Afin d'améliorer l'échantillonnage taxonomique de mes alignements de séquences, j'ai récupéré de multiples données transcriptomiques et assemblé des données de séquençage en provenance du NCBI. Parmi les données récupérées, une majorité a été fournie par le projet de séquençage de transcriptomes d'eucaryotes microscopiques marins (MMETSP : Marine Microbial Eukaryote Transcriptome Sequencing Project) [Keeling et al. 2014]. L'objectif de ce projet était de rendre disponible publiquement plus de 650 transcriptomes assemblés et annotés pour une large diversité d'eucaryotes (Figure 2.6). Ces données ont été principalement obtenues à partir de souches d'organismes provenant de cultures maintenues en laboratoire. Comme il est difficile de maintenir des cultures axéniques et que les différents laboratoires ont séquencé leurs échantillons en parallèle, il est a priori possible que ces transcriptomes contiennent des contaminations.

Afin d'évaluer la qualité des transcriptomes récoltés, nous avons conçu un protocole nous

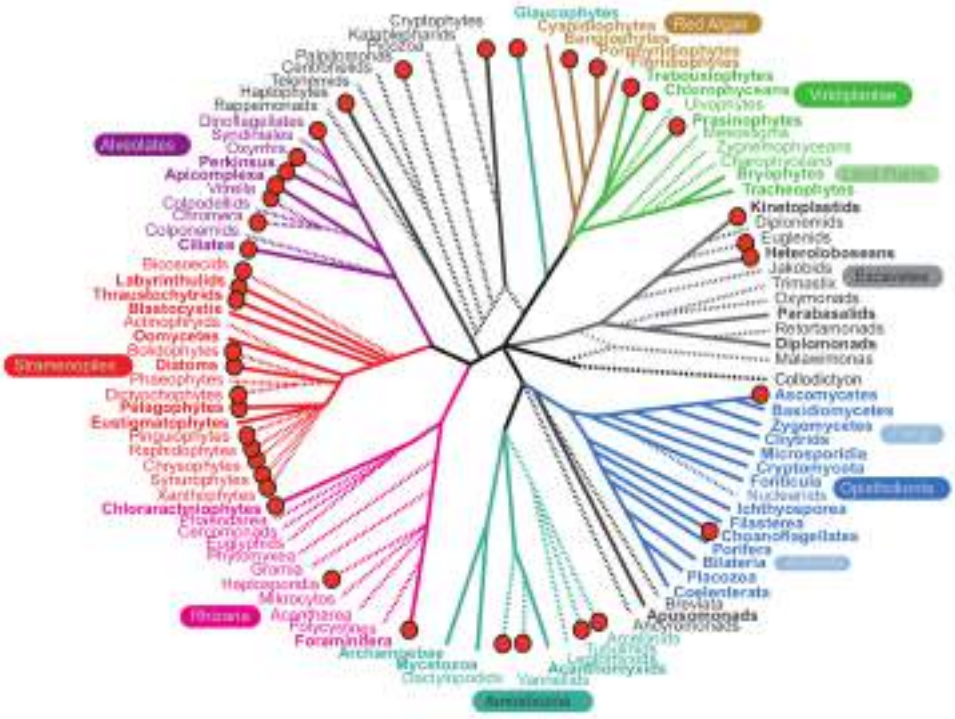


FIGURE 2.6 – Représentation de la diversité des eucaryotes. Les branches de l'arbre pour lesquels des échantillons ont été séquencés par le MMETSP sont marqués d'un cercle

permettant d'estimer la pureté des transcriptomes et d'identifier les possibles sources de contamination. Ce protocole se focalise autour d'un jeu de données de 80 protéines ribosomiques dérivé de celui de BAPTESTE et al. (2002), mis à jour en 2011 [PHILIPPE et al. 2011a], et de l'utilisation du filtre taxonomique de Forty-Two. Les protéines ribosomiques sont parmi les protéines les plus exprimées par les cellules, malgré l'existence de variations possibles entre tissus [GUIMARAES et ZAVOLAN 2016]. Une contamination de l'échantillon séquencé a donc une haute probabilité d'être observée lors de l'examination des contigs de ces protéines. Ainsi, les nouveaux transcriptomes sont analysés avec Forty-Two afin d'en extraire les orthologues qui correspondent aux 80 gènes. Sur base de l'échantillonnage déjà disponible (actuellement plus de 7000 souches eucaryotes), la (ou les) bonne(s) séquence(s) pour chaque gène est déterminée manuellement pour tous les transcriptomes, avec l'aide des mesures de similarité entre séquences et en supposant une forte similarité avec les organismes proches et une faible similarité avec les organismes éloignés, avec une validation par une analyse phylogénétique dans les cas ambigus (Figure 2.7A). Suite à la mise à jour de la base de données de protéines ribosomiques, les transcriptomes sont traités une seconde fois avec Forty-Two, mais en contraignant les séquences orthologues ajoutées à correspondre à une séquence d'un organisme du même genre. Si elles ne vérifient pas ce critère taxonomique, les séquences sont étiquetées comme contamination et affiliées à l'organisme source le plus vraisemblable, fondé sur l'échantillonnage taxonomique disponible dans l'alignement (Figure 2.7B). Finalement,

nous collectons les affiliations de chaque séquence (bonne ou mauvaise) afin de classer les transcriptomes selon leur couverture (nombre de gènes pour lesquels une bonne séquence fut trouvée) et leur taux de contamination (proportion de séquences étiquetées comme contamination) (Figure 2.7C).

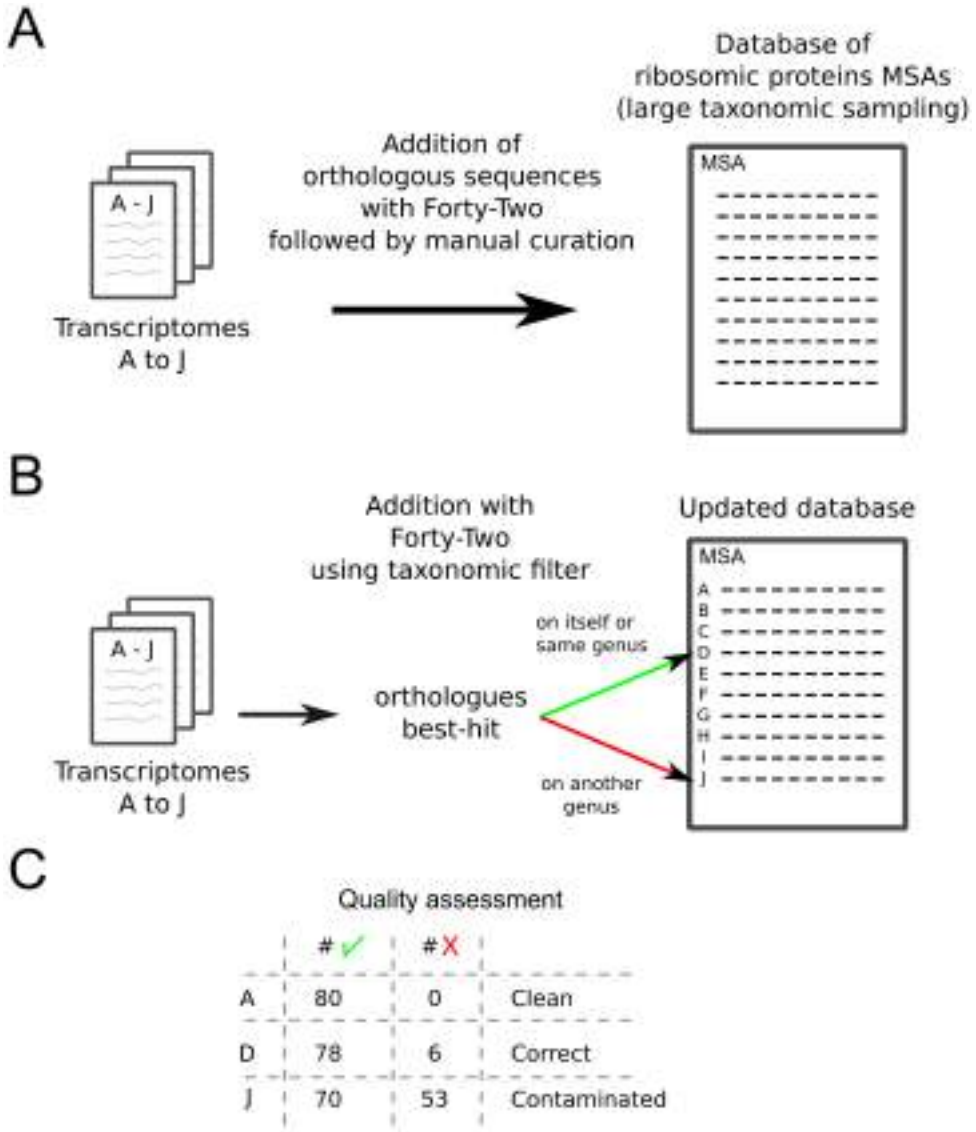


FIGURE 2.7 – Protocole d'estimation de la qualité des données transcriptomiques. A. Ajout des transcriptomes dans la base de données suivi d'une sélection manuelle des bonnes séquences. B. Ajout automatique des transcriptomes en appliquant un filtre taxonomique avec Forty-Two. C. Quantification de séquences correctes et contaminées

J'ai appliqué cette méthodologie sur 395 transcriptomes produits par le MMETSP. Sur un total de 67.087 séquences ajoutées, 15.257 furent étiquetées comme possibles contaminations. Ces dernières affectaient 325 des transcriptomes analysés dont 114 par plus de 10 séquences. Les transcriptomes les plus touchés appartenaient principalement aux dinoflagellés (e. g. 1231 pour *Symbiodinium*, 972 pour *Alexandrium*) pour lesquels le nombre de contaminations était surestimé dû à l'existence de nombreux paralogues récents. En effet,

ces paralogues divergents avaient peu de chance de respecter le filtre taxonomique, le best hit tombant par hasard sur un autre dinoflagellé proche. Cependant, la souche de *Symbiodinium* concernée (D1a) présentait également de fortes contaminations aux diatomées (89 séquences affiliées au genre *Nitzschia*) et aux Discosea (82 séquences affiliées au genre *Vannella*). Du point de vue des contaminations, les Discosea représentaient une source importante de contaminations (547 contaminations affiliées au genre *Vannella*), tout comme les Haptophyceae (374 contaminations affiliées au genre *Isochrysis*) ou les diatomées (329 séquences contaminées affiliées au genre *Nitzschia*). Outre le cas des dinoflagellés, la contamination la plus marquée était celle d'un foraminifère du genre *Sorites* aux dinoflagellés du genre *Symbiodinium* (175 contaminations). La considération de ces différents résultats ainsi que la redondance de certains genres dans les données du MMETSP m'a fait sélectionner un sous-ensemble de 254 transcriptomes (sur base de leur complétude, propreté, position supposée dans l'arbre des Eucaryotes) dans le but de réaliser un premier enrichissement propre de mes alignements.

Cependant, cette première sélection a pu passer à côté de l'ajout de taxons clés, appartenant à des clades peu échantillonnés ou permettant de réduire la longueur de certaines branches internes. Pour ajouter des données contaminées de ce type de taxons, un échantillonnage riche est nécessaire afin de profiter au mieux du filtre taxonomique de Forty-Two. Par exemple, *Gloeochaete wittrockiana* est une espèce importante pour deux raisons. Premièrement, elle appartient à la classe des Glaucophytes, un clade photosynthétique important représenté par des données moléculaires pour uniquement quatre espèces. Deuxièmement, elle est la source de contamination supposée de 16 autres transcriptomes. Par conséquent, l'ajout de ce transcriptome faciliterait à la fois l'inférence de la phylogénie à partir de la super-matrice finale et l'ajout d'autres transcriptomes contaminés lors de la construction de la super-matrice. Le premier ajout de séquences ayant enrichi l'échantillonnage des alignements en glaucophytes, j'ai pu dans un premier temps ajouter *Gloeochaete* en forçant les séquences à ressembler aux autre glaucophytes (usage d'un filtre taxonomique inclusif). Ensuite, j'ai pu utiliser la présence de ce dernier pour éviter les contaminations lui étant dues (usage d'un filtre taxonomique exclusif). En procédant par des ajouts itératifs, couplés au retrait des alignements pauvres en taxa (voir plus bas), j'ai augmenté l'échantillonnage taxonomique du jeu de données jusqu'à 500 OTUs. Celui-ci fut finalement réduit à 370 OTUs en enlevant les espèces peu représentées ou en créant des chimères entre espèces proches.

2.7 Retrait de séquences et consolidation des alignements

Après les premiers ajouts de séquences, j'ai réalisé plusieurs types de vérifications visant à assurer le bon échantillonnage des alignements et la qualité des séquences ajoutées (absence de paralogues ou xénologues). Outre la meilleure détection des cas de non orthologie, ces étapes avaient également pour but de réduire la quantité de données manquantes dans la matrice finale, une grande proportion de ces dernières pouvant affecter l'inférence phylogénétique [Roure et al. 2013]. La première étape fut de retirer les taxa procaryotes ainsi que les alignements ne contenant pas un échantillonnage suffisamment riche. Dans ce but, les alignements devaient valider 10 des 13 conditions de nombre de taxa minimum : 3 Amoebozoa, 2 Holomycota, 3 Holozoa, 2 Excavata, 10 Viridiplantae, 5 Rhodophyta, 5 Haptophyta, 5 Cryptista, 5 Rhizaria, 3 Apicomplexa, 3 Dinophyceae, 4 Ciliophora et 20 Stramenopiles. Ce filtre réduisit drastiquement le nombre d'alignements considérés (881), réduisant les temps de calcul nécessaires aux prochaines vérifications.

Pour savoir si des contaminations ont pu passer à travers mes précédents filtres, j'ai continué de vérifier si elles étaient présentes dans les alignements. Pour ce faire, j'ai réalisé une étape de BLAST pour laquelle la taxonomie des séquences fut prise en compte, avec pour objectif premier le retrait de possibles contaminations résiduelles. Pour chaque alignement, les séquences appartenant à un clade, parmi une série de clades prédéfinis, furent comparées à l'ensemble des séquences de l'alignement. L'analyse des moyennes des scores BLAST obtenus envers chaque clade prédéfini pour une séquence, permit de déterminer les séquences susceptibles d'être des contaminations, ces dernières possédant des scores moyens supérieurs pour des clades autres que le leur. Outre les contaminations, cette méthode permit également d'identifier les séquences de nucléomorphe. En effet, les cryptophytes et les chlorarachniophytes possèdent un noyau au génome réduit, nommé nucléomorphe, entre les troisième et quatrième membranes de leur chloroplaste. Les séquences de ces nucléomorphes sont très divergentes [Curtis et al. 2012] et ne vont donc pas présenter de fortes similarités à un clade particulier. Par conséquent, toute séquence de cryptophyte ou de chlorarachniophyte présentant un meilleur score moyen à un autre clade, même si cet organisme ne présente pas de signe de contaminations, a de grandes chances d'être une séquence de nucléomorphe.

J'ai appliqué une dernière méthode pour éliminer à la fois les mauvaises séquences (paralogues, xénologues) et les mauvais alignements (ceux pour lesquels le respect de l'orthologie était ambigu). Cette dernière consiste à comparer les longueurs de branches entre les arbres de gène et l'arbre des espèces. Elle nécessite l'inférence d'un arbre des espèces intermédiaire qui servira de référence. Ce dernier s'obtient en réalisant l'inférence

sur une super matrice obtenue en concaténant les différents alignements du jeu de données (Figure 2.8A). Ensuite, on infère la phylogénie de chaque alignement utilisé pour la concaténation (une séquence par OTU) en fixant la topologie de l'arbre à celle de la super matrice et en utilisant le même modèle (Figure 2.8B). Finalement, on compare les branches de l'arbre de gène à celui de l'arbre d'espèces réduit au même échantillonnage (Figure 2.8C). À topologie fixe, si une séquence est forcée à une position incorrecte de l'arbre (i.e. la position taxonomique attendue, mais qui n'est pas la bonne à cause de paralogie ou de xénologie), le modèle d'évolution des séquences ne peut expliquer sa position que par un nombre important d'événements de substitution. La taille de la branche de la séquence en question est donc augmentée de manière disproportionnée par le modèle. Par conséquent, on peut déterminer ce type de séquence en comparant la taille de sa branche à celle attendue par l'arbre d'espèce de référence. De plus, si on regarde toutes les longueurs de branche simultanément, la présence de plusieurs tailles de branche non-attendues dans l'arbre de gène affecte la valeur de corrélation des longueurs de branches (R^2) entre la référence et le gène. Une faible valeur de R^2 permet donc également de cibler des alignements possiblement riches en paralogues ou présentant un signal phylogénétique opposé à celui de la majorité des gènes (xénologies ou grave incompatibilité de la phylogénie du gène avec la phylogénie des espèces).

Ici, j'ai réalisé la concaténation sur des alignements filtrés par HmmCleaner et BMGE [Criscuolo et Gribaldo 2010] en utilisant SCaFoS [Roure et al. 2007] qui sélectionne par défaut la séquence la plus longue par OTU et j'ai inféré les arbres en utilisant le modèle LG+F+Γ4 avec RAxML. Ensuite, j'ai identifié et retiré les branches terminales cinq fois plus longues dans l'arbre de gène par rapport à l'arbre de référence. J'ai accompli ce procédé deux fois, ce qui retira d'abord 2520 séquences puis 775 séquences. Finalement, j'ai comparé une dernière fois les arbres de gènes à l'arbre de référence pour sélectionner les alignements présentant les plus mauvaises corrélations de longueur de branches. Ce dernier point réduisit le nombre d'alignements à 611 en enlevant les alignements possédant une valeur de R^2 inférieure à deux fois l'écart-type par rapport à la moyenne (supposant une distribution normale des coefficients de corrélation).

Suite à ces différentes évaluations de mes alignements, j'ai finalisé la construction de la matrice en réalignant les séquences avant de masquer les segments de séquence non-homologues avec HmmCleaner et de retirer les positions de faible entropie avec BMGE. J'ai alors conservé les alignements possédant plus de 200 OTUs avant de les concaténer avec SCaFoS. J'obtins ainsi un jeu de données final de 594 gènes et 370 OTUs comprenant 81.708 positions dont 31,46% de données manquantes. C'est à partir de ce dernier que j'ai évalué l'impact de l'échantillonnage taxonomique et du modèle d'évolution sur l'inférence de l'arbre des eucaryotes.

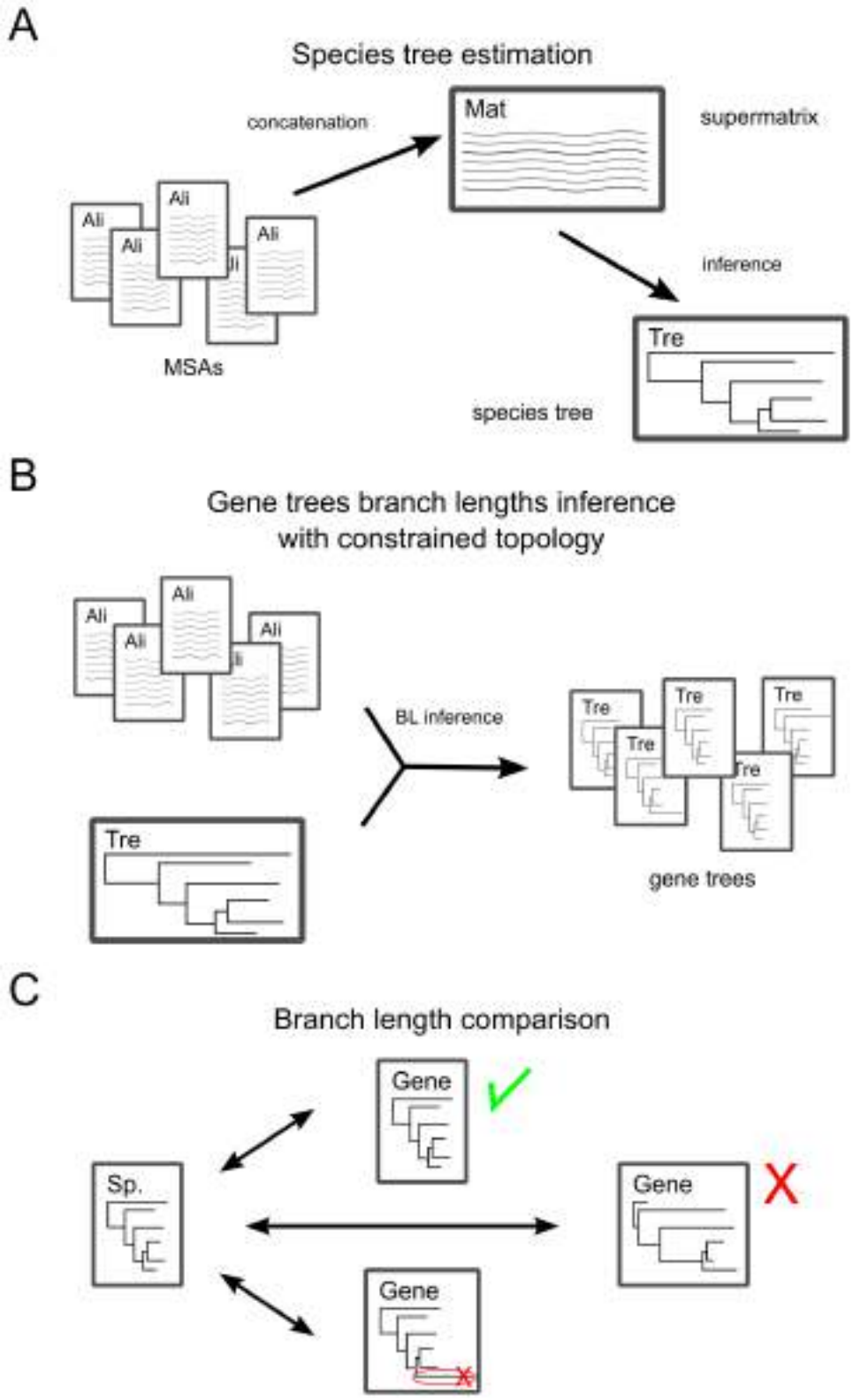


FIGURE 2.8 – Protocole de décontamination par longueur de branches. A. Création d'une référence. B. Estimation des longueurs de branches de l'arbre de gène en forçant la topologie de référence. C. Comparaison de la longueur des branches entre l'arbre de gène et l'arbre d'espèces

2.8 Inférence de la phylogénie des eucaryotes

Ce jeu de données constitue le plus grand jeu de données eucaryotes à ce jour dans le but de résoudre les noeuds incertains de la phylogénie des eucaryotes (ici, principalement ceux à l'intérieur des Diaphoretickes). En effet, il comprend un nombre de positions jusqu'à deux fois plus important que dans les précédentes études ([Burki et al. 2012] : 55.881 positions, [Brown et al. 2013] : 43.615 positions, [Katz et Grant 2014] : 36.346 positions, [Cavalier-Smith et al. 2015a] : 51.352) et est le plus riche en taxa (maximum 232 [Katz 2015]). Ces deux points m'ont permis d'évaluer l'impact du nombre d'espèces considérées sur l'inférence phylogénétique. En plus de la matrice complète, j'ai également analysé un jeu de données de 125 espèces et un autre de 63 espèces. Ces espèces ont été sélectionnées pour se rapprocher des échantillonnages utilisés par Brown et al. (2013) et Burki et al. (2016). Comme Katz et Grant (2014), j'ai pu appliquer des méthodes de tirage sans remise pour obtenir les supports statistiques. Ici, je n'ai recouru qu'au jackknife de gènes [Delsuc et al. 2008] car je le considère plus intéressant que le jackknife de sites. Comme il force l'analyse de toutes les positions des gènes échantillonnés, il permet de mieux juger de la robustesse de l'inférence par rapport aux biais qui pourraient être induits dans certains gènes [Simion et al. 2017].

Mon analyse principale a consisté en l'analyse de la matrice complète (370 espèces) avec le modèle CAT [Lartillot et Philippe 2004]. J'ai réalisé 50 tirages de jackknife de gènes, chacun comprenant environ 30.000 positions, pour chacun desquels une chaîne MCMC a tourné sous PhyloBayes MPI [Lartillot et al. 2013] jusqu'à atteindre 2000 cycles. L'arbre final correspond au consensus des 50 consensus obtenus à partir des 500 derniers arbres de chaque chaîne. L'arbre final est en accord avec les topologies observées dans la littérature (Figure 2.9). On observe notamment la division des eucaryotes en trois grands groupes, les Amorphea (JS=92%), les Diaphoretickes (JS=100%), et les Excavata (JS=100%) [Burki et al. 2008 ; Hampl et al. 2009]. Au sein des Amorphea, les Amoebozoa et les Opisthokhonta (Holozoa et Holomycota) sont retrouvés monophylétiques (support de jackknife, JS=100%). Ce dernier clade est groupe-frère (JS=92%) du regroupement de *Pygsuia biforma* (Breviatea) et *Thecamonas trahens* (Apusomonadida) (JS=74%), un groupe connu sous le nom de Obazoa [Brown et al. 2013]. En supposant la racine à la base des Diaphoretickes+Excavata, *Nutomonas longa* se retrouve groupe-frère des Amorphea (JS=86%) comme proposé récemment par Brown et al. (2018). Les Excavata forment un groupe monophylétique représenté uniquement par des Heterolobosea et des Euglenozoa, clades évoluant les plus lentement [Rodríguez-Ezpeleta et al. 2007b].

De l'autre côté de l'arbre, les différents clades composant les Diaphoretickes sont tous retrouvés monophylétiques (Rhodophyta, Viridiplantae, Glaucophyta, Haptophyta, Cryptista, Stramenopiles, Alveolata et Rhizaria). Ces derniers sont séparés en deux groupes

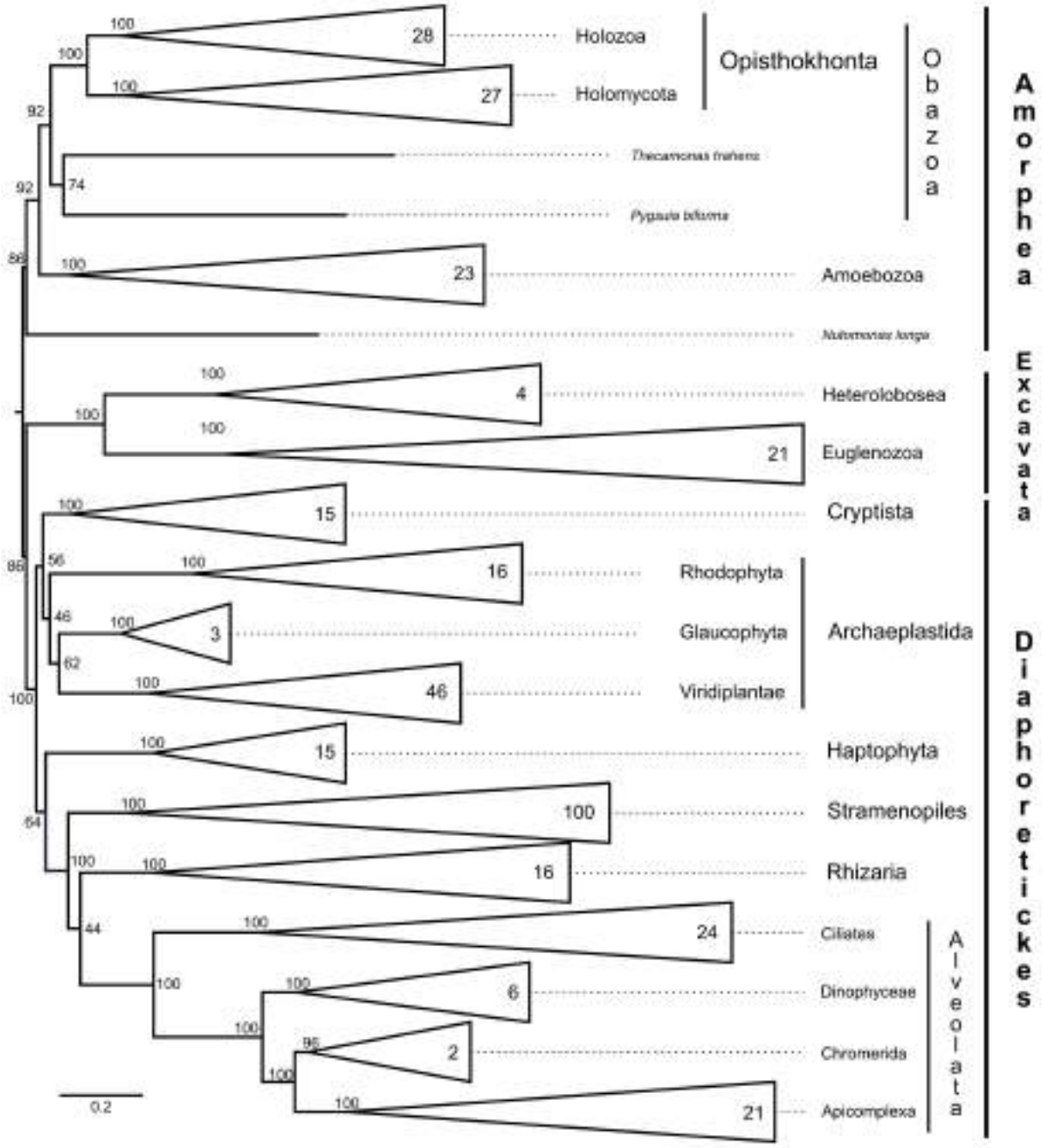


FIGURE 2.9 – Arbre phylogénétique des eucaryotes obtenu sur base de 370 espèces et 50 réplicats de jackknife de gènes de 30000 avec le modèle CAT.

comme obtenu par BURKI et al. (2012). D'une part, on retrouve les Cryptista associés aux Archaeplastida (JS=56%) et d'autre part, les Haptophyta regroupés avec l'ensemble monophylétique des SAR (JS=64%). L'absence de résolution au sein de Diaphoretickes témoigne de la problématique posée par ARCHIBALD (2009). La monophylie des Archaeplastida est retrouvée mais avec un faible support (JS=46%), la topologie retrouvée favorisant la monophylie des Viridiplantae et des Glaucophyta (JS=62%). Du côté des SAR, bien que l'ensemble soit indiscutablement monophylétique, le support pour regrouper Al-

veolata et Rhizaria est faible (JS=44%). Ce regroupement est en accord avec l'hypothèse de monophylie de ce groupe soutenue par He et al. (2016) mais va à l'encontre des résultats obtenus dans les dernières études phylogénétiques favorisant plutôt la monophylie entre Alveolata et Stramenopiles [Brown et al. 2013 ; Burki 2016 ; Brown et al. 2018 ; Heiss et al. 2018 ; Lax et al. 2018].

Suite à ces observations, j'ai testé si la topologie inférée avec le jeu de données complet était également retrouvée en considérant moins d'espèces. Mon attente était que le modèle CAT se montrerait robuste à l'échantillonnage taxonomique mais que les supports statistiques pourraient être affectés. Dans l'ensemble, ces attentes furent respectées (Table 2.1) car les trois topologies sont identiques à l'exception de l'ordre de spéciation à l'intérieur des SAR. Celui-ci est différent pour les trois échantillonnages mais reste faiblement supporté (44%, 40% et 60% pour 370, 125 et 63 espèces respectivement). Cependant, le jeu de données à 63 espèces montre un support plus élevé que les deux autres en faveur de la monophylie entre Alveolata et Stramenopiles. Comme ce nombre d'espèces est semblable à celui des études en faveur de cette topologie, l'utilisation d'un faible échantillonnage pourrait en être à l'origine. A l'intérieur des Archaeplastida, la topologie reste inchangée bien que le plus faible échantillonnage présente le plus faible support pour leur monophylie. Finalement, le support pour regrouper les Haptophyta avec les SAR semble augmenter légèrement avec la diminution de l'échantillonnage taxonomique.

Noeuds d'intérêts	CAT jack 50, 370sp ∼30.000 sites	CAT jack 50, 125sp ∼30.000 sites	CAT jack 50, 63sp ∼30.000 sites
Amorphea	92	98	100
Nutomonas	86	92	98
Diaphoretickes	100	100	98
Excavata	Amorphea-Diaphoretickes	Amorphea-Diaphoretickes	Amorphea-Diaphoretickes
Archaeplastida	((G,V)62,R)46	((G,V)64,R)57	((G,V)62,R)32
SAR	((R,A)44,S)100	((R,S)40,A)100	((A,S)60,R)100
Cryptista	56 avec Archaeplastida	54 avec Archaeplastida	64 avec Archaeplastida
Haptophyta	64 avec SAR	72 avec SAR	80 avec SAR

Les arbres correspondent aux consensus obtenus à partir de 50 réplicas de jackknife de gènes avec le modèle CAT. Pour Archaeplastida, G=Glaucophyta, V=Viridiplantae et R= Rhodophyta. Pour SAR, S=Stramenopiles, A=Alveolata et R=Rhizaria

Table 2.1 – Noeuds principaux et supports obtenus lors de variation de l'échantillonnage taxonomique (370,124 et 63 espèces).

Comme l'utilisation d'un plus grand échantillonnage taxonomique permet d'améliorer l'inférence de l'arbre, ce premier devrait affecter d'autres paramètres de l'arbre si sa topologie est peu affectée, notamment la taille de ses branches. Afin de pouvoir comparer la taille de l'arbre (i.e. la somme de toutes ses branches) à 370 espèces avec les arbres à 125 et 63 espèces, je lui ai retiré des branches afin de rendre son échantillonnage identique à ceux des deux autres arbres. La taille fut toujours plus grande pour l'arbre à 370 espèces réduit que pour les deux arbres inférés avec un échantillonnage plus faible (49,80 vs 43,92

et 33,07 vs 28,95 pour 125 et 63 espèces respectivement). Ces résultats illustrent bien le fait que le nombre de substitutions multiples est mieux inféré en présence de nombreux taxa. Le modèle CAT est donc robuste à l'échantillonnage en terme de topologie mais reste affecté par l'erreur systématique quand il s'agit d'inférer la taille des branches. Cependant, on observe que la différence de taille n'est pas proportionnelle à la différence du nombre d'espèces utilisées, suggérant un impact important du choix des taxa dans l'inférence des longueurs de branches.

Si la topologie n'est pas impactée avec CAT, est-ce également le cas avec d'autres modèles ? Pour répondre à cette question, j'ai également inféré les mêmes trois arbres en employant le modèle LG4X [Le et al. 2012] avec RAxML en réalisant 100 tirages de fast-bootstrap (Table 2.2). Au contraire de CAT, les topologies inférées furent relativement différentes à l'exception des relations à l'intérieur des Amorphea. Le positionnement des Excavata fut différent pour chaque arbre, ceux-ci se retrouvant groupe-frère des SAR, des Rhodophyta ou de Nutomonas. Les Cryptista et Haptophyta se positionnèrent de manière très variable entre l'intérieur des Archaeplastida ou groupes-frère des SAR. Les Archaeplastida ne furent jamais retrouvés monophylétiques. Les importantes incongruences topologiques suggèrent que les résultats obtenus avec LG4X sont impactés par l'erreur systématique. Les positions observées pour les Excavata font notamment penser à des attractions entre longues branches.

Noeuds d'intérêts	LG4X boot100 370sp	LG4X boot100 124sp	LG4X boot100 63sp
Amorphea	99	100	100
Nutomonas	78 avec Amorphea	58 avec Amorphea	91 avec Excavata
Diaphoretickes	Non retrouvé	Non retrouvé	96
Excavata	55 avec SAR	48 avec Rhodophyta	Amorphea-Diaphoretickes
Archaeplastida	Non retrouvé Inclus H et C (60)	Non retrouvé	Non retrouvé Inclus C (90)
SAR	((R,A)78,S)100	((R,A)85,S)100	((R,A)70,S)100
Cryptista	((G,V)98,C)43	(C,H)59	(G,C) 58
Haptophyta	(R,H)52	(C,H)59	76 SAR

Les arbres correspondent aux meilleurs arbres inférés sous RAxML avec le modèle LG4X et 100 réplicas de bootstrap. Pour Archaeplastida, G=Glaucophyta, V=Viridiplantae et R= Rhodophyta. Pour SAR, S=Stramenopiles, A=Alveolata et R=Rhizaria.

Table 2.2 – Noeuds principaux et supports obtenus lors de variation de l'échantillonnage taxonomique (370,124 et 63 espèces).

Les résultats obtenus avec le modèle CAT présentaient des supports relativement faibles à l'intérieur des Diaphoretickes. Ces derniers pourraient être dus à un signal phylogénétique trop faible lié à la non-utilisation de l'ensemble des positions du jeu de données. Il n'est pas possible de réaliser des tirages de bootstrap avec CAT dans des temps raisonnables en utilisant 370 espèces. Cependant, ceci est envisageable avec un échantillonnage réduit. De même, il devient possible d'utiliser d'autres modèles site-hétérogènes dérivés de CAT

comme C60 [Si Quang et al. 2008] voire de réaliser des bootstraps non-paramétriques grâce à la méthode PMSF [Wang et al. 2018]. Ainsi, j'ai utilisé l'entièreté du jeu de données de 63 espèces (80.000 sites) pour inférer trois arbres : l'arbre obtenu par le modèle LG+C60+Γ4 sous IQ-TREE [Nguyen et al. 2015] avec 1000 ultra fast bootstraps [Hoang et al. 2018], l'arbre obtenu par le modèle LG+C60+Γ4-PMSF sous IQ-TREE avec 50 tirages de bootstrap non-paramétrique et l'arbre obtenu par le modèle CAT+Γ4 sous PhyloBayes MPI avec 50 tirages de bootstrap non-paramétrique. Comme pour les variations d'échantillonnage taxonomique avec CAT, les trois modèles donnèrent des topologies très semblables (Table 2.3). Le seul noeud affecté fut celui positionnant les Cryptista. Les analyses les placèrent groupe-frère des Rhodophyta avec IQ-TREE et des Archaeplastida avec PhyloBayes MPI. Chaque topologie fut en faveur de la monophylie de l'ensemble Alveolata+Stramenopiles, le noeud étant supporté à plus de 90%. Les supports obtenus avec IQ-TREE furent tous très élevés à l'exception de ceux relatifs aux Cryptista. Concernant CAT, ceux-ci furent également faibles pour les noeuds internes aux Archaeplastida.

Noeuds d'intérêts	LG+C60+Γ4 UFBoot1000 ∼80.000 sites	LG+C60+Γ4-PMSF boot50 ∼80.000 sites	CAT boot50 ∼80.000 sites	CAT jack50 ∼30.000 sites
Amorphea	100	100	100	100
Nutomonas	100	100	98	98
Diaphoretickes	100	100	98	98
Excavata	Amorphea-Diaphoretickes	Amorphea-Diaphoretickes	Amorphea-Diaphoretickes	Amorphea-Diaphoretickes
Archaeplastida	Non retrouvé ((G,V)79,(C,R)39)92	Non retrouvé ((G,V)100,(C,R)68)100	((G,V)50,R)66	((G,V)62,R)32
SAR	((A,S)97,R)100	((A,S)100,R)100	((A,S)92,R)100	((A,S)60,R)100
Cryptista	39 avec Rhodophyta	68 avec Rhodophyta	66 avec Archaeplastida	64 avec Archaeplastida
Haptophyta	92 avec SAR	100 avec SAR	100 avec SAR	80 avec SAR

Les deux premiers arbres correspondent aux meilleurs arbres inférés sous IQTREE avec les modèle LG+C60+Γ4 (1000 Fast-Bootstraps) et LG+C60+Γ4-PMSF (50 bootstraps non-paramétriques) et 100 réplicas de bootstrap. Le troisième correspond au consensus de 50 réplicas de bootstrap non-paramétriques avec le modèle CAT. Le dernier est présent pour comparaison (voir table 2.1) Pour Archaeplastida, G=Glaucophyta, V=Viridiplantae, R= Rhodophyta et C=Cryptista. Pour SAR, S=Stramenopiles, A=Alveolata et R=Rhizaria

Table 2.3 – Noeuds principaux et supports obtenus par bootstrap en utilisant l'entiereté du jeu de données 63 espèces avec des modèles site-hétérogènes.

De manière générale, les supports sont supérieurs avec les bootstraps obtenus à partir de 80.000 positions qu'avec les jackknifes de 30.000 positions, comme attendu. Les deux topologies CAT à 63 espèces sont identiques et présentent des améliorations pour l'ordre d'émergence à l'intérieur des SAR (Stramenopiles+Alveolata, JS=60% vs BS=92%) et le positionnement des Haptophyta en tant que groupe-frère des SAR (JS=80% vs BS=100%). Cependant, il est étonnant de voir que l'apport de 50.000 positons ne semble pas affecter les supports pour le positionnement des Cryptista. Cela suggère que le signal existant est très faible ou que ce dernier est affecté par un important signal non-phylogénétique découlant des violations de modèle. Concernant ces dernières, je suis étonné de voir un support si élevé pour regrouper Glaucophyta et Viridiplantae avec la méthode PMSF (BS=100%)

alors que les supports à l'intérieur des Archaeplastida sont faibles dans toutes les autres analyses. Le modèle se présentant comme une heuristique du modèle CAT, il me parait improbable que ce résultat ne soit pas biaisé par le choix de l'arbre de départ LG+C60+Γ4, qui trouve la même topologie avec de plus faibles supports.

2.9 Conclusion

Les différentes méthodes semi-automatisées présentées ci-dessus m'ont permis d'obtenir un grand jeu de données eucaryotes tout en cherchant à minimiser la présence de séquences paralogues et xénologues. Cependant, malgré les multiples contrôles qualité que j'ai réalisés, les faibles supports observés en jackknife de gènes peuvent suggérer que certaines séquences de ce type auraient subsisté. Notamment, la présence de gènes transférés en provenance des endosymbiontes (EGT) pourraient être une source possible de gènes non-orthologues. Celles-ci est d'autant plus crédible que les faibles supports se concentrent autour des espèces photosynthétiques. Vérifier cette hypothèse nécessiterait de rechercher concrètement les gènes présentant des incongruences topologiques avec l'arbre d'espèces. Ceci pourrait être réalisé de plusieurs manières comme par exemple, en vérifiant la présence de clan prédéfini ou en analysant directement les bipartitions retrouvées pour l'ensemble des arbres de gènes. Deux scénarios pourraient émerger de ces résultats : (i) l'existence d'un nombre restreint de solutions bien supportées pour les noeuds à la base des Diaphoretickes, suggérant l'application de méthode de réconciliation ; (ii) l'émergence d'aucune solution en particulier, mettant l'accent sur la présence d'un faible signal phylogénétique et, par conséquent, l'importance de minimiser la création d'erreur systématique.

S'il est clair que l'erreur systématique a affecté l'inférence avec le modèle LG4X, il est plus difficile, en l'état, de quantifier son impact sur les inférences réalisées avec les modèles CAT et C60 et la méthode PMSF. Même si les arbres sont très similaires, la différence de supports obtenus pour les Archaeplastida et les Cryptista me parait anormale, surtout le degré de certitude de PMSF. Concernant ce dernier, il faudrait vérifier que le résultat n'est pas influencé par le choix de l'arbre de départ. Un autre moyen de vérifier l'existence d'un possible biais sur ces analyses serait de tester la congruence des topologies obtenues suite à divers retraits de clade. Cependant, ceci impliquerait la réalisation de nouvelles inférences très coûteuses en ressources computationnelles.

Malgré la possible présence d'erreurs dans le jeu de données, les topologies obtenues sont en accord avec les phylogénies réalisées après la construction de ce dernier [Strassert et al. 2019 ; Brown et al. 2018 ; Heiss et al. 2018 ; Lax et al. 2018 ; Cenci et al. 2018]. De manière intéressante, le regroupement des Haptophyta avec les SAR et celui des Cryptista avec les Archaeplastida ne sont pas toujours retrouvés dans ces différentes analyses, malgré

l'emploi du même modèle (i.e. LG+C60+Γ4-PMSF). Il me semble donc important de bien considérer l'incongruence entre gènes, notamment par l'emploi du jackknife et, à terme, par l'emploi des méthodes de réconciliation.

Une autre piste, limitant les problèmes de transfert, serait la réalisation d'une phylogénie des eucaryotes minimisant la présence des organismes photosynthétiques comme envisagé récemment [Cenci et al. 2018]. Cela nécessiterait l'intégration de plusieurs clades affiliés aux Diaphoretickes commes les centrohelidés, les picozoaires ou les télonemidés mais également un bon échantillonnage en Stramenopiles et Rhizaria non-photosynthétiques, ainsi que de leur groupe extérieur probable, les Excavata. Cependant, ceci constituerait une autre étude car une telle modification de l'échantillonnage taxonomique du jeu de données actuel entraînerait de refaire une grande partie des vérifications pré inférence. D'ailleurs, il serait surement plus logique d'inférer les OGs sans organismes photosynthétiques.

Chapitre 3

Phylogénomique des Stramenopiles

3.1 Résumé du chapitre 3

Ce troisième et dernier chapitre se base sur des résultats cristallisés en un article en cours de soumission. Dans cette article, nous étudions l'impact du choix de modèle d'évolution des séquences lors de l'inférence de phylogénies complexes, illustrées ici par la radiation des straménopiles photosynthétiques (Ochrophyta). Pour ce faire, nous avons construit un jeu de données phylogénomique pour chaque compartiment possédant un matériel génétique chez ces organismes (la mitochondrie, le chloroplaste et le noyau). Nous avons comparé les phylogénies obtenues avec ces différents jeux de données et observé des incongruences. Grâce à différentes expériences, nous avons mis en évidence que ces dernières pouvaient s'expliquer par la présence de violation du modèle d'évolution des séquences menant à l'inférence de topologie artéfactuelle. Au final, nous obtenons un arbre résolu des ochrophytes et nous mettons en avant l'importance des gènes riches en signal phylogénétique pour les noeuds étudiés.

Lower statistical support with larger datasets: insights from the Ochrophyta radiation

Arnaud Di Franco[1], Denis Baurain[2], Gernot Glöckner[3], Michael Melkonian[4], and Hervé Philippe*[1,5]

[1]Station d'Ecologie Théorique et Expérimentale de Moulis, UMR CNRS 5321, Moulis, France
[2]InBioS–PhytoSYSTEMS, Unité de Phylogénomique des Eucaryotes, Université de Liège, Liège, Belgium
[3]Institut für Biochemie I,Medizinische Fakultät, Universität zu Köln, Köln, Germany
[4]Botanisches Institut, Biozentrum Köln, Universität zu Köln, Köln, Germany
[5]Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, Montréal, Québec, Canada
*Corresponding author: E-mail: herve.philippe@sete.cnrs.fr

Abstract

It is commonly assumed that increasing the number of characters is key to resolving radiations. We studied photosynthetic stramenopiles (Ochrophyta) using alignments of heterogeneous size and origin (6,762 sites for mitochondrion, 21,692 sites for plastid and 209,105 sites for nucleus). While statistical support for the relationships between the six major Ochrophyta lineages increases when comparing the mitochondrion and plastid trees, it decreases in the nuclear tree. Statistical support is not simply related to the dataset size but also to the total quantity of phylogenetic signal available at each position and our ability to extract it. Here, we show that proper signal extraction is difficult to attain, as demonstrated by conflicting results obtained when varying taxon sampling. Even though the use of a better fitting model improved signal extraction and reduced conflicts, the plastid dataset still resolved the ochrophyte radiation better than the larger nucleus dataset. We propose that the higher support observed in the plastid tree is due to an acceleration of the evolutionary rate in one short deep internal branch, implying that more phylogenetic signal per position is available to resolve the Ochrophyta radiation in the plastid than in the nuclear dataset. Our work therefore suggests that, in order to resolve radiations, beyond the obvious use of datasets with more positions, we need to continue developing models of sequence evolution that better extract the phylogenetic signal and to design methods to search for genes/characters that contain more signal specifically for short internal branches.

Keywords: Phylogenomics, Radiations, Phylogenetic signal, Model of sequence evolution, Long Branch Attraction

Introduction

One of the last major challenges of phylogenetics is to resolve ancient rapid radiations. They combine two main difficulties encountered during phylogenetic inference. First, they are characterized by short internal branches, meaning a scarce genuine (historical) phylogenetic signal, which resides in the rare substitutions accumulated during a short amount of time (i.e., very few synapomorphies). Second, their ancient nature makes the sites subject to substitutional saturation (multiple substitutions at the same site). As a result, an ancient synapomorphy can easily be masked by subsequent substitutions, leading the tree reconstruction method to possibly interpret it as a convergence. The misinterpretation of site history, if not random (i.e., biased), creates a non-phylogenetic signal that conflicts with the genuine phylogenetic signal. To sum up, the phylogenetic signal left behind by ancient rapid radiations is both scarce and difficult to extract (Whitfield and Lockhart 2007). Because increasing the size of the dataset is a necessary condition to resolve such radiations, phylogenomics, the use of large datasets, initially generated great hope (Gee 2003). However, the failure of phylogenomics to resolve most of those complex cases suggests that more data may not be sufficient (see Philippe et al. 2011).

This failure might be due to limitations of existing tree reconstruction methods. Let us consider a dataset D containing n sites. The genuine phylogenetic signal (PS) contained in it for a given branch B is:

$(1)\ PS_B(D) = n * \lambda_B(D)$,

with $\lambda_B(D)$ being the mean substitution rate in branch B for dataset D. Yet, $PS_B(D)$, the total number of synapomorphies of the dataset supporting this branch, is actually unknown, as we only have access to the apparent phylogenetic signal (PS') (Baurain and Philippe 2010) inferred by a specific method m:

$(2)\ PS'_B(D|m) = PS_B(D) * EE(m|D,B) = n * \lambda_B(D) * EE(m|D,B)$,

with $EE(m|D,B)$ being the extraction efficiency of method m for branch B and dataset D. This efficiency depends not only on the properties of the radiation (e.g., age) and the dataset (e.g., global rate of evolution or taxon sampling), but also ultimately on how correctly the model of evolution infers the substitution history at each position. Limitations of tree reconstruction methods generally lead to values of $EE(m|D,B)$ smaller than one. $EE(m|D,B)$ might even be negative when a bias favors an alternative branching order (e.g., a long branch attraction between two unrelated fast-evolving taxa) and larger than one when such a bias favors the correct solution (e.g., a long branch attraction between two fast-evolving taxa that are really sisters). The failure to resolve most radiations (i.e., PS'(D|m) not significantly different from 0, as often revealed by a low bootstrap support) is due to the fact that $n*\lambda_B(D)$ and/or $EE(m|D,B)$ are too small. In phylogenomics, n tends to be maximal and improvements on $n*\lambda_B(D)$ become asymptotically smaller. This is especially true when considering the finite collection of orthologous sequences relevant to the issue at hand. While alternative approaches based on other types of characters may exist (e.g., retrotransposon insertions or intron positions), they will not be considered here. The hope of

supermatrix-based phylogenomics to resolve radiations now completely hinges on improving the extraction efficiency of the tree inference methods.

Since phylogenetic inference should be viewed as a statistical problem (Felsenstein 1983), it requires the formalization of an explicit model. The extraction efficiency of probabilistic tree reconstruction methods ultimately depends on the validity of model assumptions. In other words, model violations decrease extraction efficiency. Two main types of violations exist: (1) violations of the model of sequence evolution and (2) violations of the model of gene evolution. The first type of violation is unavoidable, as we fail to fully apprehend the complexity of sequence evolution, and it affects all tree reconstruction methods. The second violation is due to the fact that, because of incomplete lineage sorting, gene duplication/loss and horizontal gene transfer, single-gene trees might be different from the species tree (Maddison and Wiens 1997), which is not taken into account by the concatenation approach. In theory, gene duplication and horizontal gene transfer are not present, given that only orthologs should be included in the supermatrix. However, incomplete lineage sorting affects orthologs and is expected to be all the more frequent in phylogenies with short branch lengths. These violations can be addressed by coalescent-based methods that jointly infer gene and species trees (e.g., *BEAST) (Heled and Drummond 2010). However, they are still not accessible to phylogenomics because of their computational burden. If proxies to this joint inference do exist (e.g., ASTRAL) (Mirarab and Warnow 2015; Zhang et al. 2018), they are too sensitive to single-gene tree estimation errors to be considered accurate (Gatesy and Springer 2014). Consequently, only the effect of violations of the model of sequence evolution can currently be studied at the scale of phylogenomics, through the comparison of trees inferred by models fitting data at different degrees. Moreover, their universal impact, which is stronger for more ancient events like the radiation of Metazoa (Philippe et al. 2011; Simion et al. 2017), make model violations a central issue of phylogenetic inference.

To study the impact of extraction efficiency on the power of phylogenomics to resolve ancient radiations, we chose the diversification of Ochrophyta (i.e., photosynthetic Stramenopiles). Stramenopiles (also known as heterokonts) is a eukaryotic clade composed mostly of unicellular species, and is closely related to Alveolata and Rhizaria, the three clades forming the supergroup SAR (Burki et al. 2007). Inside Stramenopiles, Ochrophyta is a monophyletic group of photosynthetic organisms that appeared around 500 Mya (Brown and Sorhannus 2010). The diversity of this clade is large, ranging from the picoplanktonic *Nannochloropsis* (Eustigmatophyceae) to ecologically important diatoms (Bacillariophyta) and multicellular brown algae (Phaeophyceae). As photosynthetic eukaryotes, they harbor three genomic compartments, a nucleus (nu), a mitochondrion (mt) and a plastid (cp), the latter inherited from a red algal endosymbiont (Archibald 2015).

The diversification of the major ochrophyte lineages seems to have occurred relatively rapidly, as demonstrated by the corresponding short internal branches (Yang et al. 2012; Derelle et al. 2016). The apparent phylogenetic signal for these branches (B) is expected to vary across compartments. First, the three genomes have a quite different size ($n_{mt} < n_{cp} << n_{nu}$), suggesting that $PS'_B(nu|m) >> PS'_B(cp|m) > PS'_B(mt|m)$. Second, they have evolved under very different mutation/selection pressures (e.g., different G+C content, presence of recombination or not), leading to different mean substitution rates ($\lambda_B(mt) \neq \lambda_B(cp) \neq \lambda$

$_B$(nu)) and different extraction efficiencies (EE(m|mt,B) ≠ EE(m|cp,B) ≠ EE(m|nu,B)), due to differences in types and levels of model violations. Although the values of λ_B(D)*EE(m|D,B) vary across the three genomes, it is difficult to predict whether these variations are major. Ochrophyta constitute an interesting case study to evaluate the relative importance of n and λ_B(D)*EE(m|D,B) in our ability to resolve ancient radiations. In particular, it allows us to test whether the insistence of most phylogeneticists on increasing n rather than improving EE(m|D,B) is well-advised.

For this study, we sequenced mitochondrial and plastid genomes from five key ochrophyte species. From these new data, we built three supermatrices, one for each genomic compartment, all representing most of the major ochrophyte clades. Each dataset was carefully constructed, so as to maximize matrix size and completeness, while minimizing erroneous inclusion of non-orthologous genes, contaminated sequences and sequencing errors. With these three largest stramenopiles supermatrices to date, we studied how extraction efficiency affects phylogenetic inference. We first observed incongruent topologies between the three genomes for deep ochrophyte relationships, along with surprisingly lower bootstrap supports for the largest dataset when using the conventional model LG4X (Le et al. 2012). We then studied the impact of model violations by varying taxon sampling and by using an alternative model of sequence evolution. Finally, we explored ways to resolve the deep ochrophyte radiation.

Results & Discussion

Recovery of the major ochrophyte clades by all compartments

We carefully assembled three supermatrices from mitochondrial, plastid and nuclear genome sequences, containing 6,672, 21,692 and 209,105 amino acid positions, respectively (Table 1). They included species from eleven major clades of Ochrophyta: Bacillariophyta, Bolidophyceae, Chrysophyceae, Dictyochophyceae, Eustigmatophyceae, Pelagophyceae, Phaeophyceae, Pinguiophyceae, Raphidophyceae, Synurophyceae and Xanthophyceae. The nuclear dataset also included *Synchroma pusilla*, a species of Synchromophyceae. Mitochondrial, plastid and nuclear phylogenies inferred using the LG4X model (Fig. 1A-C) retrieved the monophyly of all major clades but Chrysophyceae with maximal bootstrap support (BS) (see Supplementary Figures 1-3). Chrysophyceae came out as a monophyletic group in the mitochondrion and plastid datasets (BS=56% and 88%, respectively), but were represented by two species only. In the nuclear phylogeny, which includes 12 chrysophycean species, they were paraphyletic, with Synurophyceae nested within Chrysophyceae, in agreement with previous studies (Yang et al. 2012). Monophyly of Synurophyceae+Chrysophyceae (SC clade) was always maximally supported. In the nuclear dataset, their grouping with Synchromophyceae (SSC clade) was highly supported, as in Yang et al. (2012) and Derelle et al. (2016). This SSC clade was also recovered in a plastid phylogeny built with partial *Synchroma* sequences obtained from RNAseq data (Keeling et al. 2014) (data not shown). In the what follows, we will consider the SSC clade as one of the now ten major ochrophyte clades contained in our analyses. The PX clade (Phaeophyceae and Xanthophyceae) (Kai et al. 2008) was recovered in all three datasets, as well as their sister relationship with Raphidophyceae (PXR clade), but with limited support in the

mitochondrial dataset (BS=92% and 71%, respectively). Two other previously recovered relationships (Yang et al. 2012; Derelle et al. 2016; Han et al. 2018) were highly supported – monophyly of Pelagophyceae and Dictyochophyceae (PD clade) and monophyly of Bolidophyceae and Bacillariophyta (BB clade) – but again mitochondrial support was not maximal for the PD clade (BS=85%). Inside Bacillariophyta, in our nuclear tree, Coscinodiscophyceae were paraphyletic at the base, followed by a monophyletic group composed of Mediophyceae and Bacillariophyceae, as in Parks et al. (2017). Overall, our results were thus in excellent agreement with existing knowledge.

Incongruence between compartments for deep ochrophyte relationships

Although the monophyly of the ten major clades was consistently recovered across the three datasets, the basal phylogeny of Ochrophyta showed incongruent relationships depending on which dataset was used. While the plastid tree strongly grouped Eustigmatophyceae with the SSC clade (BS=100), nuclear and mitochondrial trees separated them, and instead supported the grouping of Pinguiophyceae with SSC on one side (BS=68% and 74%, respectively) and the grouping of Eustigmatophyceae with PXR on the other side (BS=81% and 22%, respectively). Our plastid phylogeny (Fig. 1B) was in agreement with the work of Yang et al. (2012), which was based on nuclear SSU rRNA and four plastid-encoded genes, suggesting that their inferences were dominated by the plastid signal. Comparison with more recently published plastid (Ševčíková et al. 2015), mitochondrial (Ševčíková et al. 2016) and nuclear trees (Derelle et al. 2016) was more difficult, as those datasets lacked some major clades: Bolidophyceae, Dictyochophyceae and Pinguiophyceae (Ševčíková et al. 2015); Bolidophyceae, Dictyochophyceae, Pinguiophyceae and Xanthophyceae (Ševčíková et al. 2016); Eustigmatophyceae and Pinguiophyceae (Derelle et al. 2016). Still, the plastid tree of Ševčíková et al. (2015) was congruent with our plastid tree. However, the mitochondrial tree of Ševčíková et al. (2016) did not recover the monophyly of the PXR clade, contrary to our mitochondrial tree (Fig. 1A), but both trees recovered a basal position for Chrysophyceae. In our nuclear tree (Fig. 1C), we observed the dichotomy between Diatomista (PD+BB) and Chrysista (PXR+SSC), first proposed in Derelle et al. (2016). Although phylogenies based on the three genomic compartments yielded incongruent deep ochrophyte relationships (Fig. 1A-C), they were each in good agreement with previously published trees based on the same compartment (Yang et al. 2012; Ševčíková et al. 2015; Derelle et al. 2016; Ševčíková et al. 2016).

Unexpectedly, statistical support for the eight deep nodes that connect the ten major lineages, displayed a surprising pattern with respect to the number of positions (Fig. 1D). The average BS for these eight nodes increased from 73% in the mitochondrion tree (6,762 positions) to 97% in the plastid tree (21,692 positions), disregarding the fact that these two tree differed for basal relationships. With ~3 times more positions than the mitochondrial dataset, the plastid dataset thus confirmed the expectation that the apparent phylogenetic signal (as measured by bootstrap support) increases with dataset size. In sharp contrast, the nuclear dataset (209,105 positions), which is ~10 times larger than the plastid dataset (~7.5 times larger if taking into account missing data, see Table 1), did not follow that expectation, with an average bootstrap support of 86%. The deep ochrophyte phylogeny inferred from the

three compartments therefore showed not only incongruent relationships but also unexpected statistical supports.

Comparison of the apparent phylogenetic signal across the three genomes when controlling for the number of sites and the number of OTUs

There was more apparent signal in the plastid than in the nuclear dataset, PS'(cp|LG4X)>PS'(nu|LG4X), as shown by bootstrap support (Fig. 1D). According to equation (2), this would indicate that λ_B(cp)*EE(LG4X|cp,B) > λ_B(nu)*EE(LG4X|nu,B), since n_{cp} << n_{nu}. However, differences in taxon sampling (64/63 species in the mitochondrion/plastid datasets versus 124 in the nuclear dataset) can affect extraction efficiency, making our comparison of the three datasets difficult to interpret. To cancel out the impact of taxon sampling on extraction efficiency, we reduced the sampling of each dataset to a set of 23 common species (22 for the mitochondrion). The phylogenies (inferred with the same LG4X model as above) (Supplementary Figure 4) were virtually identical to those inferred with more species (Supplementary Figures 1-3). Yet, we observed a decrease in the apparent phylogenetic signal when reducing the number of species: the average BS for the eight deep nodes went down from 73% to 64% for the mitochondrion, from 97% to 93% for the plastid and from 86% to 69% for the nucleus. This is in agreement with the widely recognized idea that the use of a large number of species improves phylogenetic accuracy (Zwickl and Hillis 2002). The higher apparent phylogenetic signal in plastid versus nuclear compartment was thus still recovered, in spite of controlling for taxon sampling (i.e., making EE(LG4X|D) closer for the three genomes).

$$(2)\ PS'_B(D|m) = PS_B(D) * EE(m|D,B) = n * \lambda_B(D) * EE(m|D,B),$$

To better characterize the apparent phylogenetic signal of the three compartments, we used the variable length bootstraps (VLBs), or partial bootstrap, approach with the set of common species (Lecointre et al. 1994; Springer et al. 2001; Baurain et al. 2010). VLB analyses are designed to evaluate the speed at which maximal support is achieved. Usually, VLB curves are used to define the number of sites needed to reach a given level of apparent phylogenetic signal (e.g., BS=95%), in order to compare the resolving power of different datasets (Springer et al. 2001; Baurain et al. 2010). Here, they allowed us to study the variation in apparent phylogenetic signal between the three compartments without being affected by the different sizes of the datasets.

Interestingly, VLBs revealed that the apparent phylogenetic signal of most nodes was highly similar for the nucleus, the plastid and, to a lesser extent, the mitochondrion (Fig. 2). For the monophyly of the major clades (Fig. 2A-F), VLB curves always reached 100% BS below 1000 positions. For the five higher order groupings (BB, PX, PD, PXR, and PD+BB) that were recovered with the three genomes (Fig. 1A-C), the curves displayed similar increasing trends between compartments (Fig. 2G-K). The mitochondrial dataset required more sites to reach a given BS, which could be due to a reduced extraction efficiency related to the high rate of evolution in this compartment (Neiman and Taylor 2009). Yet, nucleus and plastid curves were virtually identical, sometimes the plastid increasing slightly faster (Fig. 2H) or slower (Fig. 2I) than the nucleus. In sharp contrast, support for the monophyly of E+SSC

(Fig. 2L), as well as their subsequent grouping with Pinguiophyceae and PXR (Fig. 2M), rose much faster and higher in the plastid dataset than in the two other compartments. None of the bipartitions conflicting with E+SSC (Fig. 2N-O) showed the same rapid increase in the mitochondrion or the nucleus.

Hypotheses to explain the conflict between plastid and nucleus

When controlling for the number of species and the number of sites, the apparent phylogenetic signal in plastid and nuclear compartments was almost identical for most nodes, meaning that λ_B(cp)*EE(LG4X|cp,B) ~= λ_B(nu)*EE(LG4X|nu,B). The higher average support observed over the eight deep nodes with the plastid dataset (Fig. 1D) was therefore mainly due to a few nodes (e.g., E+SSC and E+SSC+Ping+PXR), for which λ_B(cp)*EE(LG4X|cp,B) > λ_B(nu)*EE(LG4X|nu,B). The comparison of branch lengths in Fig. 1 allowed us to formulate two hypotheses. In the plastid tree (Fig. 1B), Eustigmatophyceae and SSC are connected by a long internal branch; assuming that E+SSC is correct, this implies a high value of λ_{E+SSC}(cp). However, Eustigmatophyceae and SSC evolve much faster than all other taxa. Their grouping may thus be the result of a long branch attraction. Note that this is not the long branch attraction artifact (LBA) originally described by Felsenstein (1978) in the case of maximum parsimony, because probabilistic methods used here do take branch lengths into account. Nevertheless, fast evolving lineages not only evolve faster but also evolve differently from other lineages, being subject to heterotachy (e.g., differences in the sets of sites free to vary (Lockhart 1996; Germot and Philippe 1999)) and/or heteropecilly (different substitution processes at play) (Roure and Philippe 2011), which violate the stationarity assumption made by almost all models. For the sake of simplicity, in what follows, we will present our results in terms of LBA, without always expliciting that LBA is due to model violations. In the case of fast evolving Eustigmatophyceae and SSC, such hypothetical model violations might be so important that they would lead to a negative value of extraction efficiency (EE), creating a false signal for an incorrect node. Hence, an erroneous branch B* (accounting for E+SSC) would be supported by an apparent phylogenetic signal equal to $PS'(B*) = n * \lambda_B(D) * (-EE(m|D,B)$), with B being the correct (unknown) branch.

Two hypotheses can therefore explain the large difference in apparent signal for E+SSC between plastid and nuclear compartments: (1) the apparent phylogenetic signal in the plastid is mostly based on a genuine signal and more of it is available per position due to a rate acceleration in the common ancestor of Eustigmatophyceae and SSC (high value of λ_{E+SSC}(cp)), (2) the apparent phylogenetic signal in the plastid is driven by model violations resulting in the artifactual grouping of the long branches leading to Eustigmatophyceae and SSC (EE(LG4X|cp,B)<0). As both compartments likely share the same vertical history in Ochrophyta, hypothesis (1) would imply that the groupings E+PXR and Ping+SSC inferred with the nucleus supermatrix are artifactual (EE(LG4X|nu,B)<0) while hypothesis (2) would imply they are correct. Since both hypotheses imply an erroneous branching due to a negative value of extraction efficiency (EE), distinguishing between them requires estimating whether the unavoidable model violations are sufficient to generate erroneous trees with the plastid or the nucleus datasets. Here, we applied two commonly used approaches against the LBA artifact (i.e., to reveal the effect of model violations): varying taxon sampling (to

favor or disfavor LBA) and using different models of sequence evolution (more or less sensitive to model violations).

Evidence for the presence of model violations

First, we evaluated the impact of major variations of the taxon sampling. The rationale was to reveal a possible inconsistency of the tree reconstruction method (i.e., model violations sufficiently important to make EE(m|D,B)<0) through the discovery of incongruence between phylogenies inferred from different subsets of species. We investigated two strategies: (1) use of only a distant outgroup (to increase LBA by creating a long unbroken branch) and (2) independant removal of highly supported ochrophyte lineages. We selected the six clades that were strongly supported by the three datasets: Eustigmatophyceae (E), Pinguiophyceae (Ping), SSC clade (represented by SC in the plastid), PXR clade, PD clade and BB clade. Since the phylogenetic signal is more accurately extracted with many taxa, we focus on the analyses with complete (albeit different, see above) taxon sampling. Analyses with the common set of 23 species returned comparable results, but with weaker BS (Supplementary Table 1). All groupings of the six clades of interest observed through the 14 taxon sampling variations (2 compartments * 1+6 taxon samplings) are reported in Table 2.

For the plastid, only two taxon samplings produced an incongruent topology, the use of a distant outgroup and the removal of Pinguiophyceae (3 incompatibilities, BS shown in boldface in Table 2). Both resulted in the same artifactual topological move: the attraction of the fast-evolving E+SSC group by the outgroup. Attractions were explained by the presence of a longer unbroken branch, either the distant outgroup or the branch of E+SSC in the absence of their slowly-evolving sister-group Pinguiophyceae. Taxon sampling variations of the plastid dataset revealed that model violations of LG4X sometimes produced LBA artifacts. This indicated a limited extraction efficiency with this combination of model and dataset, suggesting that hypothesis (2) may be correct. Yet, it is important to notice that the grouping E+SSC was always observed.

Variations of the nucleus dataset showed more incompatibilities with the tree inferred from all species (10, in boldface in Table 2, corresponding to 6 alternative groupings) than the plastid (only 3). The incompatibilities were more difficult to understand, as the six clades evolved at a more homogeneous rate in the nucleus than in the plastid (Fig. 1B-C). Pinguiophyceae appeared to be the most unstable clade: they emerged as sister-group to the remaining ochrophytes (BS=100%) with the distant outgroup, as sister-group to BB (BS=95%) when SSC was removed, and as sister-group to SSC (BS=100) when BB was removed. Pinguiophyceae were only represented by two closely related species (*Phaeomonas* and *Pinguiococcus*), leaving a long unbroken branch at their base (Fig. 1C). As BB and SSC are the fastest evolving clade in the nucleus, the placement of Pinguiophyceae can be explained by LBA with the longest branch available in each of the three cases: the outgroup, BB, and SSC respectively. The limited support for Ping+SSC (BS=68%) with the complete dataset would then result from an average among these contradictory attractions. The removal of PXR, a relatively slowly evolving clade, had the most dramatic effect, all the deep relationships becoming incongruent. It may have allowed the grouping of E+SSC (BS=91%), reducing the attraction between SSC and

Pinguiophyceae, the latter being then attracted by BB (BS=88%). Interestingly, E+SSC was also recovered through the removal of Pinguiophyceae (BS=56%). Altogether, these results suggest the presence of a high amount of non-phylogenetic signal under the LG4X model in the nuclear dataset (i.e., a very limited extraction efficiency), thereby supporting hypothesis (1).

While the taxon sampling variations of the nuclear dataset argued in favor of hypothesis (1), as E+PXR and Ping+SSC groupings failed to be robustly recovered, the plastid dataset also showed incongruence that may argue in favor of hypothesis (2). The higher number of incompatibilities with the nucleus than the plastid (10 versus 3) argue for a less efficient extraction efficiency for the former (EE(LG4X|nu,B)<EE(LG4X|cp,B)), hence a less reliable tree. For instance, whereas plastid samplings consistently recovered two high-level clades (E+SSC and PD+BB), the nuclear dataset failed to recover any such clade consistently. However, the main result of taxon sampling variations was the evidence for a major impact of model violations with LG4X, especially for the nuclear compartment (i.e., limited and/or negative extraction efficiency). These observations were in agreement with the sensitivity to LBA of models that do not fully incorporate the heterogeneity of the substitution process across sites (Lartillot et al. 2007; Philippe et al. 2011; Simion et al. 2017) and suggest that we need to use a better model to separate between the two hypotheses.

Impact of using a better fitting model of sequence evolution

We tested the site-heterogeneous CAT model (Lartillot and Philippe 2004) that has been shown to be less sensitive to LBA (Lartillot et al. 2007). Direct model comparison between CAT and LG4X was not possible because the latter model is not implemented in PhyloBayes-MPI. This is why we first compared LG4X to GTR using ModelFinder (Kalyaanamoorthy et al. 2017) from IQTREE (Nguyen et al. 2015), which showed GTR to be better than LG4X with both the plastid and nuclear datasets (Supplementary Table 2). Second, cross-validation demonstrated CAT to have a better fit than GTR for both datasets (plastid: likelihood difference between CAT and GTR of 234 +/- 169; nucleus: 3467 +/- 257). Consequently, the combination of these two tests showed CAT to have a better fit than LG4X for our datasets. Given its moderate size (63 x 21,692), we were able to compute CAT bootstrap support for the complete plastid dataset. In contrast, this computationally expensive model could not be used on the much larger nuclear dataset (124 x 209,105). To make it computationally tractable, we resorted to a gene jackknife approach instead (Delsuc et al. 2008). We chose to generate datasets of ~20,000 positions to be comparable in sample size with the plastid and to run 50 replicates. The resulting jackknife supports (JS) are expected to be lower than BS. Although one cannot directly compare BS of LG4X and JS of CAT, we can use JS to evaluate the effect of taxon sampling on the nucleus-based phylogeny inferred with the better fitting CAT model. Since the jackknife replicates were relatively small, we could expect a larger stochastic error and a lower statistical power to learn site-specific amino acid propensities, the key aspect of the CAT model. Therefore, the use of JS instead of BS was expected to lead to more variance in the results of the taxon sampling variations for the nucleus, i.e., more incongruence.

The plastid tree inferred using the CAT model (Fig. 3A) gave the same topology as LG4X, but with lower statistical support, especially for E+SSC (BS=84% versus 100%) and to position Pinguiophyceae sister of E+SSC (BS=58% versus 95%). Lower support for E+SSC can be explained by the fact that the CAT model is more cautious when it has to group two long branches (Eustigmatophyceae and SSC) together. In other words, it assumes more shared amino acids to be due to convergence than LG4X, the very reason of its reduced sensitivity to LBA (Lartillot et al. 2007). In contrast, the topology of the phylogeny inferred from the nucleus supermatrix using CAT (Fig. 3B) was different from LG4X (Fig. 1C). Only the monophyly of PD+BB (JS=76%) was common among the high-level relationships between the six major clades. In the CAT tree, SSC was sister of Eustigmatophyceae (JS=40%) instead of Pinguiophyceae, while the latter was sister of PD+BB (JS=48%). Finally, PXR was weakly grouped with E+SSC (JS=20%). Overall, the use of a better fitting model, which likely improves extraction efficiency, increased the congruence between plastid and nucleus trees, as shown by the common recovery of the relationship between Eustigmatophyceae and SSC (E+SSC), a key relationship that allows us to distinguish between the two hypotheses explaining the conflicts observed between the two compartments when using the LG4X model.

To confirm that using a better fitting model increases extraction efficiency and therefore reduces incongruence, we performed the same variations of taxon sampling as above, but using the CAT model. The results (Table 3) showed less incompatibilities within each compartment (1 versus 3 for the plastid and 8 versus 10 for the nucleus) and better congruence between the two compartments. In particular, CAT recovered PD+BB in all analyses of the two compartments. Also, while LG4X rarely supported E+SSC in the nucleus, it was consistently found by CAT, except when using a distant outgroup, a situation likely to maximize the effect of model violations. Yet, in that case, the alternative groupings (Ping+SSC and E+PXR) received a low JS (30% and 20%, respectively). The position of Pinguiophyceae and PXR remained unstable, displaying various sister relationships to one of the two previous clades with limited support (maximum JS of 55%). Despite the use of fewer sites (20,000) than LG4X (i.e., increased stochastic error), CAT thus turned out to be more robust to taxon sampling variations, demonstrating its success in reducing the amount of non-phylogenetic signal (i.e., increasing extraction efficiency).

Finally, we studied the effect of limiting the nuclear dataset to 20,000 positions. Reducing dataset size to one tenth of its actual size likely affects the total amount of apparent phylogenetic signal. We performed a taxon sampling experiment using gene jackknife replicates of 80,000 positions, but with only the 23 species common to the plastid and nucleus datasets, since this combination was computationally tractable. Results (Supplementary Table 3) were highly similar, indicating that the use of only 20,000 sites was not misleading. As expected, JS were higher with 80,000 than with 20,000 positions. For instance E+SSC received a support of 84% instead of 40% in the presence of the close outgroup. Overall the incongruence generated by taxon sampling variations was similar, with 3 incompatibilities in common. The only difference was for the removal of BB: with 20,000 positions, Eustigmatophyceae were sisters of PXR (JS=29%), whereas with 80,000, they remained sisters of SSC (JS=78%). This confirmed our assumption that the use of only 20,000 sites increases sensitivity to taxon sampling, hence disfavoring CAT when comparing

it to LG4X. Even with this disadvantage, CAT was more robust to taxon sampling variations, demonstrating its higher efficiency to extract an ancient phylogenetic signal. Of note, CAT always recovered the monophyly of E+SSC when using 80,000 sites (minimal JS of 78%).

Improving extraction efficiency favored hypothesis (1), that is the grouping of Eustigmatophyceae and SSC (E+SSC) was correct and more highly supported in the plastid than in the nucleus compartment because of an acceleration of the substitution rate in the branch at the base of E+SSC in plastid loci. First, E+SSC was always recovered by the two compartments for sixteen (2*8) different taxon samplings but one. Second, this relationship is also supported by a common split of the plastid encoded gene clpC, which is involved in the protein degradation pathway mediated by the ClpP protease (Ševčíková et al. 2015). Third, we observed five common losses of plastid genes in the two clades (ATP synthase CF1 delta subunit, hypothetical protein Ycf39/Isoflavone reductase, PSI reaction center subunit XII, hypothetical protein Ycf35 and cytochrome b6-f complex subunit 6/petL, data not shown), although it is difficult to exclude the possibility of convergence since the plastid genome is reduced in both cases. Overall, the use of the CAT model demonstrated the key importance of the extraction efficiency in resolving ancient radiations. Our experiments proved that improving extraction efficiency through selection of a better fitting model reduced incongruence and increased robustness to taxon sampling variations.

Using branch length heterogeneity to tackle radiation

Hypothesis (1) postulates an acceleration of the evolutionary rate in the (genuine) internal branch connecting Eustigmatophyceae and SSC (i.e., high value of $\lambda_{E+SSC}(cp)$). However, owing to the important model violations that favor LBA with LG4X (Table 2), the grouping of the fast evolving Eustigmatophyceae and SSC obtained with this model (and the corresponding internal branch length) is likely inflated (i.e., EE(LG4X|cp,E+SSC)>1). It is thus difficult to assess the relative value of $\lambda_{E+SSC}(cp)$ versus $\lambda_{E+SSC}(nu)$. Even with the better fitting CAT model, model violations can inflate the apparent phylogenetic signal for E+SSC. If we neglect the confounding effect of model violations in estimating the true value of λ_{E+SSC}, this underlines the interest of finding markers with relatively long internal branches (i.e., possible high value of λ) to resolve radiations.

Obviously, looking for such genes is difficult because it requires to be able to accurately infer the value of λ. However, testing the potential of such an approach is possible by assuming the knowledge of the correct phylogeny. More precisely, we can estimate branch lengths for each gene, constrained to a candidate topology, and select those displaying the longest (or shortest) length for the branch of interest. Finally, we can infer a phylogeny using a concatenation of the resulting set of markers and compare it to the phylogeny obtained without such a selection to study the effect of filtering the dataset by the signal of interest.

We applied this protocol, using the LG4X model, to the nucleus dataset by selecting the 200 genes with the longest internal branch at the base of E+SSC, yielding a supermatrix of 47,386 positions ($LONG_{nu}$) and, as a negative control, the 200 genes with the shortest internal branch, yielding a supermatrix of 39,867 positions ($SHORT_{nu}$). Not surprisingly, the phylogeny inferred from $SHORT_{nu}$ with the LG4X model (Supplementary Figure 5A) did not

recover E+SSC, but strongly grouped Pinguiophyceae and SSC (BS=96%) and Eustigmatophyceae and PXR (BS=100%), in agreement with the topology observed with the full dataset (Fig 1C). In the absence of a strong genuine phylogenetic signal (for E+SSC), the misleading non-phylogenetic signal dominated, and the apparent phylogenetic signal for two erroneous groupings (P+SSC and E+PXR) increased (BS from 68/81 to 96/100, respectively). In contrast, the phylogeny inferred from $LONG_{nu}$ with LG4X (Supplementary Figure 5B) strongly supported E+SSC (BS=100%) with the same complete taxon sampling. This suggests that the genuine phylogenetic signal was now stronger than the non-phylogenetic signal created by the serious violations affecting this model, thereby leading to a strong apparent signal in favor of E+SSC. As the full dataset did not support E+SSC, the non-phylogenetic signal produced over 209,105 positions is probably stronger than the corresponding signal in the 47,386 positions of the $LONG_{nu}$ set of genes. This protocol cannot be used to resolve radiations, because it assumes the species phylogeny to be known, but it can be used to reveal the contradictory attractions present in a large dataset (here SSC attracted either by Eustigmatophyceae or by Pinguiophyceae), these attractions stemming either from model violations or from the genuine (historical) signal. More importantly, it validates the idea of looking for innovative methods to detect genes with a high signal for internal nodes of a species phylogeny, disregarding the global topologies of the gene trees. Such approaches could be another avenue to alleviate the impact of model violations when trying to resolve radiations, without designing ever-more complex evolutionary models.

Towards resolving the Ochrophyta phylogeny

Our analysis showed that the resolution of the deep ochrophyte relationships was extremely difficult, because of short internal branches and serious model violations. Interestingly, the small plastid dataset appeared to contain a relatively large amount of phylogenetic signal, in particular because of its high value of $\lambda_{E+SSC}(cp)$. Since the CAT model did not show evidence of a negative value of extraction efficiency for the nucleus or the plastid, it should be interesting to combine the high number of positions of the nucleus and the high λ of the plastid to increase the apparent phylogenetic signal of the ochrophyte radiation. Indeed, by combining the plastid and nuclear datasets, we should lengthen at least one of the difficult branches, i.e., $\lambda_{E+SSC}(nu+cp) > \lambda_{E+SSC}(nu)$, thus making the problem easier to resolve for any method of phylogenetic inference. However, there are potential drawbacks to this approach, such as the fact that combining those datasets would reduce the taxon sampling down to 23 common species, along with the potential introduction of additional model violations (in particular branch length heterogeneity across compartments) (Kolaczkowski and Thornton 2004). Whereas these drawbacks might both decrease extraction efficiency, reducing the number of species allowed us to use more sites (80,000) with the best fitting model (CAT).

Such a combined phylogeny inferred with the nu+cp supermatrix using the CAT model (Fig. 4) showed a much higher support for deep ochrophyte relationships: PD+BB (JS=100%), Ping+PD+BB (JS=100%), E+SSC (JS=99%) and PXR+E+SSC (JS=90%). However, a higher statistical support is not necessarily incontrovertible evidence for a given grouping, as the inference method might be inconsistent. Therefore, we applied the same taxon sampling variations as above (i.e., the use of a distant outgroup and the removal of each major

ochrophyte clade) to the nu+cp supermatrix. Interestingly, all seven variations returned trees fully compatible with the phylogeny of Fig. 4 (Table 4). In contrast, the use of a less fitting model (LG4X) on the same supermatrix yielded a lower support and displayed sensitivity to taxon sampling (Supplementary Table 4), thereby confirming the key role of the model of sequence evolution in the accurate resolution of short internal branches. Even under difficult phylogenetic inference conditions (limited extraction efficiency due to a small number of taxa and residual violations affecting the CAT model), the robustness to taxon sampling variations argued for the nu+cp phylogeny (Fig. 4) to be a credible working hypothesis for the deep ochrophyte relationships.

Conclusion

A common belief is that increasing the number of positions (n) is key to resolving evolutionary radiations. Our work confirms that this is a necessary condition (Supplementary Table 3), but that the two other terms of the apparent phylogenetic formula proposed here, i.e., branch length (λ) and extraction efficiency (EE), cannot be neglected. In particular, extraction efficiency is a major limiting factor, because our models are necessarily over-simplified with respect to the complexity of biological evolution. The accumulation of data, not of positions but of species, is certainly useful, as the use of more taxa generally helps in the extraction of phylogenetic signal. Yet, this approach has some serious limitations, (1) some branches are unavoidably unbroken because of extinction, (2) some (rogue) species decrease extraction efficiency, and (3) the resulting increase in computational time limits us to the use of simplest models. Studies are thus needed to evaluate what are the best compromises between number of species and complexity of models to optimize extraction efficiency.

The reduction of model violations achieved when dropping LG4X in favor of the CAT model allowed us to reduce the incongruence revealed by taxon sampling variations and to improve the resolution of the ochrophyte radiation, especially for the nucleus dataset. However, the CAT model is still far from perfect. For instance, it does not take into account the genetic code to weight amino acid substitutions (see Rodrigue et al. 2010), and it assumes that the evolutionary process is the same all over the phylogeny (e.g., ignoring compositional biases, heterotachy or heteropecilly). These simplifications are bound to result in model violations that could lead to an incorrect phylogeny. The improvement of models of sequence evolution, both in terms of fit and of computational efficiency, should thus be a priority to resolve ancient radiations. For recent radiations, the impact of these model violations is expected to be more limited (fewer multiple substitutions at the same position), and it is key to address another kind of model violation (not studied in our work), the presence of incomplete lineage sorting, using coalescent methods such as *BEAST (Heled and Drummond 2010). However, when a non-negligible fraction of gene trees is different from the species tree, the interest of resolving the radiation is limited because hemiplasy is so frequent that the species tree is no longer useful to study the evolution of characters and organisms (Hahn and Nakhleh 2016).

In addition to the number of positions and the extraction efficiency, the strength of the apparent phylogenetic signal is also dependent on branch length. The length of a given

branch is variable across loci, e.g., being longer at a locus that underwent reduced purifying selection or positive directional selection. In the case of the plastid dataset, we were lucky to have had a large number of loci that underwent a substitution rate acceleration in the E+SSC basal branch. This acceleration likely explains the observation that the small plastid dataset (21,692 positions) is able to strongly recover the monophyly of this clade otherwise very difficult to resolve, while the large nuclear dataset (209,105 positions) cannot. The difference was more pronounced with the LG4X than with the CAT model, probably because the long branches of Eustimatophyceae and SSC were further artifactually attracted (i.e., EE(LG4X|cp,E+SSC)>1). This “lucky” rate acceleration suggests a new approach to resolve ancient radiations: searching for loci having accelerated in the short internal branches of interest, so as to facilitate extraction of a signal that is scarce for other, more regular, loci.

Finally, combining the nuclear and plastid datasets, along with the use of the CAT model, allowed us to simultaneously increase n, λ and EE, leading to a well-supported tree, robust to taxon sampling variations. Given the difficulties to resolve the ochrophyte radiation, this phylogeny needs to be confirmed with a richer taxon sampling and with a better model. It nevertheless constitutes a working hypothesis to understand in which order the remarkably diverse phenotypes of Ochrophyta emerged, from the picoplanctonic *Nannochloropsis* to the silica frustule-bearing diatoms and giant marine kelps.

Materials and Methods

Cultures, organelle genome sequencing and assembling

Cultures of *Chromulina chionophila* (CCAP 909/9), *Pseudopedinella elastica* (SAG B43.88), *Synura petersenii* (CCAC 0052), *Phaeomonas parva* (CCMP 2877), and *Florenciella parvula* (RCC 446) were obtained from their respective algal culture collections (CCAP: https://www.ccap.ac.uk/, SAG: http://www.uni-goettingen.de/en/culture+collection+of+algae+%28sag%29/184982.html, CCAC: http://www.ccac.uni-koeln.de/, CCMP: https://ncma.bigelow.org/, RCC: http://roscoff-culture-collection.org/). Algae were grown in the culture media recommended by the collections in aerated 1 L Erlenmeyer flasks at 15 °C and 20 µmol photons/m2/s in a 14:10 hr L/D cycle. They were harvested by centrifugation and, after grinding in liquid nitrogen, total DNA was extracted using either the NucleoSpin® Plant II Midi Kit (Macherey-Nagel, Düren, Germany) or a modified CTAB protocol (Rogers and Bendich 1985; see Suppl. Materials).

Sequencing and assembly of organelle genomes

DNA samples were converted to Illumina sequencing libraries according to the manufacturer’s protocols and sequenced in paired end mode (150 bases sequencing length). The resulting reads were assembled using ABySS (Simpson et al. 2009). Organellar contigs were extracted using gene sequences from the respective *Ectocarpus* genomes as queries in BLAST searches. Gaps were closed with GapFiller (Nadalin et al 2012) and annotation was carried out with the sequin tool from NCBI. Organelle genomes and their

corresponding annotations are available online in the NCBI databases with the following accessions: XXX (to be provided).

Creation of phylogenomic datasets

For each compartment, we assembled the datasets following a semi-automatic protocol similar to the one described in our previous phylogenomic studies (Simion et al. 2017, Irisarri et al. 2017). In summary, we used protein annotations obtained from genomic data to define orthologous groups with OrthoFinder version 1.4 (Emms and Kelly 2015). Sequence similarity matrices were computed with BLAST for mitochondrial and plastid datasets and with USEARCH (Edgar 2010) for the nuclear dataset (e-value threshold = 1e-5) before being divided with the MCL algorithm using the default inflation value (1.5). We filtered the resulting orthogroups for minimal taxonomic representation before validating their orthology relationships. Then we improved their taxon sampling by adding species from transcriptomic and genomic data using Forty-Two (https://bitbucket.org/ dbaurain/42/). Detailed description for each compartment, as well as on the computational treatments undertaken to remove paralogous and xenologous sequences from the multiple sequence alignments, can be found in the Supplementary Information. Finally, our analyses focused on the three datasets summarised in Table 1 and available at **https://doi.org/10.6084/m9.figshare.7680395.v1**.

Phylogenetic inferences

All supermatrices used in our analyses were concatenated using SCaFoS (Roure et al. 2007). We inferred phylogenetic trees using RAxML version 8.2 (Stamatakis 2014) with the LG4X mixture model (Le et al. 2012) using 100 fast bootstrap replicates. Inferences under the CAT+Γ4 mixture model (Lartillot and Philippe 2004) were carried out using PhyloBayes-MPI version 1.8 (Lartillot et al. 2013), either on bootstrap replicates for mitochondrial and plastid datasets or on gene jackknife replicates for the nucleus dataset. Preliminary analyses demonstrated that convergence was not reachable for a dataset of 124 species and 209,105 amino-acid positions with the current implementation of PhyloBayes-MPI. Following Delsuc et al. (2008), we thus used a gene jackknife approach and generated replicates of ~20,000 or ~80,000 positions with a custom script. Convergence assessment and consensus tree construction were performed as in Simion et al. (2017).

VLB analyses

We reduced each dataset to an ingroup taxon sampling of 22 comparable species (21 for the mitochondrion as one out of four Pelagophyceae species was missing), i.e., identical or closely related (Supplementary Table 5). For the outgroup, we used *Guillardia theta* for the plastid and *Phytophthora sojae* for the mitochondrion and *Phytophthora parasitica* for the nucleus. We used distinct species to have a similar branch length leading to the outgroup in each compartment, whereas using the same species (e.g., *G. theta*) would have generated a much longer branch in the mitochondrion/nucleus than in the plastid. Out of the three resulting supermatrices, we drew 1000 variable length bootstrap (VLB) replicates of different sizes (100, 250, 500, 1000, 1500, 2000, 2500 sites) and 100 replicates of 5000 sites using seqboot from the PHYLIP package (Felsenstein 1989). The best tree was obtained for each VLB replicate with RAxML under the LG4X mixture model. Finally, we retrieved the bootstrap

proportion of each bipartition for each matrix length with consense from the PHYLIP package, and further analyzed them using a custom R script.

Model comparison

AIC, AICc and BIC between LG4X and GTR+Γ4 models were computed using ModelFinder (Kalyaanamoorthy et al. 2017) from IQTREE version 1.6 (Nguyen et al. 2015), with the constrained topology previously obtained under the LG4X model with RAxML. Cross-validations between GTR+Γ4 and CAT+Γ4 were carried out using PhyloBayes version 4.1. For both plastid and nuclear datasets, ten training datasets of 10,000 positions were used, and likelihoods were computed on ten test datasets of 2,000 positions.

Acknowledgments

We thank Zehra Çebi for growing up algal strains and for extraction of DNA, and Paul Simion and Rik Verdonck for critical reading of the manuscript. Sequencing was carried out by the Cologne Center for Genomics (CCG). Computations were performed on the supercomputers Mp2 and Ms2 from the Université de Sherbrooke, managed by Calcul Québec and Compute Canada. The operation of this supercomputer is funded by the Canada Foundation for Innovation (CFI), the ministère de l'Économie, de la science et de l'innovation du Québec (MESI), and the Fonds de recherche du Québec - Nature et technologies (FRQ-NT). This work was supported by the TULIP Laboratory of Excellence (ANR-10-LABX-41).

References

Archibald JM. 2015. Endosymbiosis and eukaryotic cell evolution. Curr. Biol. 25:R911–R921.

Baurain D, Brinkmann H, Petersen J, Rodríguez-Ezpeleta N, Stechmann A, Demoulin V, Roger AJ, Burger G, Lang BF, Philippe H. 2010. Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles. Mol. Biol. Evol. 27:1698–1709.

Baurain D, Philippe H. 2010. Current Approaches to Phylogenomic Reconstruction. In: Evolutionary Genomics and Systems Biology. Hoboken, NJ, USA: John Wiley & Sons, Inc. p. 17–41.

Brown JW, Sorhannus U. 2010. A molecular genetic timescale for the diversification of autotrophic stramenopiles (Ochrophyta): Substantive underestimation of putative fossil ages. PLoS One 5:1–11.

Burki F, Shalchian-Tabrizi K, Minge M, Skjæveland Å, Nikolaev SI, Jakobsen KS, Pawlowski J. 2007. Phylogenomics Reshuffles the Eukaryotic Supergroups.Butler G, editor. PLoS One 2:e790.

Delsuc F, Tsagkogeorga G, Lartillot N, Philippe H. 2008. Additional molecular support for the new chordate phylogeny. genesis 46:592–604.

Derelle R, López-García P, Timpano H, Moreira D. 2016. A Phylogenomic Framework to Study the Diversity and Evolution of Stramenopiles (=Heterokonts). Mol. Biol. Evol. 33:2890–2898.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. 26:2460–2461.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16:1–14.

Felsenstein J. 1978. Cases in which Parsimony or Compatibility Methods Will be Positively Misleading. Syst. Zool. 27:401.

Felsenstein J. 1983. Parsimony in Systematics: Biological and Statistical Issues. Annu. Rev. Ecol. Syst. 14:313–333.

Felsenstein J. 1989. PHYLIP - Phylogeny inference package - v3.2. Cladistics.

Gatesy J, Springer MS. 2014. Phylogenetic analysis at deep timescales: Unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. Mol. Phylogenet. Evol. 80:231–266.

Gee H. 2003. Ending incongruence. Nature 425:782–782.

Germot A, Philippe H. 1999. Critical Analysis of Eukaryotic Phylogeny: A Case Study Based on the HSP70 Family. J. Eukaryot. Microbiol. 46:116–124.

Hahn MW, Nakhleh L. 2016. Irrational exuberance for resolved species trees. Evolution (N. Y). 70:7–17.

Han KY, Graf L, Reyes CP, Melkonian B, Andersen RA, Yoon HS, Melkonian M. 2018. A Re-investigation of Sarcinochrysis marina (Sarcinochrysidales, Pelagophyceae) from its Type Locality and the Descriptions of Arachnochrysis, Pelagospilus, Sargassococcus and Sungminbooa genera nov. Protist 169:79–106.

Heled J, Drummond AJ. 2010. Bayesian Inference of Species Trees from Multilocus Data. Mol. Biol. Evol. 27:570–580.

Irisarri I, Baurain D, Brinkmann H, Delsuc F, Sire JY, Kupfer A, Petersen J, Jarek M, Meyer A, Vences M, et al. 2017. Phylotranscriptomic consolidation of the jawed vertebrate timetree. Nat. Ecol. Evol. 1:1370–1378.

Kai A, Yoshii Y, Nakayama T, Inouye I. 2008. Aurearenophyceae classis nova, a New Class of Heterokontophyta Based on a New Marine Unicellular Alga Aurearena cruciata gen. et sp. nov. Inhabiting Sandy Beaches. Protist 159:435–457.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14:587–589.

Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler L a., Armbrust EV, Archibald JM, Bharti AK, Bell CJ, et al. 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. PLoS Biol. 12.

Kolaczkowski B, Thornton JW. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431:980–984.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095–1109.

Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol. Biol. 7:S4.

Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes MPI : Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment. Syst. Biol. 62:611–615.

Le SQ, Dang CC, Gascuel O. 2012. Modeling protein evolution with several amino acid replacement matrices depending on site rates. Mol. Biol. Evol. 29:2921–2936.

Lecointre G, Philippe H, Vân Lê HL, Le Guyader H. 1994. How Many Nucleotides Are Required to Resolve a Phylogenetic Problem? The Use of a New Statistical Method Applicable to Available Sequences. Mol. Phylogenet. Evol. 3:292–309.

Lockhart P, Novis P, Milligan BG, Riden J, Rambaut A, Larkum T. 2006. Heterotachy and Tree Building: A Case Study with Plastids and Eubacteria. Mol. Biol. Evol. 23:40–45.

Maddison WP, Wiens JJ. 1997. Gene Trees in Species Trees.Wiens JJ, editor. Syst. Biol. 46:523–536.

Mirarab S, Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31:i44–i52.

Nadalin F, Vezzi F, Policriti A. 2012. GapFiller: a de novo assembly approach to fill the gap within paired reads. BMC Bioinformatics 13:S8.

Neiman M, Taylor DR. 2009. The causes of mutation accumulation in mitochondrial genomes. Proc. R. Soc. B Biol. Sci. 276:1201–1209.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol. Biol. Evol. 32:268–274.

Parks MB, Wickett NJ, Alverson AJ. 2018. Signal, Uncertainty, and Conflict in Phylogenomic Data for a Diverse Lineage of Microbial Eukaryotes (Diatoms, Bacillariophyta). Mol. Biol. Evol. 35:80–93.

Philippe H, Brinkmann H, Lavrov D V., Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: Why more sequences are not enough. PLoS Biol. 9.

Rodrigue N, Philippe H, Lartillot N. 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. Proc. Natl. Acad. Sci. U. S. A. 107:4629–4634.

Rogers SO, Bendich AJ. 1985. Extraction of DNA from milligram amounts of fresh, herbarium and mummified plant tissues. Plant Mol. Biol. 5:69–76.

Roure B, Rodriguez-Ezpeleta N, Philippe H. 2007. SCaFoS: A tool for selection, concatenation and fusion of sequences for phylogenomics. BMC Evol. Biol. 7:1–12.

Roure B, Philippe H. 2011. Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. BMC Evol. Biol. 11:17.

Ševčíková T, Horák A, Klimeš V, Zbránková V, Demir-Hilton E, Sudek S, Jenkins J, Schmutz J, Pribyl P, Fousek J, et al. 2015. Updating algal evolutionary relationships through plastid genome sequencing: Did alveolate plastids emerge through endosymbiosis of an ochrophyte? Sci. Rep. 5:1–12.

Ševčíková T, Klimeš V, Zbránková V, Strnad H, Hroudová M, Vlček Č, Eliáš M. 2016. A Comparative Analysis of Mitochondrial Genomes in Eustigmatophyte Algae. Genome Biol. Evol. 8:705–722.

Simion P, Philippe H, Baurain D, Jager M, Richter DJDJ, Di Franco A, Roure B, Satoh N, Quéinnec É, Ereskovsky A, et al. 2017. A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. Curr. Biol. 27:958–967.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19:1117–1123.

Springer MS, DeBry RW, Douady C, Amrine HM, Madsen O, de Jong WW, Stanhope MJ. 2001. Mitochondrial versus nuclear gene sequences in deep- level mammalian phylogeny reconstruction. Mol. Biol. Evol.:132–143.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Whitfield JB, Lockhart PJ. 2007. Deciphering ancient rapid radiations. Trends Ecol. Evol. 22:258–265.

Yang EC, Boo GH, Kim HJ, Cho SM, Boo SM, Andersen RA, Yoon HS. 2012. Supermatrix Data Highlight the Phylogenetic Relationships of Photosynthetic Stramenopiles. Protist 163:217–231.

Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19:153.

Zwickl DJ, Hillis DM. 2002. Increased Taxon Sampling Greatly Reduces Phylogenetic Error.Crandall K, editor. Syst. Biol. 51:588–598.

Table 1. Dataset composition summary

Genomic compartment	number of species	number of amino acid positions	number of genes	missing data (%)
mitochondrion	64	6,762	32	5.84
plastid	63	21,692	99	4.12
nucleus	124	209,105	797	25.56

Table 2. Bootstrap support of high-level ochrophyte clades with varying taxon sampling under the LG4X model

	Plastid								Nucleus							
Groupings	All	Out	E	SSC	Ping	PXR	PD	BB	All	Out	E	SSC	Ping	PXR	PD	BB
PD+BB	100	100	98	90	100	99	.	.	68	100	97		100		.	.
E+SSC	100	100	.	.	100	100	100	100			.	.	**56**	**91**		
E+PXR			.			.			81	82	.	96		.	91	100
E+SSC+Ping	95		.	.	.	91	80	66			.	.	.			
E+SSC+Ping+PXR	82		.	.	.	.	77	53	67		.	.	.	.		100
Ping+SSC			*95*	.	.				68		97	.	.			100
Ping+BB					.			.				**95**	.	**88**	**83**	.
PXR+PD+BB		**98**			**56**	.	.	.						.	.	.
Ping+SSC+PXR			*87*	.	.	.					*96*	.	.	.		
Ping+PD+BB					.		.	.				**95**	.	**88**	.	.
E+SSC+PXR			.	.		.					.	.	*83*	.	**83**	
Ping+PXR+PD+BB		**95**			.	.	.	.					.	.	.	.
E+Ping			.	*67*	.						.		.			
E+Ping+PXR			.	*71*	.	.					.		.	.		
E+PXR+PD+BB			.			.	.	.		**78**	.			.	.	.
E+PXR+PD+BB+SSC			.			.	.	.		**100**	.			.	.	.

Rows correspond to the observed high-level groupings and columns to major clades that were left out from the taxon sampling (All means that all species are considered). Dots (.) indicate groupings not testable with the corresponding taxon sampling of the column, italics indicate groupings that are compatible, but not directly comparable, to the grouping formed when all the species are considered, boldface indicates groupings that are not observed when all the species are considered. Abbreviations are as in Fig. 2, and Out means use of a distant outgroup (i.e., removal of the close outgroup).

Table 3. Support of high-level ochrophyte clades with varying taxon sampling under the CAT+Γ4 model

	Plastid (Bootstrap)								Nucleus (Jackknife)							
Groupings	All	Out	E	SSC	Ping	PXR	PD	BB	All	Out	E	SSC	Ping	PXR	PD	BB
PD+BB	93	95	88	90	94	83	.	.	76	78	84	63	98	80	.	.
E+SSC	84	55	.	.	90	100	85	98	40		.	.	68	57	54	46
E+SSC+Ping	58	28	.	.	.	83	59	64			.	.	.			
E+SSC+PXR+ Ping	87	89	.	.	.	.	89	89				.	.	.		**38**
E+PXR			.	**80**		.				**20**	.	*48*		.		
Ping+SSC			*86*	.	.					**30**	**55**	.	.			
PXR+PD+BB						.	.	.			**26**		**44**	.	.	.
Ping+PD+BB					.		.	.	48			59	.	56	.	.
E+SSC+PXR			.	.	*86*	.			20		.	.		.		32
Ping+BB					.			.					.		*60*	.
Ping+SSC+PXR			*96*	.	.	.						.	.	.		
Ping+E+PXR			.	*90*	.	.					.		.	.		
E+PXR+PD+BB			.			.	.	.		**16**	.			.	.	.
PXR+Ping+BB					.	.		.					.	.	**32**	.

See legend of Table 2 for details

Table 4. Jackknife support of high-level ochrophyte clades of the fusion (nu+cp) dataset with varying taxon sampling under the CAT+Γ4 model

	gene jack-knife of 80000 sites							
Bipartitions	All	Out	E	SSC	Ping	PXR	PD	BB
PD+BB	100	100	98	99	100	99	.	.
Ping+PD+BB	100	96	88	99	.	79	.	.
E+SSC	99	99	.	.	100	100	100	98
E+SSC+PXR	90	98	.	.	96	.	96	86
E+PXR			.	*100*		.		
Ping+BB					.		*99*	.
SSC+Ping+PD+BB				.	.		.	.
Ping+PD					.		.	*100*
PXR+SSC			*70*	.		.		

See legend of Table 2 for details

Figure 1

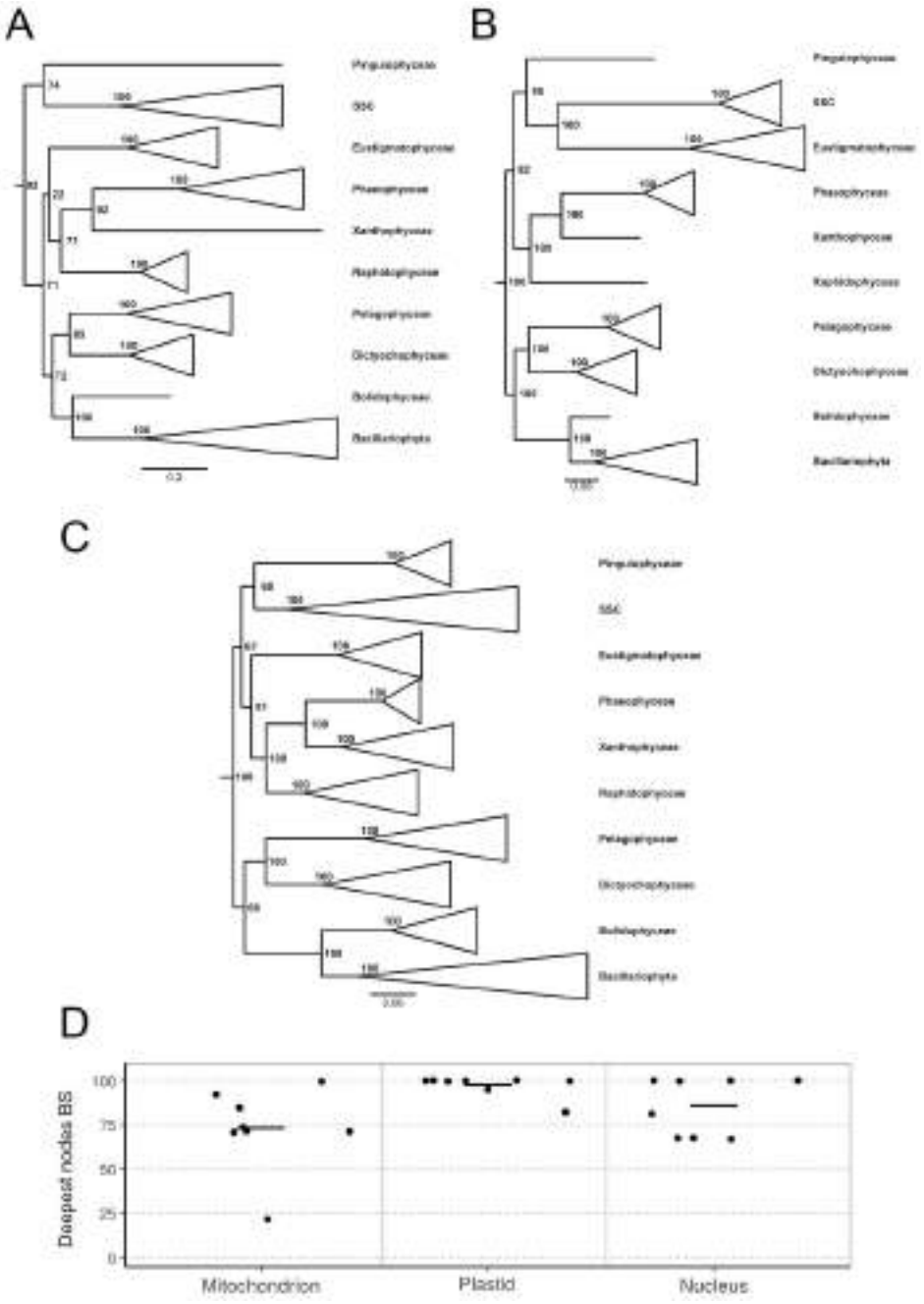


Figure 1

Collapsed trees inferred under the LG4X model from the three different datasets. Support comes from 100 fast bootstrap replicates in RAxML. The ten major clades of Ochrophyta were collapsed when more than two species were present. SSC stands for the monophyly of Synchromophyceae, Synurophyceae and Chrysophyceae, which is only represented by species of Synurophyceae and Chrysophyceae clades in the mitochondrial and plastid datasets A. mitochondrial dataset; B. plastid dataset; C. nuclear dataset; D. Bootstrap support for internal nodes from the three datasets; average values are indicated by a line.

Figure 2

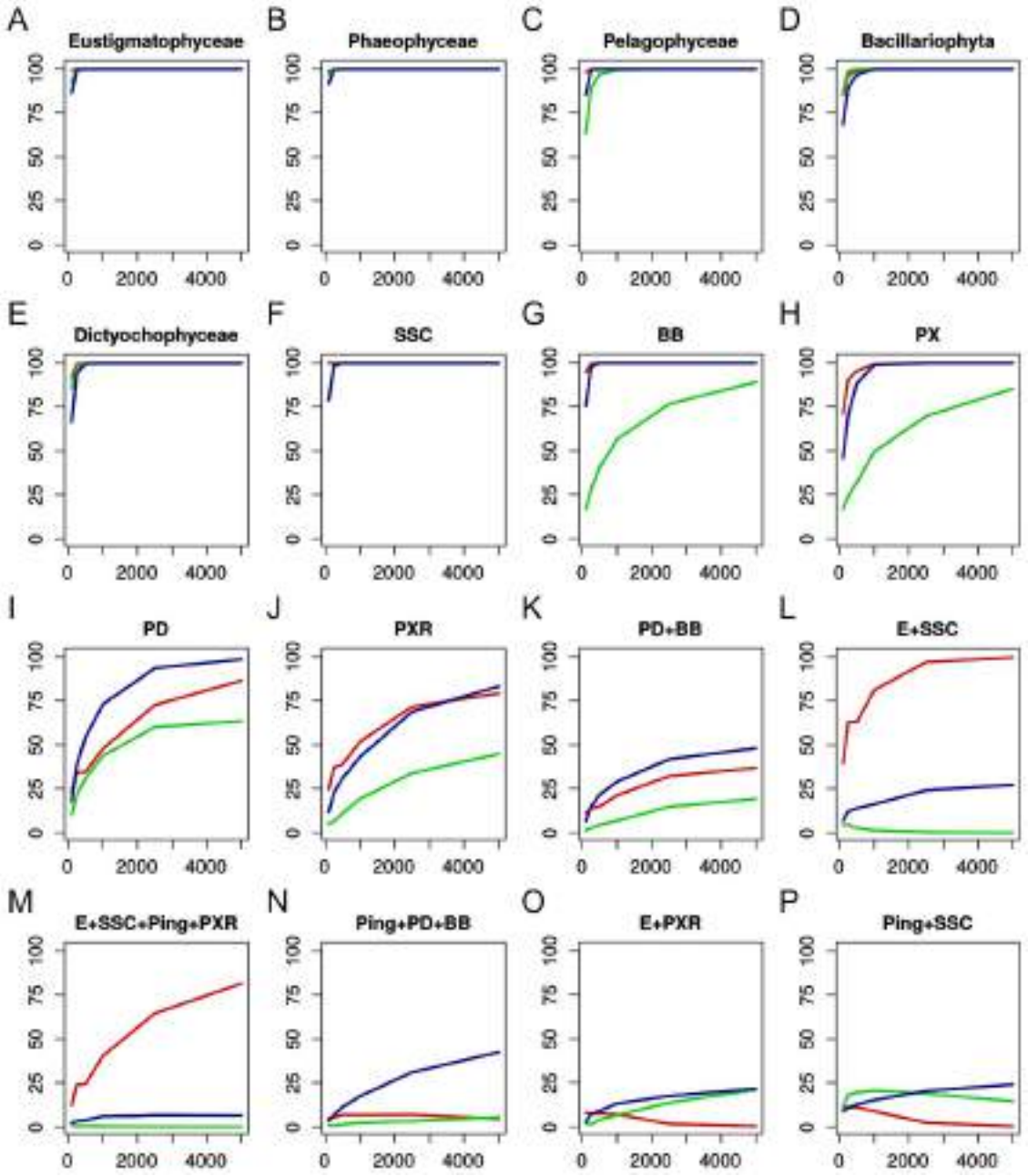


Figure 2

Variable length bootstrap results for a set of groupings in the three compartments under the LG4X model. X-axes represent the number of sites used to infer phylogeny, whereas Y-axes represent the bootstrap support observed for the grouping of interest. Line colors represent the compartments: nucleus (blue), plastid (red) and mitochondrion (green). (SSC) Synchromophyceae, Synurophyceae and Chrysophyceae, (BB) Bolidophyceae and Bacillariophyta, (PX) Phaeophyceae and Xanthophyceae, (PD) Pelagophyceae and Dictyochophyceae, (PXR) Phaeophyceae, Xanthophyceae and Raphidophyceae (E) Eustigmatophyceae, (Ping) Pinguiophyceae.

Figure 3

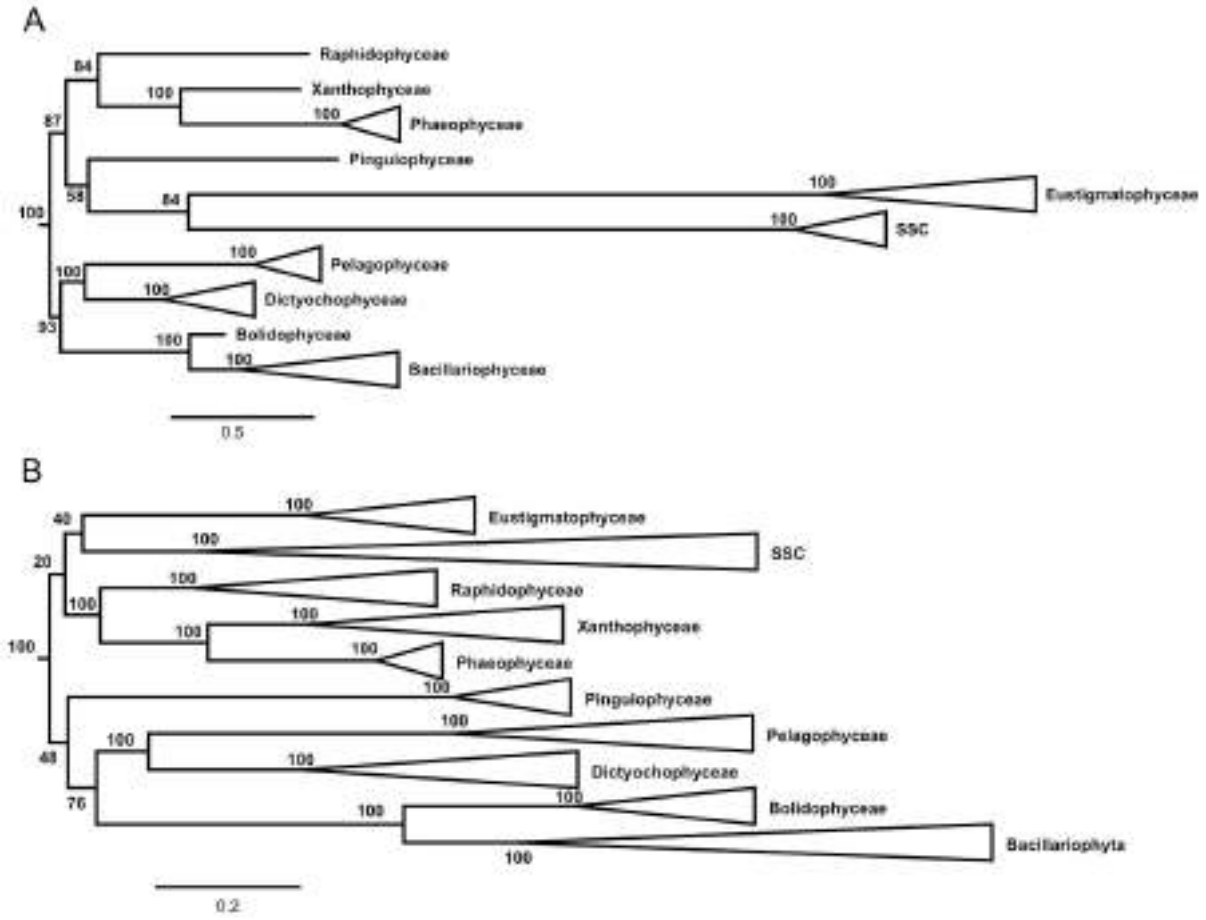


Figure 3

Phylogenetic trees inferred using PhyloBayes-MPI under the CAT+Γ4 model. The ten major clades of Ochrophyta were collapsed when more than two species were present. Statistical support values are displayed next to their relative nodes. (A) Plastid dataset (63 species and 21,692 positions). Statistical support based on 100 non-parametric bootstrap replicates. (B) Nuclear dataset (124 species and 209,105 positions). Statistical support based on 50 gene jackknife replicates of about 20,000 positions.

Figure 4

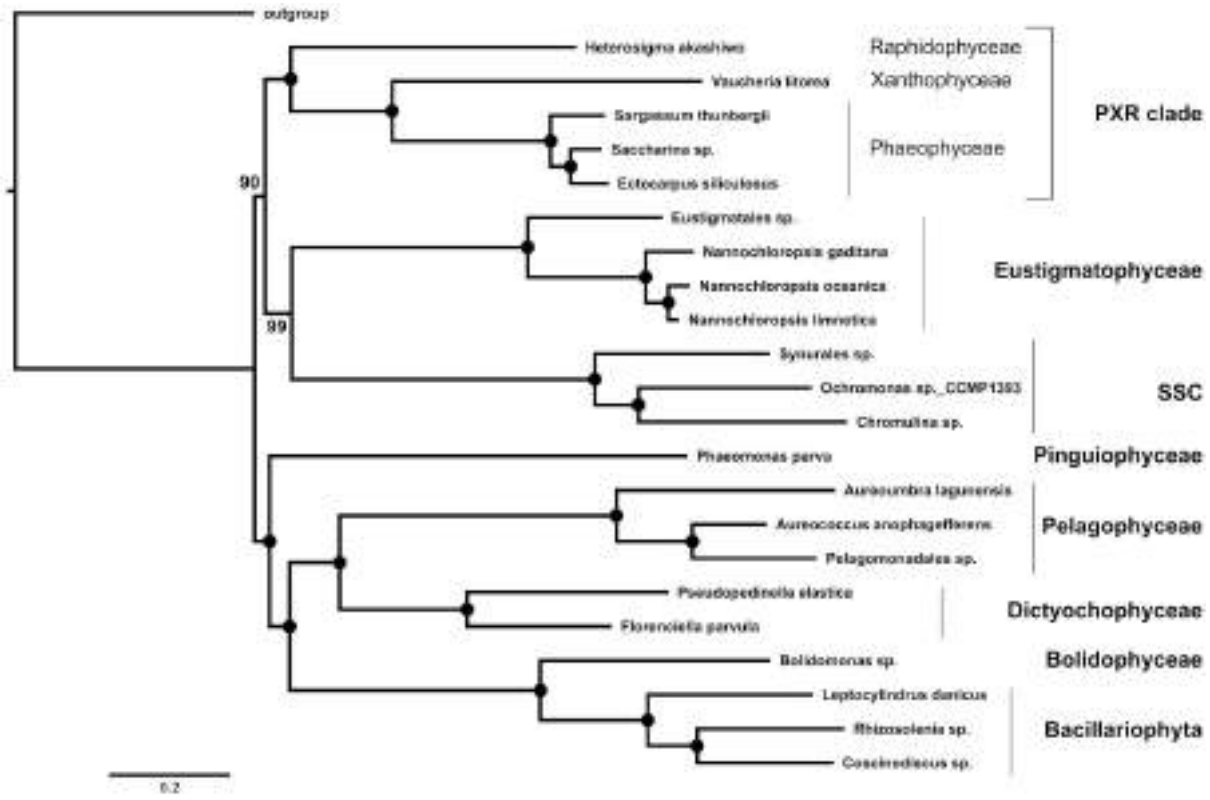


Figure 4

Consensus phylogenetic tree of the fusion (nu+cp) dataset, inferred from 100 jackknife replicates (~80,000 positions) under the CAT+Γ4 model using PhyloBayes-MPI. Statistical support corresponds to jackknife support (JS), with black circles meaning 100% JS. Species named *sp.* correspond to chimeras between the corresponding species of the plastid and nuclear dataset presented in Supplementary Table 5.

Supplementary information

Methods

1. Creation of the plastid dataset

We retrieved the protein annotations for 75 selected plastid genomes of Rhodophyta, Cryptophyta, Haptophyceae and Ochrophyta from the NCBI RefSeq database (https://www.ncbi.nlm.nih.gov/) (Supplementary Table 6). We used OrthoFinder (Emms and Kelly 2015) with a BLASTP E-value threshold of 1e-5 and an MCL inflation parameter of 1.5 to produce orthogroups (OGs). We filtered the 504 resulting OGs to retain those (108) containing ≥20 species (of which ≥1 Rhodophyta, ≥1 Stramenopiles, and either ≥1 Cryptophyta or ≥1 Haptophyceae). We first aligned the selected OGs with MAFFT (L-INS-i algorithm, 5000 iterations) (Katoh and Standley 2013), then enriched them by adding more species from genomic data (such as the five new species sequenced in this study) with Forty-Two (https://bitbucket.org/ dbaurain/42/). We checked for possible paralogy using methods that are described in the section about the construction of the nuclear dataset (see below) and found only one dubious OG, from which we manually removed four paralogous sequences. We further discarded 9 additional OGs with <30 species. Finally, to select unambiguously aligned positions, we applied a loose BMGE (Criscuolo et al. 2010) filter (entropy cutoff of 0.6 and gap cutoff of 0.4) on each aligned OG.

2. Creation of the mitochondrial dataset

As for the plastid, we retrieved all the protein annotations available for stramenopiles mitochondrial genomes from the NCBI website (Supplementary Table 7). To this first set, we added the annotations of the five new species generated in this study, as well as some identified from genomic scaffolds of Labyrinthulomycetes and Xanthophyceae species presenting a high similarity to mitochondrial genomes, using MFannot server (Beck and Lang 2010; Beck N, Lang B (2010) MFannot, organelle genome annotation websever. http://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl). We chose to integrate all annotations before delineating OGs because it can be more difficult to retrieve orthologs for fast evolving mitochondrial sequences with Forty-Two. We generated OGs using the same protocol as with the plastid dataset and retained the 33 OGs containing ≥66% of the species of the dataset. We aligned the clusters with MAFFT (L-INS-i algorithm, 5000 iterations) and manually fixed frameshift errors for some sequences of *Fragilariopsis* and *Aurantiochytrium* in the resulting alignments. We also split the sequence of *Phaeodactylum* fusion protein nad9-rps14 and added each half to its respective alignment. Finally, we removed one OG showing a too ambiguous alignment and applied BMGE on each aligned OG as above.

3. Generation and selection of orthogroups for the nuclear dataset

We retrieved the complete proteomes for 53 species across Stramenopiles, Alveolata and Rhizaria from the NCBI website (Supplementary Table 8). We performed pairwise similarity

searches between all proteomes using USEARCH (Edgar 2010) with a minimal value of 1e-5 (instead of BLAST to speed up computations). We then generated OGs with OrthoFinder, as for the plastid and mitochondrial datasets. Out of a total of 212,849 OGs, we retained only those containing ≥15 species, with at least one species from each of the three clades (Stramenopiles, Alveolata and Rhizaria) and ≤600 sequences to avoid retrieving large protein families. These filters left us with 3,063 OGs.

At this stage, we observed that some OGs contained highly divergent sequences (possibly non-homologous), dragged into the clusters by a single similarity link. To address this issue, we removed the sequences that were not similar to a minimum percentage of the other sequences in each OG (BLAST E-value threshold of 1e-10). We proceeded in two steps, first removing the sequences matching <30% of the other sequences, then 50%. These two steps removed 17,734 and 6,335 sequences, respectively. Finally, we filtered the OGs anew, so as to retain only those with ≥15 species, hence reducing their number to 2,892.

To classify OGs among those close to true orthogroups and those corresponding to more complex multigene families, we used an automated phylogenetic analysis of single-gene trees (Simion et al., 2017). Briefly, we first aligned OGs with MAFFT v7 (L-INS-I algorithm, 5000 iterations) (Katoh and Stanley 2013), then filtered out columns with <5% of amino acid residues and sequences with <50 parsimony informative positions. We then inferred trees with RAxML v8 (Stamatakis 2014) using the LG+F+Γ4 model (Le and Gascuel 2008), and ran a custom script aimed at detecting cases of ancient paralogy. As in Simion et al. (2017), we computed how many of seven predefined clades (Rhizaria, Ciliophora, Myzozoa, Oomycetes, Labyrinthulomycetes, *Blastocystis* and Ochrophyta) were affected by out-paralogy (i.e., at least two clans containing sequences exclusively from a given clade). This allowed us to separate the OGs containing ≤3 out-paralogs (1904) from the other, more complex OGs (988) having more out-paralogs.

To split the OGs showing too many cases of out-paralogy (which may correspond to an ancient gene duplication), we used the software root-max-div (Simion et al. 2017). This program searches for the branch maximizing the taxonomic diversity on both sides of the bipartition and splits the tree on the branch if (i) the number of sequences on each side satisfies a minimal threshold (two first parameters), (ii) the number of common species on both sides is above a minimal threshold (third parameter) and (iii) the branch length is among the top percentile of the tree branches (fourth parameter). We applied four different parameter sets in the following order (30-30-0-5, 30-10-0-5, 10-10-0-5, and 40-10-0-20), retrieving the two sub-alignment files of the first successful parameter set for each OG. We repeated the whole procedure until no more gene tree could be cut. Finally, we filtered the split OGs again, to retain only those with ≥15 species, thereby reducing their number to 336.

Finally, we pooled the two groups of OGs, yielding a total of 2,240 OGs, of which we reduced the redundancy with a custom script targeting highly similar subsequences of the same organism inside individual OGs, as in Simion et al. (2017). This step removed 5,573 sequences.

4. Assessment of transcriptomes quality

Before improving the taxon sampling of our nucleus dataset using a combination of transcriptomic and genomic data, we evaluated the contamination level of the available transcriptomes. They consisted in assemblies from MMETSP (Keeling 2014), TSA retrieved from the NCBI website, and SRA raw reads also retrieved from the NCBI website, which were assembled using Trinity v2.6 (Haas 2013) with the trimmomatic and jaccard clip options (--jaccard_clip --trimmomatic). A set of 80 highly expressed gene alignments (ribosomal proteins), on which a large diversity of eukaryotic sequences are regularly added and manually curated, was used as a reference. To estimate the contamination level, we took advantage of Forty-Two and its taxonomic filters to add back previously incorporated organisms to this dataset. Forty-Two was designed to search transcriptomes for orthologous sequences and add them to existing alignments. At this last step, it can verify if the added orthologous sequence satisfies a user-defined positive and/or negative taxonomic filter (i.e., belonging to Stramenopiles or not belonging to Xanthophyceae). Here, we checked that added sequences indeed matched an organism of the same genus in the alignment, which allowed us to distinguish between orthologous sequences genuinely belonging to the transcriptome from orthologous sequences belonging to contaminants. We considered a transcriptome to be clean when we found <5 contaminant sequences over the 80 alignments. For each contaminant sequence, we further retrieved the most closely related organism in the alignment, so as to design optimal taxonomic filters for contaminated transcriptomes. Overall, this approach allowed us to exploit taxonomically interesting transcriptomes that were contaminated without adding contaminated sequences in our OGs.

5. *Vaucheria litorea* transcriptome decontamination

Whereas *Vaucheria litorea* was one of the only Xanthophyceae for which a large amount of data was available, we observed that its transcriptome was contaminated by a large array of organisms. Because of its isolated phylogenetic position in our current sample of the eukaryotic diversity, combined to a relatively fast evolutionary rate, it was difficult to only rely on sequence similarity for decontamination. Thus, to tackle this issue, we implemented a strategy based on k-mer distributions (Teeling et al. 2004) to identify and remove the largest part of *Vaucheria* contaminant sequences. Briefly, we assembled two sets of *Vaucheria* sequences (i.e., genuine and contaminant), for which we computed the frequencies of all possible 6-nt k-mers. Then, the k-mer composition of each transcript was compared against those reference distributions using an Euclidean distance, and we discarded the transcripts closer to the contaminant than the genuine sequences. To define these two sets of reference, we used eukaryotic orthologs (594 nuclear genes from 370 species covering the diversity of eukaryotes) from a non-published study, in which we added *Vaucheria* transcripts with Forty-Two. (We used these orthologs instead of the 80 ribosomal proteins to maximize the number and the variety of sequences in the reference sets.) Then, we inferred the single-gene phylogeny of all orthologs using RAxML (LG+F+Γ4 model) and retrieved the taxonomy of the sister clan of each *Vaucheria* sequence. We considered transcripts added close to Ochrophyta species as genuine reference sequences, whereas the other transcripts were pooled separately as different sources of contaminants. Those were mainly

representatives of Labyrinthulomycetes, Discosea and Viridiplantae. To identify contaminants in the full *Vaucheria* transcriptome, we first tested the transcripts against Labyrinthulomycetes contaminated sequences then to Discosea sequences, and finally to Viridiplantae sequences. We verified the effectiveness of our approach with the protocol described in section 4, which confirmed that the majority of the contaminants had been properly removed after using the Labyrinthulomycetes and Discosea references. Thus, we used the resulting transcriptome in our dataset construction without excluding the sequences that would have been removed after testing against Viridiplantae.

6. Enrichment of OGs to increase taxonomic diversity

Starting from OGs generated at section 3, we improved our taxonomic sampling by adding sequences from genomic and transcriptomic data with Forty-Two. We worked in two consecutive steps, first adding non-contaminated transcriptomic data (127 species) (see section 4. Assessment of transcriptomes quality) and genomic data (13 species), and then contaminated transcriptomic data (54 species). As explained in section 4, we took advantage of the presence of contaminants relatives or sequenced organisms relatives in our OGs to design custom taxonomic filters for the added sequences (see YAML files in Supplementary Archive). After each run of Forty-Two, we ran the custom script described at the end of section 3 to reduce the number of redundant sequences per species. Finally, we filtered the enriched OGs to retain only those with ≥5 species of Rhizaria, ≥10 species of Alveolata, ≥10 non-photosynthetic Stramenopiles and ≥20 photosynthetic Stramenopiles (Ochrophyta), leaving us with 1330 enriched OGs and a total of 244 species.

7. Targeted decontamination

Some transcriptomes showed evidence of contamination by organisms outside of SAR, which cannot be handled by the taxonomic filters of Forty-Two (due to the lack of related sequences in our OGs). To remove these contaminant sequences from our OGs, we used a custom script to BLAST each sequence against two reference databases of wanted (i.e., SAR species) and unwanted (i.e., contaminant species) proteomes. We then discarded the sequences better matching the unwanted database over the wanted database. Contamination sources detected at this step were as various as red algae (70 sequences), green algae (302 sequences), green plants (297 sequences), animals (247 sequences), fungi (12 sequences) or alpha-proteobacteria (3481 sequences, probably containing some genuine sequences of mitochondrial origin).

8. Detection and elimination of remaining paralogs

After targeted removal of contaminant sequences, we realigned each OG with MAFFT and searched for possibly remaining paralogs that could not have been handled by our tree splitting step and/or that appeared during OG enrichment. We inferred single-gene trees as before (RAxML, LG+F+Γ4 model) and used them to split alignments with the previously described script, but using new values of the four parameters (50-50-50-10). We also split trees for recent paralogs specific to one clade (at least 150 species on one side, 5 on the

other side, 3 in common, and branch length among the top-10%, 150-5-3-10), discarding the smallest sub-alignment file. This step mainly allowed us to detect and remove potential nucleomorph sequences. The latter sequences were more specifically targeted by isolating Rhizaria sequences, splitting them with a 10-1-1-5 parameter set and discarding the smallest group of rhizarian sequences. Finally, we reassessed the presence of ancient paralogs between the 7 clades, as described in section 3 but with a small modification. We inferred single gene trees without the sequences with <100 parsimony informative positions, as misplacement of these fragments could inflate the separation of monophyletic clades (Di Franco et al. 2019). We kept the clusters with less than 16 out-paralogs and filtered them for ≥93 species, leaving us with a total of 1,119 alignments of orthologous sequences.

9. Branch length decontamination and filtering

To remove remaining outlier sequences of various origins (contaminants, paralogs or xenologs), we used the same protocol as in Simion et al. (2017). Instead of using a block-oriented BMGE filter on our OGs as above, we applied HmmCleaner v1.8 (Di Franco et al. 2019) with default parameters and filtered columns with <5% of amino acid residues, in order to keep as much signal as possible for inferring single-gene trees. We created a supermatrix with those OGs using SCaFoS (Roure et al. 2007), picking the longest sequence per OTU. On one side, we inferred the supermatrix tree with RAxML and LG+F+Γ4 model to serve as the reference tree. On the other side, we inferred the branch lengths of each individual OG while constraining the tree topology to the supermatrix topology. Finally, we compared the terminal branch lengths of single-gene trees to those of the supermatrix tree for each OTU. We performed this procedure twice, removing sequences that were x times longer in a single-gene tree compared to the supermatrix tree. We first removed 423 sequences using a branch length ratio of x=7 and then 465 sequences using a ratio of x=5. Moreover, we discarded 53 OGs in which too many branch lengths were incoherent with those of the supermatrix tree. To identify these OGs, we computed a branch length R^2 for each OG with respect to the supermatrix tree, and set the elimination threshold to a R^2 value below the mean value of the R^2 distribution minus 1.96 times the standard deviation, assuming a normal distribution of R^2 values.

In the last step of our nuclear dataset construction, we reduced our taxon sampling to 185 SAR OTUs by either removing too closely related species or building chimeras with SCaFoS, depending on their completeness. We applied BMGE as above, followed by HmmCleaner with default parameters on each OG. We retained the 797 OGs having ≤62 missing OTUs using SCaFoS, and subsequently removed Rhizaria and Alveolata.

10. DNA extraction via CTAB method reagents (modified protocol)

reagents

- TE buffer: 10 mM Tris-HCl (pH 8), 1mM EDTA, sterile bidistilled water
- 5% SDS solution
- proteinase K solution: 10 mg/ ml
- RNAse: 10 mg/ml
- 5M NaCl

- CTAB buffer (pre-warmed at 65°C): 1.4 M NaCl, 0.1 M Tris-HCl (pH 8), 25 mM EDTA, 2% CTAB (w/v), sterile bidistilled water
- 24:1 chloroform : isoamyl alcohol
- Isopropanol/2-propanol (pre-cooled at -20°C)
- 80% ethanol (pre-cooled at -20°C)

protocol

1. harvest algae & wash pellet several times with respective medium (centrifugation at lowest-possible speed)
2. homogenize pellet (≈ 1 ml) with liquid nitrogen
3. transfer homogenized powder very quickly into a sterile falcon tube containing a mixture of 900 µl TE buffer, 700 µl 5% SDS and 25 µl proteinase K solution & vortex immediately
4. incubate in water bath at 60°C, 20‘ (vortex occasionally)
5. add 500 µl 5 M NaCl, 25 µl RNAse A and 5 ml CTAB buffer (pre-warmed at 65°C) & vortex
6. incubate in water bath at 60°C, 10‘ (vortex occasionally)
7. centrifuge the sample (3,000 g, 15‘) & transfer supernatant into a new sterile falcon tube
8. add equal volume of 24:1 chloroform : isoamyl alcohol & vortex rigorously
9. incubate at RT, 10’
10. separate phases by centrifugation (3,000 g, 15’)
11. transfer & portion the upper, aqueous phase carefully into sterile 1.5 ml eppendorf tubes (if the aqueous phase is not clear, the chloroform-isoamyl alcohol extraction has to be repeated from step 8)
12. maintain DNA extracts on ice & perpetuate this condition
13. add 2/3 volume of isopropanol (pre-cooled at -20°C) & vortex
14. incubate at -20°C, at least 1 h (or overnight)
15. centrifuge (17,000 g, 15’, 4°C) & discard the supernatant carefully
16. wash pellet with 1 ml 80% ethanol (pre-cooled at -20°C) & vortex
17. centrifuge (17,000 g, 5’, 4°C) & discard the supernatant carefully
18. repeat the washing step
19. air-dry the pellet & dissolve the pellet in 50 - 200 µl TE buffer
20. store at -20°C

References

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.

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.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. 26:2460–2461.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16:1–14.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8:1494–1512.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780.

Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler L a., Armbrust EV, Archibald JM, Bharti AK, Bell CJ, et al. 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. PLoS Biol. 12.

Le SQ, Gascuel O. 2008. An Improved General Amino Acid Replacement Matrix. Mol. Biol. Evol. 25:1307–1320.

Roure B, Rodriguez-Ezpeleta N, Philippe H. 2007. SCaFoS: A tool for selection, concatenation and fusion of sequences for phylogenomics. BMC Evol. Biol. 7:1–12.

Simion P, Philippe H, Baurain D, Jager M, Richter DJDJ, Di Franco A, Roure B, Satoh N, Quéinnec É, Ereskovsky A, et al. 2017. A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. Curr. Biol. 27:958–967.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Teeling H, Meyerdierks A, Bauer M, Amann R, Glockner FO. 2004. Application of tetranucleotide frequencies for the assignment of genomic fragments. Environ. Microbiol. 6:938–947.

Supplementary Table 1. Bootstrap support of high-level ochrophyte clades for 23 species datasets with varying taxon sampling under LG4X model

	Plastid								Nucleus							
Bipartitions	All	Out	E	SSC	Ping	PXR	PD	BB	All	Out	E	SSC	Ping	PXR	PD	BB
PD+BB	54	96	61	46	64	88	.	.	51	81	70		100	59	.	.
Ping+PD+BB					.		.	.				**98**	.	**81**	.	.
E+SSC	100	100	.	.	100	100	100	100		71	.	.		**60**	**47**	
E+PXR			.			.					.		**59**	.		
Ping+SSC			100	.	.				35		51	.	.			44
Ping+BB					.			.				**67**	.		**96**	.
PXR+PD+BB		**80**				.	.	.						.	.	.
E+SSC+PXR			.	.	99	.				71	.	.		.	**44**	
PXR+Ping+PD+BB		**54**			.	.	.	.				*97*	.	.	.	.
E+PXR+PD+BB			.			.	.	.			.		**60**	.	.	.
SSC+Ping+PD+BB				.	.		.	.	36		51	.	.		.	.
Ping+PD+SSC				.	.		.					.	.		.	47
Ping+PD+BB+ SSC+PXR				.	.	.	.	.	33			.	.	.	.	.
Ping+PD+PXR+ SSC				.	.	.	.					.	.	.	.	*47*
Ping+E+SSC	93		.	.	.	100	68	84			.	.	.			
PXR+Ping+E+SSC	99		.	.	.	.	100	100			.	.	.	.		
Ping+SSC+PXR			98	.	.	.						.	.	.		
Ping+E			.	78	.						.		.			
Ping+E+PXR			.	97	.	.					.		.	.		
E+SSC+PXR+ PD+BB			.	.		.	.	.		76	.	.		.	.	.

Rows correspond to the observed high-level groupings and columns to major clades that were left out from the taxon sampling (All means that all species are considered). Dots (.) indicate groupings not testable with the corresponding taxon sampling of the column, italics indicate groupings that are compatible, but not directly comparable, to the grouping formed when all the species are considered, boldface indicates groupings that are not observed when all the species are considered. Abbreviations are as in Fig. 2, and Out means use of a distant outgroup (i.e., removal of the close outgroup).

Supplementary Table 2. Comparison of evolutionary models

Compartment	Model	Likelihood	Degree of freedom	AIC	AICc	BIC
Plastid	LG4X	817,358.41	129	1,634,974.82	1,634,976.38	1,636,004.85
Plastid	GTR+Γ4	807,924.17	332	1,616,512.34	1,616,522.69	1,619,163.26
Nucleus	LG4X	13,173,397.30	270	26,347,334.60	26,347,335.30	26,350,102.26
Nucleus	GTR+Γ4	13,017,108.02	454	26,035,124.05	26,035,126.03	26,039,777.82

Supplementary Table 3. Jackknife support of high-level ochrophyte clades for two different replicate sizes of the 23-species nuclear dataset with varying taxon sampling under CAT+Γ4 model

	~20,000 positions								~80,000 positions							
Bipartitions	All	Out	E	SSC	Ping	PXR	PD	BB	All	Out	E	SSC	Ping	PXR	PD	BB
PD+BB	85	79	81	81	97	79	.	.	98	100	100	99	100	100	.	.
Ping+PD+BB	72	53	52	83	.	77	.	.	99	98	80	100	.	98	.	.
E+SSC	40	52	.	.	58	67	48		84	87	.	.	89	97	90	78
E+PXR			.	**54**		.		**29**			.	**79**		.		
Ping+BB					.		*71*	.					.		*94*	.
PXR+PD+BB					*35*	.	.	.					*49*	.	.	.
PXR+Ping+PD+BB	24				.	.	.	.	45				.	.	.	.
SSC+Ping+PD+BB			**54**	.	.		.	.			**81**	.	.		.	.
Ping+PD					.		.	*59*					.		.	*96*
PXR+Ping+BB					.	.	*24*	.					.	.	*55*	.
E+SSC+Ping+PD			.	.	.		.				.	.	.		.	**42**
Ping+PD+SSC				.	.		.	**39**				.	.		.	
E+SSC+PXR		**46**	.	.		.				**71**	.	.		.		

See legend of Supplementary Table 1 for details

Supplementary Table 4. Statistical support of high-level clades for the 23-species fusion (nu+cp) dataset with varying taxon sampling under LG4X and CAT+Γ4 models

	LG4X (bootstrap)							CAT+Γ4 (jackknife 80,000)						
Groupings	All	E	SSC	Ping	PXR	PD	BB	All	E	SSC	Ping	PXR	PD	BB
PD+BB		**86**	**59**	*100*	**74**	.	.	100	98	99	100	99	.	.
Ping+PD+BB	65		85	.	99	.	.	100	88	99	.	79	.	.
E+SSC	100	.	.	100	100	100	100	99	.	.	100	100	100	98
E+SSC+PXR	86	.	.	94	.	100	67	90	.	.	96	.	96	86
E+PXR		.			.				.	*100*		.		
Ping+SSC		**64**	.	.						.	.			
Ping+BB	46			.		100	.				.		*99*	.
PXR+Ping+PD+BB			**64**	.	.	.	.				.	.	.	.
SSC+Ping+PD+BB		**63**	.	.		.	.			.	.		.	.
Ping+PD				.		.	*100*				.		.	*100*
PXR+SSC			.		.				*70*	.		.		

See legend of Supplementary Table 1 for details

Supplementary Table 5. Species matches between compartments for common taxon sampling analyses

Mitochondrion	Plastid	Nucleus
Phytophthora sojae	*Guillardia theta*	*Phytophthora parasitica*
Aureococcus anophagefferens	*Aureococcus anophagefferens*	*Aureococcus anophagefferens*
Sarcinochrysidales sp._CCMP2135	*Aureoumbra lagunensis*	*Aureoumbra lagunensis_CCMP1510*
Chromulina chionophila	*Chromulina chionophila_CCAP9099*	*Chromulina nebulosa_UTEXLB2642*
Berkeleya fennica	*Coscinodiscus radiatus*	*Coscinodiscus wailesii_CCMP2513*
Ectocarpus siliculosus	*Ectocarpus siliculosus*	*Ectocarpus siliculosus*
Florenciella parvula	*Florenciella parvula_RCC446*	*Florenciella parvula_CCMP2471*
Heterosigma akashiwo	*Heterosigma akashiwo*	*Heterosigma akashiwo_CCMP2393*
Nannochloropsis gaditana	*Nannochloropsis gaditana*	*Nannochloropsis gaditana*
Nannochloropsis limnetica	*Nannochloropsis limnetica*	*Nannochloropsis limnetica*
Nannochloropsis oceanica	*Nannochloropsis oceanica*	*Nannochloropsis oceanica*
Ochromonas danica	*Ochromonas sp._CCMP1393*	*Ochromonas sp._CCMP1393*
Asterionella formosa	*Leptocylindrus danicus*	*Leptocylindrus danicus_B650*
Phaeomonas parva	*Phaeomonas parva_CCMP2877*	*Phaeomonas parva_CCMP2877*
Pseudopedinella elastica	*Pseudopedinella elastica_SAGB4388*	*Pseudopedinella elastica_CCMP716*
Phaeodactylum tricornutum	*Rhizosolenia imbricata*	*Rhizosolenia setigera_CCMP1694*
Saccharina japonica	*Saccharina japonica*	*Saccharina angustata*
Sargassum thunbergii	*Sargassum thunbergii*	*Sargassum thunbergii*
Synura peterssenii	*Synura peterssenii_CCAC0052*	*Mallomonas sp._CCMP3275*
Trachydiscus minutus	*Trachydiscus minutus*	*Eustigmatos cf._polyphem*
Triparma laevis	*Triparma laevis*	*Bolidomonas pacifica_CCMP1866*
Heterococcus sp._DN1	*Vaucheria litorea*	*Vaucheria litorea*
-	uncultured *Pelagomonas*	*Pelagomonas calceolata_CCMP1756*

Supplementary Table 6. Genomes used to construct the plastid dataset

Species
Ahnfeltia plicata
Apophlaea sinclairii
Asparagopsis taxiformis
Asterionella formosa
Asterionellopsis glacialis
Aureococcus anophagefferens
Aureoumbra lagunensis
Bangiopsis subsimplex
Calliarthron tuberculosum
Ceramium cimbricum
Ceramium japonicum
Cerataulina daemon
Chaetoceros simplex
Chondrus crispus
Choreocolax polysiphoniae
Coccophora langsdorfii
Coeloseira compressa
Coscinodiscus radiatus
Costaria costata
Cryptomonas paramecium
Cyanidioschyzon merolae_strain_10D
Cyanidium caldarium
Cylindrotheca closterium
Dasya binghamiae
Didymosphenia geminata
Durinskia baltica
Ectocarpus siliculosus
Emiliania huxleyi
Erythrotrichia carnea
Eunotia naegelii
Fucus vesiculosus
Galdieria sulphuraria
Gelidium elegans
Gelidium vagum
Gracilaria chilensis
Gracilaria chorda

Gracilaria firma
Gracilaria tenuistipitata_var._liui
Grateloupia taiwanensis
Guillardia theta
Heterosigma akashiwo
Hildenbrandia rivularis
Hildenbrandia rubra
Kumanoa americana
Leptocylindrus danicus
Lithodesmium undulatum
Mastocarpus papillatus
Membranoptera platyphylla
Membranoptera tenuis
Membranoptera weeksiae
Nannochloropsis gaditana
Nannochloropsis granulata
Nannochloropsis limnetica
Nannochloropsis oceanica
Nannochloropsis oculata
Nannochloropsis salina
Nitzschia sp._Irils04
Odontella sinensis
Palmaria palmata
Pavlova lutheri
Phaeocystis antarctica
Phaeocystis globosa
Phaeodactylum tricornutum
Pleurocladia lacustris
Plocamium cartilagineum
Porphyridium purpureum
Porphyridium sordidum
Pseudo-nitzschia multiseries
Pyropia haitanensis
Pyropia perforata
Pyropia pulchra
Rhizosolenia imbricata
Rhodochaete parvula
Rhodomonas salina
Rhodymenia pseudopalmata
Roundia cardiophora
Saccharina japonica
Sargassum thunbergii

Schimmelmannia schousboei
Schizymenia dubyi
Sebdenia flabellata
Sporolithon durum
Teleaulax amphioxeia
Thalassiosira oceanica_CCMP1005
Thalassiosira pseudonana
Thalassiosira weissflogii
Thorea hispida
Toxarium undulatum
Trachydiscus minutus
Triparma laevis
Ulnaria acus
Undaria pinnatifida
Vaucheria litorea
Vertebrata lanosa
Wildemania schizophylla

Supplementary Table 7. Genomes used to construct the mitochondrial dataset

Species
Achlya hypogyna
Aphanomyces astaci
Aphanomyces invadans
Asterionella formosa
Aurantiochytrium sp._T66
Aureococcus anophagefferens
Berkeleya fennica
Cafeteria roenbergensis
Chattonella marina
Chromulina chionophila
Colpomenia peregrina
Costaria costata
Desmarestia viridis
Dictyota dichotoma
Didymosphenia geminata
Ectocarpus siliculosus
Fistulifera solaris
Florenciella parvula
Fragilariopsis cylindrus_CCMP1102
Fucus vesiculosus
Heterococcus sp._DN1
Heterosigma akashiwo
Laminaria digitata
Laminaria hyperborea
Monodopsis sp._MarTras21
Nannochloropsis gaditana
Nannochloropsis oceanica
Navicula ramosissima
Ochromonas danica
Peronospora tabacina
Petalonia fascia
Phaeodactylum tricornutum
Phaeomonas parva
Phytophthora infestans
Phytophthora sojae
Pleurocladia lacustris

Proteromonas lacertae
Pseudo-nitzschia multiseries
Pseudopedinella elastica
Pseudoperonospora cubensis
Pylaiella littoralis
Pythium insidiosum
Pythium ultimum
Saccharina angustata
Saccharina japonica
Saprolegnia ferax
Saprolegnia parasitica_CBS_223.65
Sargassum muticum
Sargassum thunbergii
Sargassum vachellianum
Schizochytrium sp.
Scytosiphon lomentaria
Skeletonema marinoi
Synura peterssenii
Synura synuroidea
Thalassiosira pseudonana
Thraustotheca clavata
Trachydiscus minutus
Triparma laevis
Turbinaria ornata
Ulnaria acus
Undaria pinnatifida
Vischeria sp._CAUP_Q_202

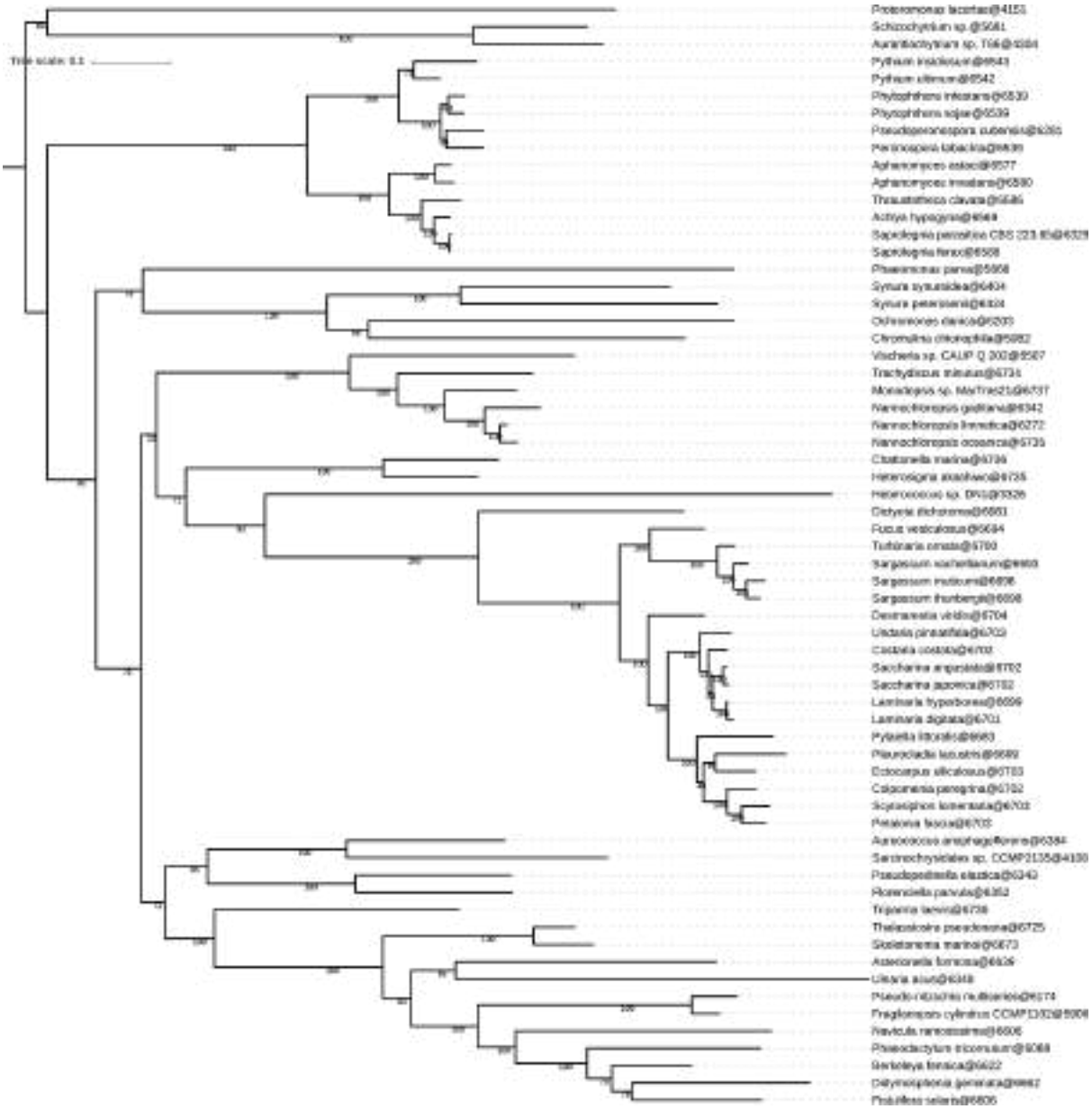
Supplementary Table 8. Genomes used to construct the nuclear dataset

Species
Albugo candida
Aphanomyces astaci
Aphanomyces invadans
Aplanochytrium kerguelense
Aurantiochytrium limacinum
Aureococcus anophagefferens
Babesia bigemina
Babesia bovis
Babesia microti_strain_RI
Bigelowiella natans
Blastocystis hominis
Blastocystis sp._NandII
Blastocystis sp._subtype_4
Cryptosporidium muris_RN66
Cryptosporidium parvum_Iowa_II
Cyclospora cayetanensis
Ectocarpus siliculosus
Eimeria acervulina
Eimeria tenella
Fragilariopsis cylindrus
Gregarina niphandrodes
Hammondia hammondi
Ichthyophthirius multifiliis
Nannochloropsis gaditana
Neospora caninum_Liverpool
Oxytricha trifallax
Paramecium tetraurelia
Perkinsus marinus_ATCC_50983
Phaeodactylum tricornutum_CCAP_1055/1
Phytophthora parasitica

Phytophthora sojae
Plasmodiophora brassicae
Plasmodium gaboni
Plasmodium ovale_curtisi
Plasmodium reichenowi
Plasmodium vinckei_petteri
Plasmodium vivax
Plasmopara halstedii
Pseudocohnilembus persalinus
Pseudo-nitzschia multiseries
Reticulomyxa filosa
Saprolegnia diclina_VS20
Saprolegnia parasitica_CBS_223.65
Schizochytrium aggregatum
Stylonychia lemnae
Tetrahymena thermophila_SB210
Thalassiosira oceanica
Thalassiosira pseudonana_CCMP1335
Theileria equi_strain_WA
Theileria orientalis_strain_Shintoku
Theileria parva
Toxoplasma gondii_ME49
Vitrella brassicaformis_CCMP3155

Suppl. Figure 1

https://itol.embl.de/tree/19350761125837153960072O#

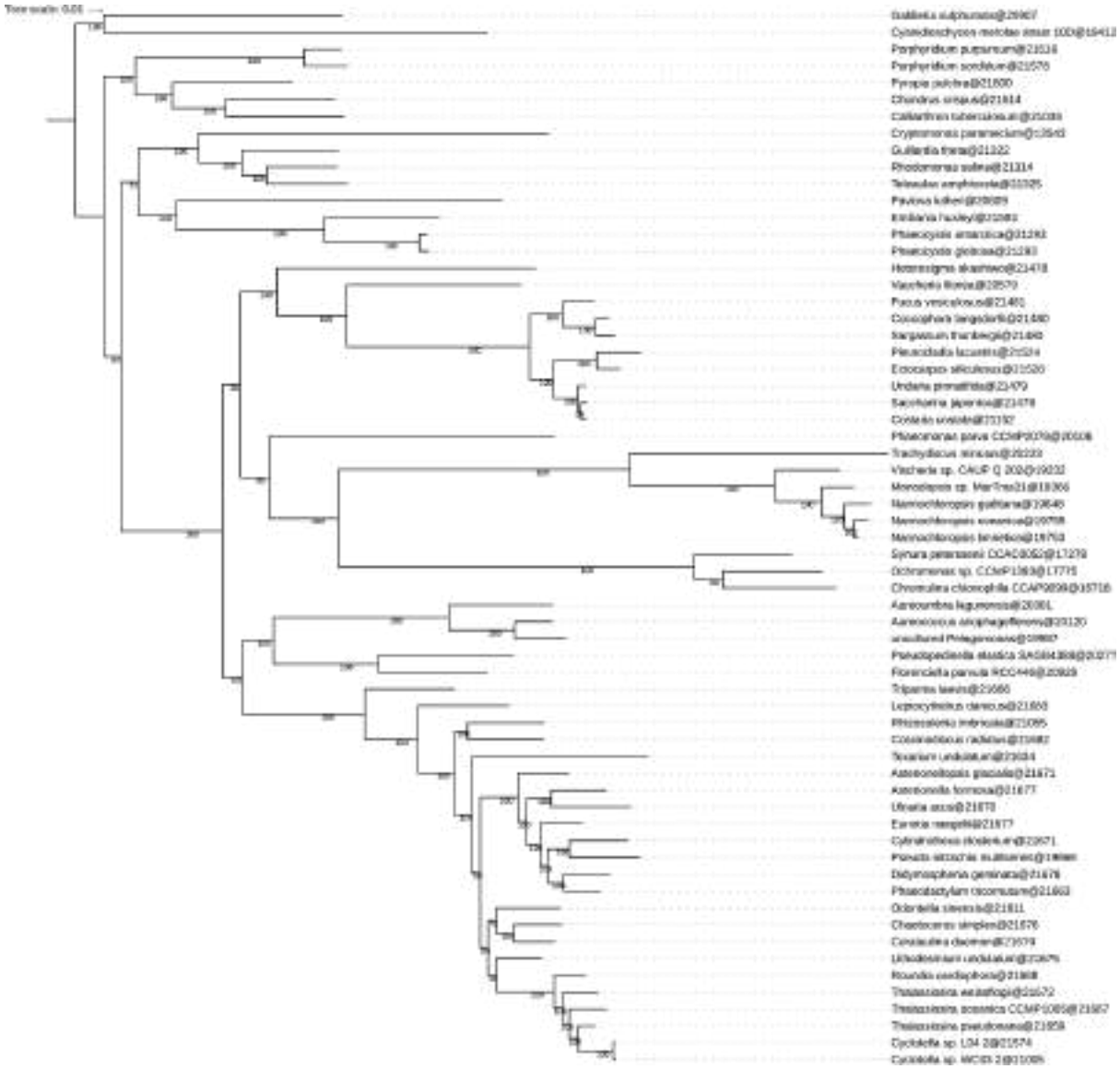


Suppl. Figure 1 Legend

Complete taxon sampling of the mitochondrial phylogenetic tree inferred using RAxML with LG4X.

Suppl. Figure 2

https://itol.embl.de/tree/19350761125757153960071 2#

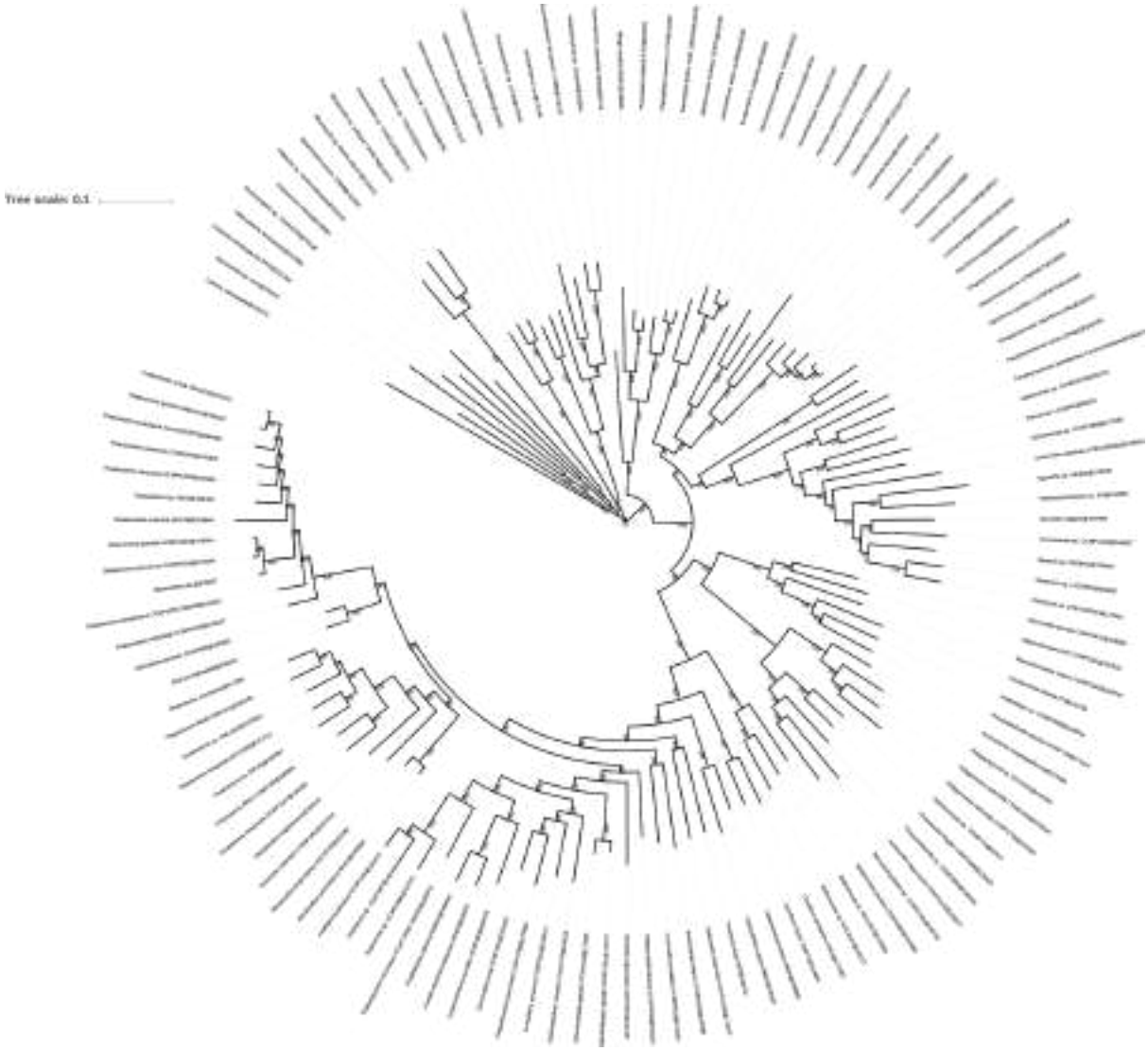


Suppl. Figure 2 Legend

Complete taxon sampling of the plastid phylogenetic tree inferred using RAxML with LG4X.

Suppl. Figure 3

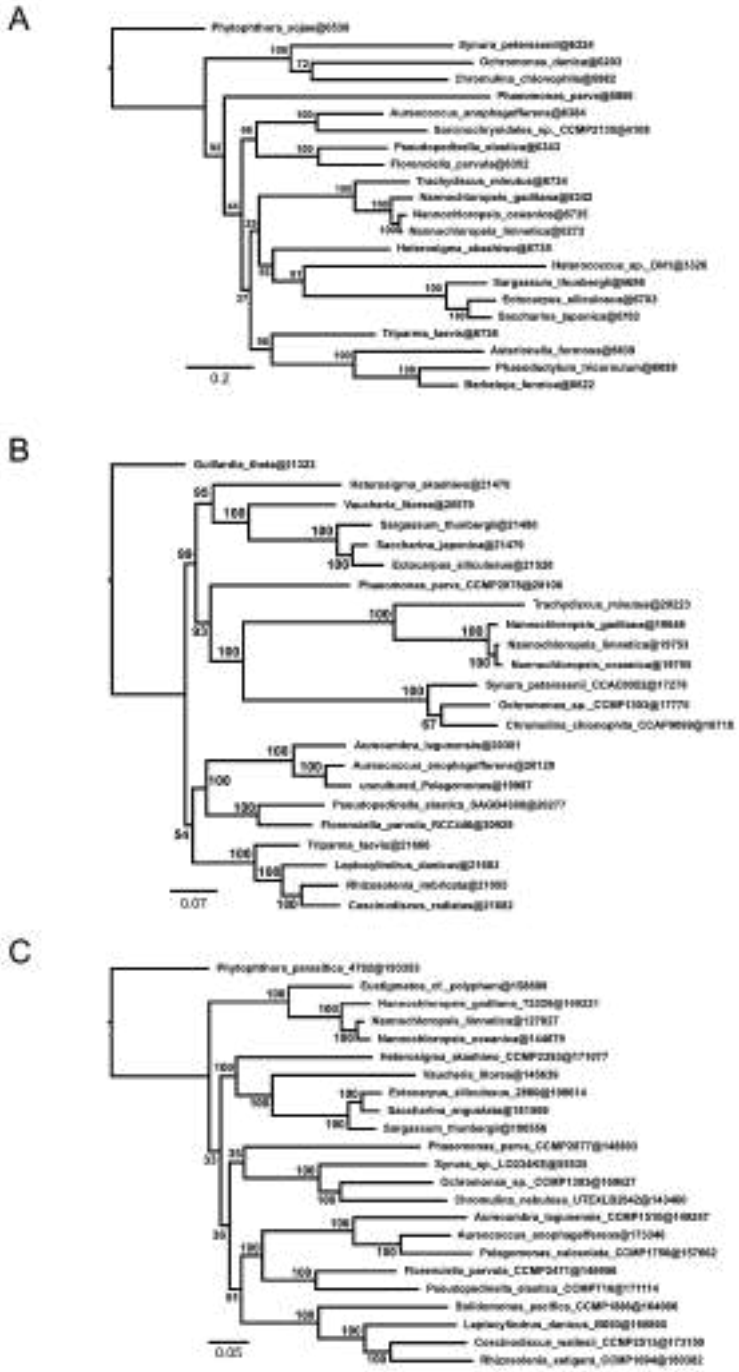
https://itol.embl.de/tree/19350761126143153919071O#



Suppl. Figure 3 Legend

Complete taxon sampling of the nuclear phylogenetic tree inferred using RAxML with LG4X.

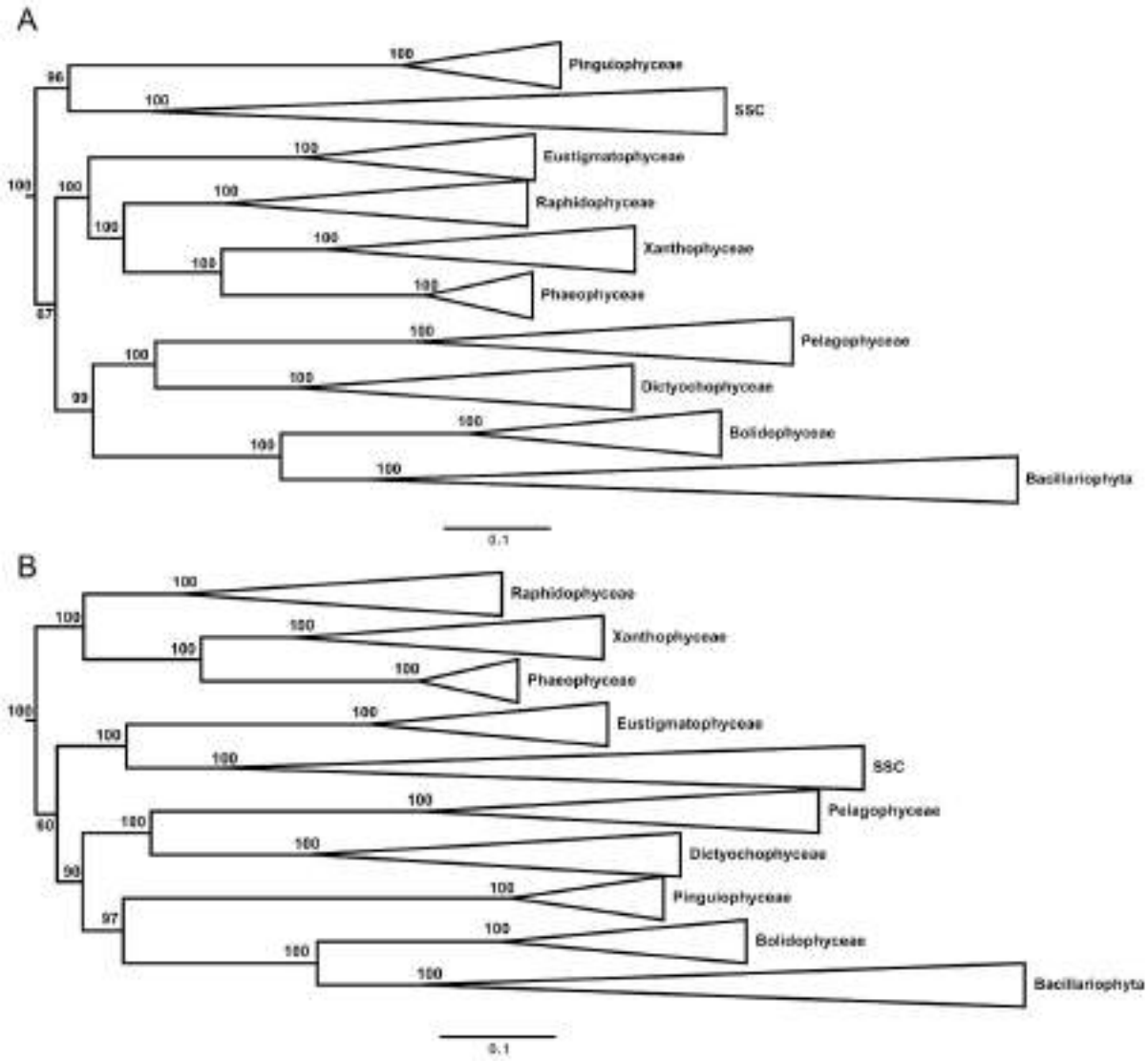
Suppl. Figure 4



Suppl. Figure 4 Legend

Phylogenetic trees of the three compartment datasets with common taxonomic sampling inferred using RAxML with LG4X. A. mitochondrion (22 species); B. plastid (23 species); C. nucleus (23 species).

Suppl. Figure 5



Suppl. Figure 5 Legend

Collapsed trees inferred from supermatrices of 200 nuclear genes selected on the basis of the size of the internal branch leading to E+SSC (RAxML with LG4X model) A. Tree inferred from the supermatrix composed of the 200 genes (39,867 positions) with the shortest E+SSC branch ($SHORT_{nu}$) B. Tree inferred from the supermatrix composed of the 200 genes (47,386 positions) with the longest E+SSC branch ($LONG_{nu}$).

Conclusion

Ce travail de thèse s'est axé autour de l'étude des eucaryotes photosynthétiques. Il avait pour premier objectif de profiter de l'apport toujours plus important de données de séquençage pour améliorer la résolution de l'arbre des eucaryotes. Cet objectif ne devait être qu'une étape menant à des questions nécessitant une connaissance de la phylogénie des espèces. Ces dernières visaient à comprendre l'impact de l'environnement sur l'évolution des communautés photosynthétiques, en mêlant des informations en provenance de la géologie, de la génétique et de la biochimie. Cependant, les difficultés et les incertitudes de l'inférence phylogénétique d'événements anciens m'ont obligé à focaliser ma thèse sur des considérations purement phylogénétiques.

Mes nouveaux objectifs ont donc concentré mon travail sur l'évaluation de la qualité de l'inférence phylogénétique. J'ai abordé ce dernier point à plusieurs échelles allant de l'alignement de séquences orthologues à l'inférence de l'arbre des espèces. J'ai notamment consacré beaucoup d'énergies à l'étude de l'impact des données erronées et de l'erreur systématique. J'ai tenté de regrouper au mieux mes résultats dans ce manuscrit de thèse bien que je ne puisse pas aborder chacune des analyses que j'ai réalisées pour orienter mes choix méthodologiques.

Dans le premier chapitre, j'ai présenté HmmCleaner, un logiciel dont l'objectif est de filtrer les segments de séquences de faible similarité dans les alignements multiples de séquences. Sa comparaison avec les programmes de filtrage de positions des alignement suggère que l'impact des erreurs d'acquisitions et d'annotations de séquences est plus important que celui des erreurs réalisées lors de l'alignement des séquences. Nos analyses ont, encore une fois, remis en avant les difficultés rencontrées lors l'inférence des arbres de gènes, à cause de la prévalence de l'erreur stochastique. L'amélioration de la qualité de cette inférence vis à vis de l'élimination de données problématiques dépend d'un équilibre fragile entre bruit et signal qui est difficile à prendre en compte par les méthodes automatiques actuelles, et requiert donc des études supplémentaires..

J'ai axé le second chapitre autour des méthodologies employées lors de la création d'un jeu de données à une échelle phylogénomique que j'ai appliquées aux eucaryotes. J'ai mis en avant la complexité du processus de validation des alignements de séquences orthologues, celle-ci étant exacerbée dans les études ciblant de larges périodes évolutives par la présence de multiples duplications et pertes de gènes. J'ai également considéré la qua-

lité des données employées en montrant que la présence de contaminations n'était pas mineure dans les projets de séquençage actuels, tout en participant au développement de méthodes pour efficacement réduire leur impact. Malgré ces différentes considérations et l'obtention d'un gros jeu de données, je ne suis pas arrivé à obtenir un arbre complètement résolu des eucaryotes. Bien que celui-ci soit en bonne adéquation avec les autres études publiées [Burki et al. 2016 ; Brown et al. 2018 ; Cenci et al. 2018 ; Lax et al. 2018], les supports obtenus pour les noeuds d'intérêt relatifs aux organismes photosynthétiques restent faibles ou potentiellement impactés par les erreurs stochastique et systématique. Il sera donc difficile d'utiliser ces résultats pour améliorer les connaissances relatives à l'acquisition de la photosynthèse.

Mon dernier chapitre s'est concentré sur l'analyse de l'impact du choix du modèle évolutif sur l'inférence phylogénétique. Nous l'avons étudié dans le cadre de la congruence entre les différents compartiments génomiques disponibles chez les straménopiles photosynthétiques. Si l'acquisition de la photosynthèse est supposée être un événement unique chez l'ancêtre de ces organismes, les incongruences observées entre les phylogénies nucléaires et mitochondriales d'une part et la phylogénie plastidiale d'autre part, n'allaient pas dans ce sens. Nous avons démontré que ces incongruences étaient le résultat du signal non-phylogénétique engendré par les violations du modèle d'évolution des séquences utilisé. En changeant de modèles, nous avons obtenu des supports faibles, en faveur d'un signal phylogénétique rare pour résoudre les noeuds à la base de ces organismes photosynthétiques. Dans ce cadre, nous avons discuté de l'importance de choisir des marqueurs présentant un fort signal pour les noeuds d'intérêts, à l'opposé de la pensée actuelle se focalisant sur le signal global. Ainsi, nous avons tiré profit du fort signal présent dans le plastide pour proposer un arbre complètement résolu des ochrophytes.

Parmi les considérations importantes soulevées à travers l'ensemble de ces trois chapitres, je noterai surtout les difficultés rencontrées pour inférer les arbres de gènes. Ces inférences interviennent à plusieurs reprises dans la construction des jeux de données et constituent une étape essentielle des analyses centrées sur la génétique ou les transferts de gènes. Cependant, en fonction de leur taille ou de leur vitesse évolutive, l'inférence est rapidement confronté à l'effet de l'erreur stochastique rendant les arbres difficilements interprétables. Je me confronte d'ailleurs encore à ces difficultés lors de la réalisation de mon projet se focalisant sur les gènes transférés en provenance de l'endosymbionte. Les faibles supports obligent à augmenter le nombre de conditions à satisfaire pour s'assurer que l'histoire évolutive du gène correspond bien au scénario recherché, ce qui, au final, réduit drastiquement le nombre de cas étudiés (e.g. [Ponce-Toledo et al. 2018 ; Dunning et al. 2019]. Des pistes pour améliorer la qualité de ces arbres se situent dans les inférences cherchant à augmenter la quantité d'informations utilisable, que ce soit par des sources extérieures aux séquences moléculaires comme les structures protéiques ou l'ordre des gènes sur les chro-

mosomes, ou en réalisant une inférence jointe des arbres de gènes et de l'arbre d'espèces [Boussau et al. 2013]. Ces dernières me semblent les plus simples à envisager dans un premier temps et auront l'avantage de considérer d'autres points problématiques comme les cas de paralogies et de transferts, mais la limitation liée au temps-calcul pourrait être rédhibitoire.

Un autre point essentiel de ma thèse est l'importance du développement des modèles d'évolution des séquences. Les ressources informatiques nécessaires à la réalisation des inférences phylogénétiques forcent à utiliser des modèles simples. Cependant, les résultats fournis par ces derniers sont très impactés par les violations de modèle. D'un autre côté, les impératifs des modèles complexes forcent à réduire la taille des jeux de données. C'est notamment une des raisons qui nous obligent à réaliser des sous-échantillonnages de nos jeux de données pour réaliser les inférences CAT. Les jackknifes nous permettent de mieux juger des incongruences entre gènes et des biais liés à l'erreur systématique mais, le fait de devoir se limiter à 20.000 ou 30.000 positions nous empêche encore de mettre complètement de côté l'impact de l'erreur stochastique. Nous sommes donc en grand besoin de modèles complexes et optimisés permettant de ne pas être limités par le matériel informatique disponible. L'arrivée de la méthode PMSF [Wang et al. 2018] et d'une implémentation plus rapide de PhyloBayes [Dang et Kishino 2019] constituent donc de bonnes nouvelles qui permettent d'envisager la comparaison future de nombreuses inférences à plein potentiel. Cependant, au moment où j'écris ces lignes, le premier est largement incorporé aux études phylogénétiques mais sans mise à l'épreuve de ses résultats dû à sa dépendance à une inférence longue avec le modèle C60 et le second suscite la recherche active d'un code source fonctionnel sur les réseaux sociaux (d'ailleurs, voir [Darriba et al. 2018] à ce sujet).

Le faible nombre d'implémentations fonctionnelles semble d'ailleurs mener à l'homogénéisation des études phylogénétiques visant à résoudre l'arbre des eucaryotes. Ainsi, peu d'études considèrent l'impact de leurs choix de données ou des modèles utilisés sur les résultats obtenus. L'analyse des études menées sur les eucaryotes va même plus loin, montrant que les marqueurs employés deviennent simplement identiques (bien que ce choix de marqueurs soit orienté par le problème de l'erreur stochastique). Dans ces conditions, les dernières études phylogénétiques publiées sur le sujet sont portées par le séquençage d'un ou de plusieurs organismes et leur placement subséquent dans l'arbre du vivant mais continuent de mener à des topologies incongruentes entre elles [Burki et al. 2016 ; Brown et al. 2018 ; Cenci et al. 2018 ; Lax et al. 2018]. S'il est assez simple d'expliquer ce point par les problématiques précédemment évoquées (i.e. difficulté de l'inférence et temps de calcul) ainsi que par les obligations qu'imposent la poursuite d'une carrière scientifique, j'aimerais voir l'application (et le développement) de méthodes permettant d'accumuler des preuves sur la véracité des hypothèses proposées. Je pense notamment à

la recherche des positions ou marqueurs violant le modèle et à une analyse de l'impact de l'échantillonnage sur l'inférence.

Si je n'ai pas su réaliser les objectifs fixés au début de ma thèse, je considère néanmoins que ma redirection aura permis de fournir des études qui, je l'espère, remettront en question certaines pratiques actuelles en phylogénomique. Cela signifie également qu'il me reste beaucoup de domaines à explorer autour de la photosynthèse chez les eucaryotes. Je pense notamment à la datation moléculaire et l'influence qu'ont le choix de la topologie et des points de calibration sur celle-ci, ou encore la recherche de gènes transférés lors des endosymbioses. Les avancées technologiques et l'élargissement constant de la biodiversité connue aidant, je ne doute pas qu'il sera un jour possible de rendre compréhensible les interactions complexes qu'y ont eu lieu entre les eucaryotes photosynthétiques et qu'il nous sera possible de comprendre comment elles ont eu lieu.

Bibliographie

A Shepherd, Daisy, et Steffen Klaere, 2019, « How Well Does Your Phylogenetic Model Fit Your Data ? », sous la dir. de Peter Foster, *Systematic Biology* 68 (*1*) : 157-167, issn : 1063-5157, doi :10.1093/sysbio/syy066.

Adl, Sina M., et al., 2018, « Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes », *Journal of Eukaryotic Microbiology* (), issn : 10665234, doi :10.1111/jeu.12691.

Altschul, SF, et al., 1997, « Gapped BLAST and PSI- BLAST : a new generation of protein database search programs », *Nucleic acids Res* 25, n° 17 () : 3389-3402, issn : 0305-1048.

Altschul, Stephen F., et al., 1990, « Basic local alignment search tool. », *Journal of molecular biology* 215, n° 3 () : 403-410, issn : 0022-2836, doi :10.1016/S0022-2836(05)80360-2.

Archibald, John M, 2009, « The Puzzle of Plastid Evolution », *Current Biology* 19, n° 2 () : 81-88, issn : 09609822, doi :10.1016/j.cub.2008.11.067.

Arnold, M L, 1992, « Natural Hybridization as an Evolutionary Process », *Annual Review of Ecology and Systematics* 23, n° 1 () : 237-261, issn : 0066-4162, doi :10.1146/annurev.es.23.110192.001321.

Baldauf, S L, et al., 2000, « A kingdom-level phylogeny of eukaryotes based on combined protein data », *Science* 290, n° 5493 () : 972-977, issn : 00368075, doi :10.1126/science.290.5493.972.

Ballenghien, Marion, Nicolas Faivre et Nicolas Galtier, 2017, « Patterns of cross-contamination in a multispecies population genomic project : detection, quantification, impact, and solutions », *BMC Biology* 15, n° 1 () : 25, issn : 1741-7007, doi :10.1186/s12915-017-0366-6.

Bapteste, Eric, et al., 2002, « The analysis of 100 genes supports the grouping of three highly divergent amoebae : Dictyostelium, Entamoeba, and Mastigamoeba », *Proceedings of the National Academy of Sciences* 99, n° 3 () : 1414-1419, issn : 0027-8424, doi :10.1073/PNAS.032662799.

Baurain, Denis, et al., 2010, « Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles », *Molecular Biology and Evolution* 27, n° 7 () : 1698-1709, issn : 07374038, doi :10.1093/molbev/msq059.

Bekker, A., et al., 2004, « Dating the rise of atmospheric oxygen », *Nature* 427, n° 6970 () : 117-120, issn : 0028-0836, doi :10.1038/nature02260.

Bengtson, Stefan, et al., 2017, « Three-dimensional preservation of cellular and subcellular structures suggests 1.6 billion-year-old crown-group red algae », sous la dir. de David Penny, *PLOS Biology* 15, n° 3 () : e2000735, issn : 1545-7885, doi :10.1371/journal.pbio.2000735.

Bhattacharya, D., et Klaus Weber, 1997, « The actin gene of the glaucocystophyte Cyanophora paradoxa : analysis of the coding region and introns, and an actin phylogeny of eukaryotes », *Current Genetics* 31, n° 5 () : 439-446, issn : 0172-8083, doi :10.1007/s002940050227.

Bhattacharya, D, et al., 1995, « Comparisons of nuclear-encoded small-subunit ribosomal RNAs reveal the evolutionary position of the Glaucocystophyta. », *Molecular Biology and Evolution* 12, n° 3 () : 415-420, issn : 1537-1719, doi :10.1093/oxfordjournals.molbev.a040216.

Bodył, Andrzej, 2018, « Did some red alga-derived plastids evolve via kleptoplastidy ? A hypothesis », *Biological Reviews* 93, n° 1 () : 201-222, issn : 14647931, doi :10.1111/brv.12340.

Boussau, Bastien, et al., 2013, « Genome-scale coestimation of species and gene trees », *Genome Research* 23 (*2*) : 323-330, issn : 10889051, doi :10.1101/gr.141978.112.

Brinkmann, H., et H. Philippe, 1999, « Archaea sister group of Bacteria ? Indications from tree reconstruction artifacts in ancient phylogenies », *Molecular Biology and Evolution* 16, n° 6 () : 817-825, issn : 0737-4038, doi :10.1093/oxfordjournals.molbev.a026166.

Brocks, J J, et al., 1999, « Archean molecular fossils and the early rise of eukaryotes. », *Science (New York, N.Y.)* 285, n° 5430 () : 1033-6, issn : 0036-8075, doi :10.1126/SCIENCE.285.5430.1033.

Brocks, Jochen J., et al., 2017, « The rise of algae in Cryogenian oceans and the emergence of animals », *Nature* 548, n° 7669 () : 578-581, issn : 0028-0836, doi :10.1038/nature23457.

Brocks, Jochen J, et al., 2003, « A reconstruction of Archean biological diversity based on molecular fossils from the 2.78 to 2.45 billion-year-old Mount Bruce Supergroup, Hamersley Basin, Western Australia », *Geochimica et Cosmochimica Acta* 67, n° 22 () : 4321-4335, issn : 0016-7037, doi :10.1016/S0016-7037(03)00209-6.

Brown, Matthew W, et al., 2013, « Phylogenomics demonstrates that breviate flagellates are related to opisthokonts and apusomonads. », *Proceedings. Biological sciences / The Royal Society* 280 :20131755, issn : 1471-2954, doi :10.1098/rspb.2013.1755.

Brown, Matthew W, et al., 2018, « Phylogenomics Places Orphan Protistan Lineages in a Novel Eukaryotic Super-Group », *Genome Biology and Evolution* 10, n° 2 () : 427-433, issn : 1759-6653, doi :10.1093/gbe/evy014.

Burki, Fabien, 2017, « The Convoluted Evolution of Eukaryotes With Complex Plastids », *Advances in Botanical Research* 84 () : 1-30, issn : 0065-2296, doi :10.1016/BS.ABR.2017.06.001.

— . 2016, « The Eukaryotic Tree of Life from a Global Phylogenomic Perspective », *Cold Spring Harbor perspectives in biology* 6, n° 5 () : a016147, issn : 19430264, doi :10.1101/cshperspect.a016147.

Burki, Fabien, Kamran Shalchian-Tabrizi et Jan Pawlowski, 2008, « Phylogenomics reveals a new 'megagroup' including most photosynthetic eukaryotes. », *Biology letters* 4, n° 4 () : 366-9, issn : 1744-9561, doi :10.1098/rsbl.2008.0224.

Burki, Fabien, et al., 2013, « Phylogenomics of the Intracellular Parasite Mikrocytos mackini Reveals Evidence for a Mitosome in Rhizaria », *Current Biology* 23, n° 16 () : 1541-1547, issn : 0960-9822, doi :10.1016/J.CUB.2013.06.033.

Burki, Fabien, et al., 2007, « Phylogenomics Reshuffles the Eukaryotic Supergroups », sous la dir. de Geraldine Butler, *PLoS ONE* 2, n° 8 () : e790, issn : 1932-6203, doi :10.1371/journal.pone.0000790.

Burki, Fabien, et al., 2012, « The evolutionary history of haptophytes and cryptophytes : phylogenomic evidence for separate origins », *Proceedings of the Royal Society of London B : Biological Sciences* 279, n° 1736 () : 2246-2254, issn : 0962-8452, doi :10.1098/RSPB.2011.2301.

Burki, Fabien, et al., 2016, « Untangling the early diversification of eukaryotes : a phylogenomic study of the evolutionary origins of Centrohelida , Haptophyta and Cryptista », *Proceedings. Biological sciences* 283, n° 1823 () : 20152802, issn : 0962-8452, doi :10.1098/rspb.2015.2802.

Butterfield, Nicholas J., 2000, « Bangiomorpha pubescens n. gen., n. sp. : implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes », *Paleobiology* 26, n° 3 () : 386-404, issn : 0094-8373, doi :10.1666/0094-8373(2000)026<0386:BPNGNS>2.0.CO;2.

Capella-Gutiérrez, Salvador, José M. Silla-Martínez et Toni Gabaldón, 2009, « trimAl : A tool for automated alignment trimming in large-scale phylogenetic analyses », *Bioinformatics* 25 (*15*) : 1972-1973, issn : 13674803, doi :10.1093/bioinformatics/btp348.

Castresana, J., 2000, « Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis », *Molecular Biology and Evolution* 17 (*4*) : 540-552, issn : 0737-4038, doi :10.1093/oxfordjournals.molbev.a026334.

Cavalier-Smith, T, 2003, « Genomic reduction and evolution of novel genetic membranes and protein-targeting machinery in eukaryote-eukaryote chimaeras (meta-algae). », *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 358, n° 1429 () : 109-33, discussion 133-4, issn : 0962-8436, doi :10.1098/rstb.2002.1194.

— . 1999, « Principles of protein and lipid targeting in secondary symbiogenesis : euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree », *The Journal of eukaryotic microbiology* 46, n° 4 () : 347-366, issn : 1066-5234.

Cavalier-Smith, Thomas, Ema E. Chao et Rhodri Lewis, 2015a, « Multiple origins of Heliozoa from flagellate ancestors : New cryptist subphylum Corbihelia, superclass Corbistoma, and monophyly of Haptista, Cryptista, Hacrobia and Chromista », *Molecular Phylogenetics and Evolution* 93 () : 331-362, issn : 10959513, doi :10.1016/j.ympev.2015.07.004.

Cavalier-Smith, Thomas, et al., 2014, « Multigene eukaryote phylogeny reveals the likely protozoan ancestors of opisthokonts (animals, fungi, choanozoans) and Amoebozoa », *Molecular Phylogenetics and Evolution* 81 () : 71-85, issn : 1055-7903, doi :10.1016/J.YMPEV.2014.08.012.

Cavalier-Smith, Thomas, et al., 2015b, « Multigene phylogeny resolves deep branching of Amoebozoa », *Molecular Phylogenetics and Evolution* 83 :293-304, issn : 10959513, doi :10.1016/j.ympev.2014.08.011.

Cembella, Allan D., 2003, « Chemical ecology of eukaryotic microalgae in marine ecosystems », *Phycologia* 42, n° 4 () : 420-447, issn : 0031-8884, doi :10.2216/i0031-8884-42-4-420.1.

Cenci, Ugo, et al., 2018, « Nuclear genome sequence of the plastid-lacking cryptomonad Goniomonas avonlea provides insights into the evolution of secondary plastids », *BMC Biology* 16, n° 1 () : 137, issn : 1741-7007, doi :10.1186/s12915-018-0593-5.

Chatzou, Maria, et al., 2016, « Multiple sequence alignment modeling : methods and applications », *Briefings in Bioinformatics* 17 (*6*) : 1009-1023, issn : 1467-5463, doi :10.1093/bib/bbv099. arXiv : 1411.3409.

Chen, F., et al., 2006, « OrthoMCL-DB : querying a comprehensive multi-species collection of ortholog groups », *Nucleic Acids Research* 34, n° 90001 () : D363-D368, issn : 0305-1048, doi :10.1093/nar/gkj123.

Chevreux, B., T. Wetter et S. Suhai, 1999, « Genome Sequence Assembly Using Trace Signals and Additional Sequence Information ». *in Computer Science and Biology : Proceedings of the German Conference on Bioinformatics (GCB)*, 45-56.

Cornet, Luc, et al., 2018, « Consensus assessment of the contamination level of publicly available cyanobacterial genomes », sous la dir. de Francisco Rodriguez-Valera, *PLOS ONE* 13, n° 7 () : e0200323, issn : 1932-6203, doi :10.1371/journal.pone.0200323.

Criscuolo, Alexis, et Simonetta Gribaldo, 2010, « BMGE (Block Mapping and Gathering with Entropy) : a new software for selection of phylogenetic informative regions from multiple sequence alignments. », *BMC evolutionary biology* 10 :210, issn : 1471-2148, doi :10.1186/1471-2148-10-210.

— . 2011, « Large-Scale Phylogenomic Analyses Indicate a Deep Origin of Primary Plastids within Cyanobacteria », *Molecular Biology and Evolution* 28, n° 11 () : 3019-3032, issn : 1537-1719, doi :10.1093/molbev/msr108.

Crotty, Stephen M., et al., 2017, « GHOST : Recovering Historical Signal from Heterotachously-evolved Sequence Alignments », *bioRxiv* () : 174789, doi :10.1101/174789.

Curtis, Bruce a, et al., 2012, « Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. », *Nature* 492 (*7427*) : 59-65, issn : 1476-4687, doi :10.1038/nature11681.

Dalquen, Daniel A., et Christophe Dessimoz, 2013, « Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals », *Genome Biol Evol* 5, n° 10 () : 1800-1806, issn : 1759-6653, doi :10.1093/gbe/evt132.

Dang, Tung, et Hirohisa Kishino, 2019, « Stochastic Variational Inference for Bayesian Phylogenetics : A Case of CAT Model », sous la dir. de Rasmus Nielsen, *Molecular Biology and Evolution* (), issn : 0737-4038, doi :10.1093/molbev/msz020.

Darriba, Diego, Tomáš Flouri et Alexandros Stamatakis, 2018, « The State of Software for Evolutionary Biology », sous la dir. de Keith Crandall, *Molecular Biology and Evolution* 35, n° 5 () : 1037-1046, issn : 0737-4038, doi :10.1093/molbev/msy014.

Darwin, Charles, 1859, *The origin of species by means of natural selection*, London : Murray, isbn : 145381468X, doi :10.1126/science.146.3640.51-b. arXiv : arXiv:1011.1669v3.

Dayton, P K, 1985, « Ecology of Kelp Communities », *Annual Review of Ecology and Systematics* 16, n° 1 () : 215-245, issn : 0066-4162, doi :10.1146/annurev.es.16.110185.001243.

De Oliveira Martins, Leonardo, Diego Mallo et David Posada, 2016, « A Bayesian supertree model for genome-wide species tree reconstruction », *Systematic Biology* 65 (*3*) : 397-416, issn : 1076836X, doi :10.1093/sysbio/syu082.

Delsuc, Frédéric, Henner Brinkmann et Hervé Philippe, 2005, « Phylogenomics and the reconstruction of the tree of life », *Nature Reviews Genetics* 6, nº 5 () : 361-375, issn : 1471-0056, doi :10.1038/nrg1603.

Delsuc, Frédéric, et al., 2008, « Additional molecular support for the new chordate phylogeny », *genesis* 46, nº 11 () : 592-604, issn : 1526954X, doi :10.1002/dvg.20450.

Di Rienzi, Sara C, et al., 2013, « The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria », *eLife* 2 (), issn : 2050-084X, doi :10.7554/eLife.01102.

Dodd, Matthew S., et al., 2017, « Evidence for early life in Earth's oldest hydrothermal vent precipitates », *Nature* 543, nº 7643 () : 60-64, issn : 0028-0836, doi :10.1038/nature21377.

Doust, Alexander B., et al., 2006, « The photophysics of cryptophyte light-harvesting », *Journal of Photochemistry and Photobiology A : Chemistry* 184, **numbers** 1-2 () : 1-17, issn : 1010-6030, doi :10.1016/J.JPHOTOCHEM.2006.06.006.

Dunning, Luke T, et al., 2019, « Lateral transfers of large DNA fragments spread functional genes among grasses. », *Proceedings of the National Academy of Sciences of the United States of America* 116, nº 10 () : 4416-4425, issn : 1091-6490, doi :10.1073/pnas.1810031116.

Ebersberger, Ingo, Sascha Strauss et Arndt von Haeseler, 2009, « HaMStR : Profile hidden markov model based search for orthologs in ESTs », *BMC Evolutionary Biology* 9, nº 1 () : 157, issn : 1471-2148, doi :10.1186/1471-2148-9-157.

Eddy, Sean R., 2011, « Accelerated profile HMM searches », *PLoS Computational Biology* 7 (*10*), issn : 1553734X, doi :10.1371/journal.pcbi.1002195. arXiv : NIHMS150003.

Eddy, Sr, 1998, « Profile hidden Markov models. », *Bioinformatics* 14 (*9*) : 755-763, issn : 13674803, doi :btb114[pii].

Eisen, Jonathan A., et Claire M. Fraser, 2003, « Phylogenomics : Intersection of Evolution and Genomics », *Science* 300, nº 5626 () : 1706-1707, issn : 0036-8075, doi :10.1126/SCIENCE.1086292.

Emms, David M, et Steven Kelly, 2015, « OrthoFinder : solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy », *Genome Biology* 16 (*1*) : 1-14, issn : 1474-760X, doi :10.1186/s13059-015-0721-2.

Enright, A J, S Van Dongen et C A Ouzounis, 2002, « An efficient algorithm for large-scale detection of protein families. », *Nucleic acids research* 30, nº 7 () : 1575-84, issn : 1362-4962.

Esson, Heather J., et Brian S. Leander, 2008, « NOVEL PELLICLE SURFACE PATTERNS ON EUGLENA OBTUSA (EUGLENOPHYTA) FROM THE MARINE BENTHIC ENVIRONMENT : IMPLICATIONS FOR PELLICLE DEVELOPMENT AND EVOLUTION », *Journal of Phycology* 44, nº 1 () : 132-141, issn : 00223646, doi :10.1111/j.1529-8817.2007.00447.x.

Falkowski, Paul G., et al., 2004, « The evolution of modern eukaryotic phytoplankton. », *Science (New York, N.Y.)* 305, nº 5682 () : 354-360, issn : 0036-8075, doi :10.1126/science.1095964.

Felsenstein, Joseph, 1978a, « Cases in which Parsimony or Compatibility Methods Will be Positively Misleading », *Systematic Zoology* 27, nº 4 () : 401, issn : 00397989, doi :10.2307/2412923.

— . 1978b, « The Number of Evolutionary Trees », *Systematic Zoology* 27, nº 1 () : 27, issn : 00397989, doi :10.2307/2412810.

Felsenstein, Joseph., 2004, *Inferring phylogenies*, 664, Sinauer Associates, isbn : 9780878931774.

Felsner, Gregor, et al., 2011, « ERAD Components in Organisms with Complex Red Plastids Suggest Recruitment of a Preexisting Protein Transport Pathway for the Periplastid Membrane », *Genome Biology and Evolution* 3 () : 140-150, issn : 1759-6653, doi :10.1093/gbe/evq074.

Feuda, Roberto, et al., 2017, « Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals », *Current Biology* 27, nº 24 () : 3864-3870.e4, issn : 0960-9822, doi :10.1016/J.CUB.2017.11.008.

Fitch, Walter M., 1970, « Distinguishing homologous from analogous proteins », *Systematic Zoology* 19, nº 2 () : 99-113, issn : 00397989, doi :10.2307/2412448.

— . 2000, « Homology - a personal view on some of the problems », *Trends Genet* 16, nº 5 () : 227-231, issn : 0168-9525, doi :10.1016/S0168-9525(00)02005-9.

Foster, Peter G, 2004, « Modeling Compositional Heterogeneity », sous la dir. de Ted Schultz, *Systematic Biology* 53, nº 3 () : 485-495, issn : 1063-5157, doi :10.1080/10635150490445779.

Galtier, N., et M. Gouy, 1998, « Inferring pattern and process : maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis », *Molecular Biology and Evolution* 15, nº 7 () : 871-879, issn : 0737-4038, doi :10.1093/oxfordjournals.molbev.a025991.

Galtier, N, et M Gouy, 1995, « Inferring phylogenies from DNA sequences of unequal base compositions. », *Proceedings of the National Academy of Sciences of the United States of America* 92, nº 24 () : 11317-21, issn : 0027-8424.

Galtier, Nicolas, 2001, « Maximum-Likelihood Phylogenetic Analysis Under a Covarion-like Model », *Molecular Biology and Evolution* 18, nº 5 () : 866-873, issn : 1537-1719, doi :10.1093/oxfordjournals.molbev.a003868.

Galtier, Nicolas, et Vincent Daubin, 2008, « Dealing with incongruence in phylogenomic analyses. », *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 363, nº 1512 () : 4023-9, issn : 1471-2970, doi :10.1098/rstb.2008.0144.

Gatesy, John, et Mark S. Springer, 2014, « Phylogenetic analysis at deep timescales : Unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum », *Molecular Phylogenetics and Evolution* 80 (*1*) : 231-266, issn : 10959513, doi :10.1016/j.ympev.2014.08.013.

Germot, Agnès, et Hervé Philippe, 1999, « Critical Analysis of Eukaryotic Phylogeny : A Case Study Based on the HSP70 Family », *The Journal of Eukaryotic Microbiology* 46, nº 2 () : 116-124, issn : 1066-5234, doi :10.1111/j.1550-7408.1999.tb04594.x.

Gilbert, Walter, 1986, « Origin of life : The RNA world », *Nature* 319, nº 6055 () : 618-618, issn : 0028-0836, doi :10.1038/319618a0.

Glass, J. B., F. Wolfe-Simon et A. D. Anbar, 2009, « Coevolution of metal availability and nitrogen assimilation in cyanobacteria and algae », *Geobiology* 7, nº 2 () : 100-123, issn : 14724677, doi :10.1111/j.1472-4669.2009.00190.x.

Gómez, Fernando, 2012, « A quantitative review of the lifestyle, habitat and trophic diversity of dinoflagellates (Dinoflagellata, Alveolata) », *Systematics and Biodiversity* 10, nº 3 () : 267-275, issn : 1477-2000, doi :10.1080/14772000.2012.721021.

Gould, Sven B., Uwe-G. Maier et William F. Martin, 2015, « Protein Import and the Origin of Red Complex Plastids », *Current Biology* 25, nº 12 () : R515-R521, issn : 0960-9822, doi :10.1016/J.CUB.2015.04.033.

Gouy, Richard, Denis Baurain et Hervé Philippe, 2015, « Rooting the tree of life : the phylogenetic jury is still out », *Philosophical Transactions of the Royal Society B : Biological Sciences* 370 (*1678*) : 20140329, issn : 0962-8436, doi :10.1098/rstb.2014.0329.

Grant, Jessica R., et Laura a. Katz, 2014, « Building a Phylogenomic Pipeline for the Eukaryotic Tree of Life - Addressing Deep Phylogenies with Genome-Scale Data », *PLoS Currents*, nº APR : 1-19, issn : 21573999, doi :10.1371/currents.tol.c24b6054aebf3602748ac042ccc8f2e9.

Gray, Michael W., et John M. Archibald, 2012, « Origins of Mitochondria and Plastids ». *in Genomics of Chloroplasts and Mitochondria*, sous la dir. de Ralph Bock et Volker Knoop, 35 :103-126, Advances in Photosynthesis and Respiration, Springer Netherlands, isbn : 978-94-007-2919-3, doi :10.1007/978-94-007-2920-9. arXiv : arXiv:1011.1669v3.

Groussin, M., B. Boussau et M. Gouy, 2013, « A Branch-Heterogeneous Model of Protein Evolution for Efficient Inference of Ancestral Sequences », *Systematic Biology* 62, n° 4 () : 523-538, issn : 1076-836X, doi :10.1093/sysbio/syt016.

Guimaraes, Joao C, et Mihaela Zavolan, 2016, « Patterns of ribosomal protein expression specify normal and malignant human cells. », *Genome biology* 17 (*1*) : 236, issn : 1474-760X, doi :10.1186/s13059-016-1104-z.

Haldane, John B. S., 1929, « The Origin of Life », *The Rationalist Annual*.

Hammer, A, R Schumann et H Schubert, 2002, « Light and temperature acclimation of Rhodomonas salina (Cryptophyceae) : photosynthetic performance », *Aquatic Microbial Ecology* 29, n° 3 () : 287-296, issn : 0948-3055, doi :10.3354/ame029287.

Hampl, Vladimir, et al., 2009, « Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". », *Proceedings of the National Academy of Sciences of the United States of America* 106 (*10*) : 3859-3864, issn : 0027-8424, doi :10.1073/pnas.0807880106.

He, Ding, et al., 2014, « An Alternative Root for the Eukaryote Tree of Life », *Current Biology* 24 (*4*) : 465-470, issn : 09609822, doi :10.1016/j.cub.2014.01.036.

He, Ding, et al., 2016, « Reducing long-branch effects in multi-protein data uncovers a close relationship between Alveolata and Rhizaria », *Molecular Phylogenetics and Evolution* 101 () : 1-7, issn : 1055-7903, doi :10.1016/J.YMPEV.2016.04.033.

Heiss, Aaron A, et al., 2018, « Combined morphological and phylogenomic re-examination of malawimonads, a critical taxon for inferring the evolutionary history of eukaryotes. », *Royal Society open science* 5, n° 4 () : 171707, issn : 2054-5703, doi :10.1098/rsos.171707.

Heled, J., et A. J. Drummond, 2010, « Bayesian Inference of Species Trees from Multilocus Data », *Molecular Biology and Evolution* 27, n° 3 () : 570-580, issn : 0737-4038, doi :10.1093/molbev/msp274.

Hoang, Diep Thi, et al., 2018, « UFBoot2 : Improving the Ultrafast Bootstrap Approximation », *Molecular Biology and Evolution* 35, n° 2 () : 518-522, issn : 0737-4038, doi :10.1093/molbev/msx281.

Hrdy, Ivan, et al., 2004, « Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I », *Nature* 432, n° 7017 () : 618-622, issn : 0028-0836, doi :10.1038/nature03149.

Imanian, Behzad, Jean-François Pombert et Patrick J. Keeling, 2010, « The Complete Plastid Genomes of the Two 'Dinotoms' Durinskia baltica and Kryptoperidinium foliaceum », sous la dir. d'Anita Brandstaetter, *PLoS ONE* 5, n° 5 () : e10711, issn : 1932-6203, doi :10.1371/journal.pone.0010711.

Ishida, Ken-ichiro, Hiroko Endo et Sayaka Koike, 2011, « Partenskyella glossopodia (Chlorarachniophyceae) possesses a nucleomorph genome of approximately 1Mbp », *Phycological Research* 59, n° 2 () : 120-122, issn : 13220829, doi :10.1111/j.1440-1835.2011.00608.x.

Jackson, Christopher, et al., 2018, « Plastid phylogenomics with broad taxon sampling further elucidates the distinct evolutionary origins and timing of secondary green plastids », *Scientific Reports* 8, n° 1 () : 1523, issn : 2045-2322, doi :10.1038/s41598-017-18805-w.

Janouškovec, Jan, et al., 2015, « Factors mediating plastid dependency and the origins of parasitism in apicomplexans and their close relatives. », *Proceedings of the National Academy of Sciences of the United States of America* 112, n° 33 () : 10200-7, issn : 1091-6490, doi :10.1073/pnas.1423790112.

Javaux, Emmanuelle, 2011, « Early eukaryotes in Precambrian oceans ». *in Origins and Evolution of Life*, sous la dir. de Muriel Gargaud, Purificacion Lopez-Garcia et Herve Martin, 414-449, Cambridge : Cambridge University Press, doi :10.1017/CBO9780511933875.028.

Javaux, Emmanuelle J., et Kevin Lepot, 2018, « The Paleoproterozoic fossil record : Implications for the evolution of the biosphere during Earth's middle-age », *Earth-Science Reviews* 176 () : 68-86, issn : 0012-8252, doi :10.1016/J.EARSCIREV.2017.10.001.

Jeffroy, Olivier, et al., 2006, « Phylogenomics : the beginning of incongruence ? », *Trends in Genetics* 22 (*4*), doi :10.1016/j.tig.2006.02.003.

Johnson, Matthew D., 2011, « The acquisition of phototrophy : Adaptive strategies of hosting endosymbionts and organelles », *Photosynthesis Research* 107, n° 1 () : 117-132, issn : 01668595, doi :10.1007/s11120-010-9546-8.

Kalyaanamoorthy, Subha, et al., 2017, « ModelFinder : fast model selection for accurate phylogenetic estimates », *Nature Methods* 14, n° 6 () : 587-589, issn : 1548-7091, doi :10.1038/nmeth.4285.

KATOH, Kazutaka, et Daron M. STANDLEY, 2013, « MAFFT multiple sequence alignment software version 7 : improvements in performance and usability », *Mol Biol Evol* 30 (*4*) : 772-780, ISSN : 07374038, doi :10.1093/molbev/mst010.

KATOH, Kazutaka, et al., 2002, « MAFFT : a novel method for rapid multiple sequence alignment based on fast Fourier transform. », *Nucleic acids research* 30 (*14*) : 3059-3066, ISSN : 1362-4962, doi :10.1093/nar/gkf436. arXiv : journal.pone.0035671 [10.1371].

KATZ, L. a., et J. R. GRANT, 2014, « Taxon-Rich Phylogenomic Analyses Resolve the Eukaryotic Tree of Life and Reveal the Power of Subsampling by Sites », *Systematic Biology* 64 (*3*) : 406-415, ISSN : 1063-5157, doi :10.1093/sysbio/syu126.

KATZ, Laura A., 2015, « Recent events dominate interdomain lateral gene transfers between prokaryotes and eukaryotes and, with the exception of endosymbiotic gene transfers, few ancient transfer events persist », *Philosophical Transactions of the Royal Society B : Biological Sciences* 370 (*1678*) : 20140324, ISSN : 0962-8436, doi :10.1098/rstb.2014.0324.

KATZ, Laura A., et al., 2011, « Subulatomonas tetraspora nov. gen. nov. sp. is a Member of a Previously Unrecognized Major Clade of Eukaryotes », *Protist* 162 (*5*) : 762-773, ISSN : 14344610, doi :10.1016/j.protis.2011.05.002.

KEELING, P. J., et al., 1999, « The secondary endosymbiont of the cryptomonad Guillardia theta contains alpha-, beta-, and gamma-tubulin genes », *Molecular Biology and Evolution* 16, n° 9 () : 1308-1313, ISSN : 0737-4038, doi :10.1093/oxfordjournals.molbev.a026221.

KEELING, Patrick J., et John M. ARCHIBALD, 2008, « Organelle Evolution : What's in a Name ? », *Current Biology* 18, n° 8 () : R345-R347, ISSN : 0960-9822, doi :10.1016/J.CUB.2008.02.065.

KEELING, Patrick J., et Jeffrey D. PALMER, 2008, « Horizontal gene transfer in eukaryotic evolution », *Nature Reviews Genetics* 9, n° 8 () : 605-618, ISSN : 1471-0056, doi :10.1038/nrg2386.

KEELING, Patrick J., et al., 2014, « The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) : Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing », *PLoS Biology* 12 (*6*), ISSN : 15457885, doi :10.1371/journal.pbio.1001889.

KIM, K., W. KIM et S. KIM, 2011, « ReMark : an automatic program for clustering orthologs flexibly combining a Recursive and a Markovclustering algorithms », *Bioinformatics* 27, n° 12 () : 1731-1733, ISSN : 1367-4803, doi :10.1093/bioinformatics/btr259.

Kirkham, Amy R., et al., 2011, « Basin-scale distribution patterns of photosynthetic picoeukaryotes along an Atlantic Meridional Transect », *Environmental Microbiology* 13, n° 4 () : 975-990, issn : 14622912, doi :10.1111/j.1462-2920.2010.02403.x.

Knauth, L Paul, et Martin J Kennedy, 2009, « The late Precambrian greening of the Earth. », *Nature* 460 (*7256*) : 728-732, issn : 0028-0836, doi :10.1038/nature08213.

Knoll, a H, et al., 2006, « Eukaryotic organisms in Proterozoic oceans. », *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 361 (*1470*) : 1023-1038, issn : 0962-8436, doi :10.1098/rstb.2006.1843.

Knoll, Andrew H, 2014, « Paleobiological perspectives on early eukaryotic evolution. », *Cold Spring Harbor perspectives in biology* 6, n° 1 () : a016121, issn : 1943-0264, doi :10.1101/cshperspect.a016121.

Koski, Liisa B., et G. Brian Golding, 2001, « The closest BLAST hit is often not the nearest neighbor », *Journal of Molecular Evolution* 52, n° 6 () : 540-2., issn : 0022-2844, doi :10.1007/s002390010184.

Kudryavtsev, Anatoliy B., Andrew D. Czaja et Abhishek B. Tripathi, 2007, « Evidence of Archean life : Stromatolites and microfossils », *Precambrian Research* 158, **numbers** 3-4 () : 141-155, issn : 0301-9268, doi :10.1016/J.PRECAMRES.2007.04.009.

Kuzniar, Arnold, et al., 2008, « The quest for orthologs : finding the corresponding gene across genomes. », *Trends in genetics : TIG* 24, n° 11 () : 539-51, issn : 0168-9525, doi :10.1016/j.tig.2008.08.009.

Kuznicki, Leszek, et al., 1990, « Photobehavior of euglenoid flagellates : Theoretical and evolutionary perspectives », *Critical Reviews in Plant Sciences* 9, n° 4 () : 343-369, issn : 0735-2689, doi :10.1080/07352689009382295.

Lartillot, Nicolas, Henner Brinkmann et Hervé Philippe, 2007, « Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model », *BMC Evolutionary Biology* 7, n° Suppl 1 () : S4, issn : 14712148, doi :10.1186/1471-2148-7-S1-S4.

Lartillot, Nicolas, et Hervé Philippe, 2004, « A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process », *Molecular Biology and Evolution* 21 (*6*) : 1095-1109, issn : 07374038, doi :10.1093/molbev/msh112.

Lartillot, Nicolas, et al., 2013, « PhyloBayes MPI : Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment », *Systematic Biology* 62 (*4*) : 611-615, doi :10.1093/sysbio/syt022.

Laurin-Lemay, Simon, Henner Brinkmann et Hervé Philippe, 2012, « Origin of land plants revisited in the light of sequence contamination and missing data », *Current Biology* 22, n° 15 () : R593-R594, issn : 0960-9822, doi :10.1016/j.cub.2012.06.013.

Lax, Gordon, et al., 2018, « Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes », *Nature* () : 1, ISSN : 0028-0836, doi :10.1038/s41586-018-0708-8.

Lazcano, A, et S L Miller, 1996, « The origin and early evolution of life : prebiotic chemistry, the pre-RNA world, and time. », *Cell* 85, n° 6 () : 793-8, ISSN : 0092-8674, doi :10.1016/S0092-8674(00)81263-5.

Le, S. Q., et O. Gascuel, 2008, « An Improved General Amino Acid Replacement Matrix », *Molecular Biology and Evolution* 25, n° 7 () : 1307-1320, ISSN : 0737-4038, doi :10.1093/molbev/msn067.

Le, Si Quang, Cuong Cao Dang et Olivier Gascuel, 2012, « Modeling protein evolution with several amino acid replacement matrices depending on site rates », *Molecular Biology and Evolution* 29 (*10*) : 2921-2936, ISSN : 07374038, doi :10.1093/molbev/mss112.

Leliaert, Frederik, et al., 2012, « Phylogeny and Molecular Evolution of the Green Algae », *Critical Reviews in Plant Sciences* 31 (*1*) : 1-46, ISSN : 0735-2689, doi :10.1080/07352689.2011.615705.

Lenning, Kees Van, et al., 2004, « Pigment diversity of coccolithophores in relation to taxonomy, phylogeny and ecological preferences ». *in Coccolithophores*, 51-73, Berlin, Heidelberg : Springer Berlin Heidelberg, doi :10.1007/978-3-662-06278-4_3.

Lenton, Timothy M., et al., 2014, « Co-evolution of eukaryotes and ocean oxygenation in the Neoproterozoic era », *Nature Geoscience* 7 (*April*) : 257-265, ISSN : 1752-0894, doi :10.1038/NGEO2108.

Li, Li, et al., 2003, « OrthoMCL : identification of ortholog groups for eukaryotic genomes. », *Genome research* 13, n° 9 () : 2178-89, ISSN : 1088-9051, doi :10.1101/gr.1224503.

Liò, P, et N Goldman, 1998, « Models of molecular evolution and phylogeny. », *Genome research* 8, n° 12 () : 1233-44, ISSN : 1088-9051, doi :10.1101/GR.8.12.1233.

Lockhart, P.J., et al., 1992, « Substitutional bias confounds inference of cyanelle origins from sequence data », *Journal of Molecular Evolution* 34, n° 2 () : 153-162, ISSN : 0022-2844, doi :10.1007/BF00182392.

Longo, Mark S., Michael J. O'Neill et Rachel J. O'Neill, 2011, « Abundant Human DNA Contamination Identified in Non-Primate Genome Databases », sous la dir. de Najib El-Sayed, *PLoS ONE* 6, n° 2 () : e16410, ISSN : 1932-6203, doi :10.1371/journal.pone.0016410.

Lopez, P., D. Casane et H. Philippe, 2002, « Heterotachy, an Important Process of Protein Evolution », *Molecular Biology and Evolution* 19, n° 1 () : 1-7, ISSN : 1537-1719, doi :10.1093/oxfordjournals.molbev.a003973.

Lunter, Gerton, et al., 2005, « Bayesian coestimation of phylogeny and sequence alignment », *BMC Bioinformatics* 6, n° 1 () : 83, issn : 14712105, doi :10.1186/1471-2105-6-83.

Ma, Xiao-Nian, et al., 2016, « Lipid Production from Nannochloropsis », *Marine Drugs* 14, n° 4 () : 61, issn : 1660-3397, doi :10.3390/md14040061.

Mackiewicz, Paweł, et Przemysław Gagat, 2014, « Monophyly of Archaeplastida supergroup and relationships among its lineages in the light of phylogenetic and phylogenomic studies. Are we close to a consensus ? », *Acta Societatis Botanicorum Poloniae* 83, n° 4 () : 263-280, issn : 2083-9480, doi :10.5586/asbp.2014.044.

Mallet, James, Nora Besansky et Matthew W. Hahn, 2016, « How reticulated are species ? », *BioEssays* 38, n° 2 () : 140-149, issn : 02659247, doi :10.1002/bies.201500149.

Mann, David G., et Pieter Vanormelingen, 2013, « An Inordinate Fondness ? The Number, Distributions, and Origins of Diatom Species », *Journal of Eukaryotic Microbiology* 60, n° 4 () : 414-420, issn : 10665234, doi :10.1111/jeu.12047.

Marshall, A. T., 1996, « Calcification in Hermatypic and Ahermatypic Corals », *Science* 271, n° 5249 () : 637-639, issn : 0036-8075, doi :10.1126/science.271.5249.637.

Martin, Ronald, et Antonietta Quigg, 2013, « Tiny Plants That Once Ruled the Seas », *Scientific American* 308, n° 6 () : 40-45, issn : 0036-8733, doi :10.1038/scientificamerican0613-40.

Martin, William, 1999, « Mosaic bacterial chromosomes : a challenge en route to a tree of genomes », *BioEssays* 21, n° 2 () : 99-104, issn : 02659247, doi :10.1002/(SICI)1521-1878(199902)21:2<99::AID-BIES3>3.0.CO;2-B.

Martin, William, et Michael J Russell, 2003, « On the origins of cells : a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. », *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 358, n° 1429 () : 59-83, discussion 83-5, issn : 0962-8436, doi :10.1098/rstb.2002.1183.

Martin, William, et al., 2008, « Hydrothermal vents and the origin of life », *Nature Reviews Microbiology* 6, n° 11 () : 805-814, issn : 1740-1526, doi :10.1038/nrmicro1991.

Masquelier, Sylvie, et al., 2011, « Distribution of eukaryotic plankton in the English Channel and the North Sea in summer », *Journal of Sea Research* 66, n° 2 () : 111-122, issn : 1385-1101, doi :10.1016/J.SEARES.2011.05.004.

McCourt, Richard M, Charles F Delwiche et Kenneth G Karol, 2004, « Charophyte algae and land plant origins », *Trends in ecology & evolution* 19, n° 12 () : 661-666, issn : 0169-5347, doi :10.1016/j.tree.2004.09.013.

Merchant, Samier, Derrick E. Wood et Steven L. Salzberg, 2014, « Unexpected cross-species contamination in genome sequencing projects », *PeerJ* 2 () : e675, issn : 2167-8359, doi :10.7717/peerj.675.

El-Metwally, Sara, et al., 2013, « Next-Generation Sequence Assembly : Four Stages of Data Processing and Computational Challenges », *PLoS Computational Biology* 9 (*12*), issn : 1553734X, doi :10.1371/journal.pcbi.1003345.

Mills, Daniel B., et Donald E. Canfield, 2014, « Oxygen and animal evolution : Did a rise of atmospheric oxygen "trigger" the origin of animals ? », *BioEssays* 36 (*12*) : 1145-1155, issn : 02659247, doi :10.1002/bies.201400101.

Minge, Marianne A, et al., 2010, « A phylogenetic mosaic plastid proteome and unusual plastid-targeting signals in the green-colored dinoflagellate Lepidodinium chlorophorum », *BMC Evolutionary Biology* 10, n° 1 () : 191, issn : 1471-2148, doi :10.1186/1471-2148-10-191.

Mirarab, Siavash, et Tandy Warnow, 2015, « ASTRAL-II : coalescent-based species tree estimation with many hundreds of taxa and thousands of genes », *Bioinformatics* 31, n° 12 () : i44-i52, issn : 1367-4803, doi :10.1093/bioinformatics/btv234.

Mooers, Arne Ø., et Edward C. Holmes, 2000, « The evolution of base composition and phylogenetic inference », *Trends in Ecology & Evolution* 15, n° 9 () : 365-369, issn : 0169-5347, doi :10.1016/S0169-5347(00)01934-0.

Moore, Christa E., et al., 2012, « Nucleomorph Genome Sequence of the Cryptophyte Alga Chroomonas mesostigmatica CCMP1168 Reveals Lineage-Specific Gene Loss and Genome Complexity », *Genome Biology and Evolution* 4, n° 11 () : 1162-1175, issn : 1759-6653, doi :10.1093/gbe/evs090.

Moore, Robert B., et al., 2008, « A photosynthetic alveolate closely related to apicomplexan parasites », *Nature* 451, n° 7181 () : 959-963, issn : 0028-0836, doi :10.1038/nature06635.

Moreira, David, Hervé Le Guyader et Hervé Philippe, 2000, « The origin of red algae and the evolution of chloroplasts », *Nature* 405, n° 6782 () : 69-72, issn : 0028-0836, doi :10.1038/35011054.

Moustafa, Ahmed, et al., 2009, « Genomic footprints of a cryptic plastid endosymbiosis in diatoms. », *Science (New York, N.Y.)* 324, n° 5935 () : 1724-6, issn : 1095-9203, doi :10.1126/science.1172983.

Mukherjee, Indrani, et al., 2018, « The Boring Billion, a slingshot for Complex Life on Earth », *Scientific Reports* 8, n° 1 () : 4432, issn : 2045-2322, doi :10.1038/s41598-018-22695-x.

Nabout, João Carlos, et al., 2013, « How many species of Cyanobacteria are there ? Using a discovery curve to predict the species number », *Biodiversity and Conservation* 22, nº 12 () : 2907-2918, issn : 0960-3115, doi :10.1007/s10531-013-0561-x.

Nakayama, Takuro, et John M Archibald, 2012, « Evolving a photosynthetic organelle », *BMC Biology* 10 (*1*) : 35, issn : 1741-7007.

Nguyen, Lam-Tung T, et al., 2015, « IQ-TREE : a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies », *Molecular Biology and Evolution* 32, nº 1 () : 268-274, issn : 1537-1719, doi :10.1093/molbev/msu300.

Nikaido, I, et al., 2000, « Generation of 10,154 expressed sequence tags from a leafy gametophyte of a marine red alga, Porphyra yezoensis », *DNA Res* 7, nº 3 () : 223-7, issn : 1340-2838.

Nosenko, Tetyana, et al., 2006, « Chimeric Plastid Proteome in the Florida "Red Tide" Dinoflagellate Karenia brevis », *Molecular Biology and Evolution* 23, nº 11 () : 2026-2038, issn : 1537-1719, doi :10.1093/molbev/msl074.

Nosenko, Tetyana, et al., 2013, « Deep metazoan phylogeny : When different genes tell different stories », *Molecular Phylogenetics and Evolution* 67, nº 1 () : 223-233, issn : 1055-7903, doi :10.1016/J.YMPEV.2013.01.010.

Nowack, Eva C M, Michael Melkonian et Gernot Glöckner, 2008, « Chromatophore genome sequence of Paulinella sheds light on acquisition of photosynthesis by eukaryotes. », *Current biology : CB* 18, nº 6 () : 410-8, issn : 0960-9822, doi :10.1016/j.cub.2008.02.051.

Oborník, Miroslav, et al., 2012, « Morphology, Ultrastructure and Life Cycle of Vitrella brassicaformis n. sp., n. gen., a Novel Chromerid from the Great Barrier Reef », *Protist* 163, nº 2 () : 306-323, issn : 1434-4610, doi :10.1016/J.PROTIS.2011.09.001.

Oparin, A I, 1924, « The Origin of Life ».

Orgel, Leslie E., 1998, « The origin of life—a review of facts and speculations », *Trends in Biochemical Sciences* 23, nº 12 () : 491-495, issn : 0968-0004, doi :10.1016/S0968-0004(98)01300-0.

Parfrey, Laura Wegener, et al., 2011, « Estimating the timing of early eukaryotic diversification with multigene molecular clocks », *Proceedings of the National Academy of Sciences* 108 (*33*) : 13624-13629, issn : 0027-8424, doi :10.1073/pnas.1110633108.

Pearce, Ben K D, et al., 2017, « Origin of the RNA world : The fate of nucleobases in warm little ponds. », *Proceedings of the National Academy of Sciences of the United States of America* 114, nº 43 () : 11327-11332, issn : 1091-6490, doi :10.1073/pnas.1710339114.

Percival, Steven L., et David W. Williams, 2014, « Cyanobacteria ». Chap. 4 *in Microbiology of Waterborne Diseases*, 79-88, Academic Press, isbn : 9780124158467, doi :10.1016/B978-0-12-415846-7.00005-6.

Philippe, Hervé, et Béatrice Roure, 2011, « Difficult phylogenetic questions : more data, maybe ; better methods, certainly », *BMC Biology* 9, n° 1 () : 91, issn : 1741-7007, doi :10.1186/1741-7007-9-91.

Philippe, Hervé, et al., 2011a, « Acoelomorph flatworms are deuterostomes related to Xenoturbella », *Nature* 470, n° 7333 () : 255-258, issn : 0028-0836, doi :10.1038/nature09676.

Philippe, Hervé, et al., 2005, « Heterotachy and long-branch attraction in phylogenetics », *BMC Evolutionary Biology* 5 (*1*) : 50, issn : 14712148, doi :10.1186/1471-2148-5-50.

Philippe, Hervé, et al., 2004, « Phylogenomics of eukaryotes : impact of missing data on large alignments », *Molecular Biology and Evolution* 21, n° 9 () : 1740-1752, issn : 1537-1719, doi :10.1093/molbev/msh182.

Philippe, Hervé, et al., 2011b, « Resolving difficult phylogenetic questions : Why more sequences are not enough », *PLoS Biology* 9 (*3*), issn : 15449173, doi :10.1371/journal.pbio.1000602.

Ponce-Toledo, Rafael I., et al., 2017, « An Early-Branching Freshwater Cyanobacterium at the Origin of Plastids », *Current Biology* 27, n° 3 () : 386-391, issn : 0960-9822, doi :10.1016/J.CUB.2016.11.056.

Ponce-Toledo, Rafael I, et al., 2018, « Secondary Plastids of Euglenids and Chlorarachniophytes Function with a Mix of Genes of Red and Green Algal Ancestry », *Molecular Biology and Evolution* 35 (*9*) : 2198-2204, issn : 0737-4038, doi :10.1093/molbev/msy121.

Qiu, Huan, Hwan Su Yoon et Debashish Bhattacharya, 2013a, « Algal endosymbionts as vectors of horizontal gene transfer in photosynthetic eukaryotes. », *Frontiers in plant science* 4 (*September*) : 366, issn : 1664-462X, doi :10.3389/fpls.2013.00366.

Qiu, Huan, et al., 2013b, « Adaptation through horizontal gene transfer in the cryptoendolithic red alga Galdieria phlegrea », *Current Biology* 23 (*19*) : R865-R866, issn : 09609822, doi :10.1016/j.cub.2013.08.046.

Redelings, Benjamin D., et Marc A. Suchard, 2005, « Joint Bayesian Estimation of Alignment and Phylogeny », sous la dir. de Paul Lewis, *Systematic Biology* 54, n° 3 () : 401-418, issn : 1076-836X, doi :10.1080/10635150590947041.

Reyes-Prieto, Adrian, Ahmed Moustafa et Debashish Bhattacharya, 2008, « Multiple Genes of Apparent Algal Origin Suggest Ciliates May Once Have Been Photosynthetic », *Current Biology* 18, n° 13 () : 956-962, issn : 0960-9822, doi :10.1016/J.CUB.2008.05.042.

Rippka, R., J. Waterbury et G. Cohen-Bazire, 1974, « A cyanobacterium which lacks thylakoids », *Archives of Microbiology* 100 (*1*) : 419-436, issn : 0302-8933, doi :10.1007/BF00446333.

Rivera, Maria C., et al., 1998, « Genomic evidence for two functionally distinct gene classes », *Proceedings of the National Academy of Sciences* 95, n° 11 () : 6239-6244, issn : 0027-8424, doi :10.1073/PNAS.95.11.6239.

Rodríguez-Ezpeleta, Naiara, et al., 2007a, « Detecting and Overcoming Systematic Errors in Genome-Scale Phylogenies », sous la dir. de Frank Anderson, *Systematic Biology* 56, n° 3 () : 389-399, issn : 1076-836X, doi :10.1080/10635150701397643.

Rodríguez-Ezpeleta, Naiara, et al., 2005, « Monophyly of Primary Photosynthetic Eukaryotes : Green Plants, Red Algae, and Glaucophytes », *Current Biology* 15, n° 14 () : 1325-1330, issn : 0960-9822, doi :10.1016/J.CUB.2005.06.040.

Rodríguez-Ezpeleta, Naiara, et al., 2007b, « Toward Resolving the Eukaryotic Tree : The Phylogenetic Positions of Jakobids and Cercozoans », *Current Biology* 17 (*16*) : 1420-1425, issn : 09609822, doi :10.1016/j.cub.2007.07.036.

Rogers, Matthew B., et al., 2007, « The Complete Chloroplast Genome of the Chlorarachniophyte Bigelowiella natans : Evidence for Independent Origins of Chlorarachniophyte and Euglenid Secondary Endosymbionts », *Molecular Biology and Evolution* 24, n° 1 () : 54-62, issn : 1537-1719, doi :10.1093/molbev/msl129.

Roure, Béatrice, Denis Baurain et Hervé Philippe, 2013, « Impact of missing data on phylogenies inferred from empirical phylogenomic data sets », *Molecular Biology and Evolution* 30 (*1*) : 197-214, issn : 07374038, doi :10.1093/molbev/mss208.

Roure, Béatrice, et Hervé Philippe, 2011, « Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference », *BMC Evolutionary Biology* 11, n° 1 () : 17, issn : 1471-2148, doi :10.1186/1471-2148-11-17.

Roure, Béatrice, Naiara Rodriguez-Ezpeleta et Hervé Philippe, 2007, « SCaFoS : A tool for selection, concatenation and fusion of sequences for phylogenomics », *BMC Evolutionary Biology* 7 (*SUPPL. 1*) : 1-12, issn : 14712148, doi :10.1186/1471-2148-7-S1-S2.

Roy, Scott William, 2009, « Phylogenomics : Gene Duplication, Unrecognized Paralogy and Outgroup Choice », sous la dir. de Niyaz Ahmed, *PLoS ONE* 4, n° 2 () : e4568, issn : 1932-6203, doi :10.1371/journal.pone.0004568.

Sagan, Lynn, 1967, « On the origin of mitosing cells », *Journal of Theoretical Biology* 14, n° 3 () : 225-IN6, issn : 0022-5193, doi :10.1016/0022-5193(67)90079-3.

Schopf, J William, et al., 2018, « SIMS analyses of the oldest known assemblage of microfossils document their taxon-correlated carbon isotope compositions. », *Proceedings of the National Academy of Sciences of the United States of America* 115 (*1*) : 53-58, issn : 1091-6490, doi :10.1073/pnas.1718063115.

Seeleuthner, Yoann, et al., 2018, « Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans », *Nature Communications* 9, n° 1 () : 310, issn : 2041-1723, doi :10.1038/s41467-017-02235-3.

Sela, Itamar, et al., 2015, « GUIDANCE2 : Accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters », *Nucleic Acids Research* 43, n° W1 () : W7-W14, issn : 13624962, doi :10.1093/nar/gkv318.

Shih, Patrick M., 2015, « Cyanobacterial evolution : Fresh insight into ancient questions », *Current Biology* 25 (*5*) : R192-R193, issn : 09609822, doi :10.1016/j.cub.2014.12.046.

Si Quang, Le, Olivier Gascuel et Nicolas Lartillot, 2008, « Empirical profile mixture models for phylogenetic reconstruction », *Bioinformatics* 24, n° 20 () : 2317-2323, issn : 1460-2059, doi :10.1093/bioinformatics/btn445.

Sibbald, Shannon J., et John M. Archibald, 2017, « More protist genomes needed », *Nature Ecology & Evolution* 1, n° 5 () : 0145, issn : 2397-334X, doi :10.1038/s41559-017-0145.

Simion, Paul, et al., 2017, « A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals », *Current Biology* 27 (*7*) : 958-967, issn : 09609822, doi :10.1016/j.cub.2017.02.031.

Simion, Paul, et al., 2018, « A software tool 'CroCo' detects pervasive cross-species contamination in next generation sequencing data », *BMC Biology* 16, n° 1 () : 28, issn : 1741-7007, doi :10.1186/s12915-018-0486-7.

Soo, Rochelle M., et al., 2014, « An Expanded Genomic Representation of the Phylum Cyanobacteria », *Genome Biology and Evolution* 6, n° 5 () : 1031-1045, issn : 1759-6653, doi :10.1093/gbe/evu073.

Soo, Rochelle M., et al., 2017, « On the origins of oxygenic photosynthesis and aerobic respiration in Cyanobacteria », *Science* 355, n° 6332 () : 1436-1440, issn : 0036-8075, doi :10.1126/science.aal3794.

Stamatakis, Alexandros, 2014, « RAxML version 8 : A tool for phylogenetic analysis and post-analysis of large phylogenies », *Bioinformatics* 30 (*9*) : 1312-1313, issn : 14602059, doi :10.1093/bioinformatics/btu033. arXiv : bioinformatics/btu033 [10.1093].

Stamatakis, Alexandros, et Andre J. Aberer, 2013, « Novel Parallelization Schemes for Large-Scale Likelihood-based Phylogenetic Inference ». *in 2013 IEEE 27th International Symposium on Parallel and Distributed Processing*, 1195-1204, IEEE, isbn : 978-1-4673-6066-1, doi :10.1109/IPDPS.2013.70.

Stibitz, Thomas B., Patrick J. Keeling et Debashish Bhattacharya, 2000, « Symbiotic Origin of a Novel Actin Gene in the Cryptophyte Pyrenomonas helgolandii », *Molecular Biology and Evolution* 17, nº 11 () : 1731-1738, issn : 1537-1719, doi :10.1093/oxfordjournals.molbev.a026271.

Stiller, John W, et al., 2014, « The evolution of photosynthesis in chromist algae through Serial Endosymbioses », *Nature Communications* 5 :1-7, doi :10.1038/ncomms6764.

Strassert, Jürgen F H, et al., 2019, « New phylogenomic analysis of the enigmatic phylum Telonemia further resolves the eukaryote tree of life », sous la dir. de Beth Shapiro, *Molecular Biology and Evolution* (), issn : 0737-4038, doi :10.1093/molbev/msz012.

Struck, Torsten H., 2013, « The Impact of Paralogy on Phylogenomic Studies – A Case Study on Annelid Relationships », sous la dir. de Zhanjiang Liu, *PLoS ONE* 8, nº 5 () : e62892, issn : 1932-6203, doi :10.1371/journal.pone.0062892.

Sullivan, Jack, et Paul Joyce, 2005, « Model Selection in Phylogenetics », *Annual Review of Ecology, Evolution, and Systematics* 36, nº 1 () : 445-466, issn : 1543-592X, doi :10.1146/annurev.ecolsys.36.102003.152633.

Susko, E., et A. J. Roger, 2007, « On Reduced Amino Acid Alphabets for Phylogenetic Inference », *Molecular Biology and Evolution* 24, nº 9 () : 2139-2150, issn : 0737-4038, doi :10.1093/molbev/msm144.

Szöllősi, Gergely J., et al., 2013a, « Efficient Exploration of the Space of Reconciled Gene Trees », *Systematic Biology* 62, nº 6 () : 901-912, issn : 1076-836X, doi :10.1093/sysbio/syt054.

Szöllősi, Gergely J., et al., 2013b, « Lateral Gene Transfer from the Dead », *Systematic Biology* 62, nº 3 () : 386-397, issn : 1076-836X, doi :10.1093/sysbio/syt003.

Szöllősi, Gergely J., et al., 2015, « The Inference of Gene Trees with Species Trees », *Systematic Biology* 64, nº 1 () : e42-e62, issn : 1076-836X, doi :10.1093/sysbio/syu048.

Talavera, Gerard, et Jose Castresana, 2007, « Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments », *Systematic Biology* 56 (*4*) : 564-577, issn : 1063-5157, doi :10.1080/10635150701472164.

Tan, Ge, et al., 2015, « Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference », *Systematic Biology* 64 (*5*) : 778-791, issn : 1076836X, doi :10.1093/sysbio/syv033.

Tirok, K, et U Gaedke, 2007, « Regulation of planktonic ciliate dynamics and functional composition during spring in Lake Constance », *Aquatic Microbial Ecology* 49, n° 1 () : 87-100, issn : 0948-3055, doi :10.3354/ame01127.

Tonini, Jo ? ?o, et al., 2015, « Concatenation and species tree methods exhibit statistically indistinguishable accuracy under a range of simulated conditions », *PLoS Currents* 7 (*TREEOFLIFE*) : 1-15, issn : 21573999, doi :10.1371/currents.tol.34260cc27551a527b124ec5f6334b6be.

Uzzell, T, et K W Corbin, 1971, « Fitting discrete probability distributions to evolutionary events. », *Science (New York, N.Y.)* 172, n° 3988 () : 1089-96, issn : 0036-8075, doi :10.1126/SCIENCE.172.3988.1089.

Wächtershäuser, G, 1990, « Evolution of the first metabolic cycles. », *Proceedings of the National Academy of Sciences of the United States of America* 87, n° 1 () : 200-4, issn : 0027-8424, doi :10.1073/PNAS.87.1.200.

Wall, D. P., H. B. Fraser et A. E. Hirsh, 2003, « Detecting putative orthologs », *Bioinformatics* 19, n° 13 () : 1710-1711, issn : 1367-4803, doi :10.1093/bioinformatics/btg213.

Wang, Huai-Chun, et al., 2018, « Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation », *Systematic Biology* 67, n° 2 () : 216-235, issn : 1063-5157, doi :10.1093/sysbio/syx068.

Whelan, Simon, et Nick Goldman, 2001, « A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach », *Molecular Biology and Evolution* 18, n° 5 () : 691-699, issn : 1537-1719, doi :10.1093/oxfordjournals.molbev.a003851.

Whelan, Simon, Pietro Liò et Nick Goldman, 2001, « Molecular phylogenetics : state-of-the-art methods for looking into the past », *Trends in Genetics* 17, n° 5 () : 262-272, issn : 0168-9525, doi :10.1016/S0168-9525(01)02272-7.

Whitfield, James B., et Peter J. Lockhart, 2007, « Deciphering ancient rapid radiations », *Trends in Ecology and Evolution* 22 (*5*) : 258-265, issn : 01695347, doi :10.1016/j.tree.2007.01.012.

Wodniok, Sabina, et al., 2011, « Origin of land plants : Do conjugating green algae hold the key ? », *BMC Evolutionary Biology* 11, n° 1 () : 104, issn : 1471-2148, doi :10.1186/1471-2148-11-104.

Woese, C R, O Kandler et M L Wheelis, 1990, « Towards a natural system of organisms : proposal for the domains Archaea, Bacteria, and Eucarya. », *Proceedings of the National Academy of Sciences of the United States of America* 87, n° 12 () : 4576-9, issn : 0027-8424, doi :10.1073/PNAS.87.12.4576.

Woese, C.R., et al., 1991, « Archaeal Phylogeny : Reexamination of the Phylogenetic Position of Archaeoglohus fulgidus in Light of Certain Composition-induced Artifacts », *Systematic and Applied Microbiology* 14, n° 4 () : 364-371, issn : 0723-2020, doi :10.1016/S0723-2020(11)80311-5.

Wolf, Yuri I., et Eugene V. Koonin, 2012, « A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes », *Genome Biol Evol* 4, n° 12 () : 1286-1294, issn : 1759-6653, doi :10.1093/gbe/evs100.

Wu, Martin, Sourav Chatterji et Jonathan A. Eisen, 2012, « Accounting for alignment uncertainty in phylogenomics », *PLoS ONE* 7 (*1*) : 1-10, issn : 19326203, doi :10.1371/journal.pone.0030288.

Yang, Ziheng, 1994, « Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites : Approximate methods », *Journal of Molecular Evolution* 39, n° 3 () : 306-314, issn : 0022-2844, doi :10.1007/BF00160154.

Yoon, Hwan Su, et al., 2004, « A Molecular Timeline for the Origin of Photosynthetic Eukaryotes », *Molecular Biology and Evolution* 21, n° 5 () : 809-818, issn : 07374038, doi :10.1093/molbev/msh075.

Yoon, Hwan Su, et al., 2009, « A single origin of the photosynthetic organelle in different Paulinella lineages. », *BMC evolutionary biology* 9 :98.

Yoon, Hwan Su, et al., 2002, « The single, ancient origin of chromist plastids. », *Proceedings of the National Academy of Sciences of the United States of America* 99, n° 24 () : 15507-12, issn : 0027-8424, doi :10.1073/pnas.242379899.

Zapata, M, et al., 2004, « Photosynthetic pigments in 37 species (65 strains) of Haptophyta : implications for oceanography and chemotaxonomy », *Marine Ecology Progress Series* 270 :83-102, issn : 0171-8630, doi :10.3354/meps270083.

Zhang, Chao, et al., 2018, « ASTRAL-III : polynomial time species tree reconstruction from partially resolved gene trees », *BMC Bioinformatics* 19, n° S6 () : 153, issn : 1471-2105, doi :10.1186/s12859-018-2129-y.

ZIMBA, Paul V., et al., 2010, « Identification of euglenophycin – A toxin found in certain euglenoids », *Toxicon* 55, n° 1 () : 100-104, ISSN : 0041-0101, doi :10.1016/J.TOXICON.2009.07.004.

Annexe

TABLE 3.1 – Liste des 25 espèces utilisées pour la comparaison entre OrthoMCL et OrthoFinder.

Acanthamoeba castellanii
Arabidopsis thaliana
Batrachochytrium dendrobatidis
Bigelowiella natans
Capsaspora owczarzaki
Chlamydomonas reinhardtii
Chondrus crispus
Cyanidioschyzon merolae
Dictyostelium purpureum
Emiliania huxleyi
Galdieria sulphuraria
Giardia lamblia
Guillardia theta
Homo sapiens
Ichthyophthirius multifiliis
Monosiga brevicollis
Mucor circinelloides
Naegleria gruberi
Nannochloropsis gaditana
Perkinsus marinus
Phaeodactylum tricornutum
Physcomitrella patens
Phytophthora infestans
Toxoplasma gondii
Trichomonas vaginalis

TABLE 3.2: Liste des 91 espèces (33 eucaryotes, 28 bactéries, 30 archées) utilisées pour inférer les orthougroupes afin de réaliser le jeu de données.

Domaine	Souches
Eukaryota	*Acanthamoeba castellanii*
Eukaryota	*Arabidopsis thaliana*
Eukaryota	*Batrachochytrium dendrobatidis_JAM81*
Eukaryota	*Bigelowiella natans_CCMP2755*
Eukaryota	*Chondrus crispus*

Eukaryota	*Cyanidioschyzon merolae_strain_10D*
Eukaryota	*Capsaspora owczarzaki*
Eukaryota	*Chlamydomonas reinhardtii*
Eukaryota	*Emiliania huxleyi_CCMP1516*
Eukaryota	*Giardia lamblia_ATCC_50803*
Eukaryota	*Galdieria sulphuraria*
Eukaryota	*Guillardia theta_CCMP2712*
Eukaryota	*Homo sapiens*
Eukaryota	*Monosiga brevicollis_MX1*
Eukaryota	*Mucor circinelloides*
Eukaryota	*Nannochloropsis gaditana_CCMP526*
Eukaryota	*Naegleria gruberi*
Eukaryota	*Phytophthora infestans_T304*
Eukaryota	*Perkinsus marinus_ATCC_50983*
Eukaryota	*Physcomitrella patens*
Eukaryota	*Phaeodactylum tricornutum_CCAP_1055*
Eukaryota	*Toxoplasma gondii_ME49*
Eukaryota	*Trichomonas vaginalis_G3*
Eukaryota	*Dictyostelium discoideum*
Eukaryota	*Ectocarpus siliculosus*
Eukaryota	*Nematostella vectensis*
Eukaryota	*Tetrahymena thermophila*
Eukaryota	*Trichoplax adherens*
Eukaryota	*Chrysochromulina sp._CCMP291*
Eukaryota	*Drosophila melanogaster*
Eukaryota	*Micromonas pusilla_RCC299*
Eukaryota	*Stylonychia lemnae*
Eukaryota	*Thecamonas trahens*
Bacteria	*Actinomyces graevenitzii_435830*
Bacteria	*Bacteroides clarus_762984*
Bacteria	*Bacillus subtilis_1147161*
Bacteria	*Chlorobaculum parvum_517417*
Bacteria	*Eubacterium cellulosolvens_633697*
Bacteria	*Escherichia coli_574521*
Bacteria	*Erysipelothrix rhusiopathiae_650150*
Bacteria	*Flavobacterium columnare_1041826*
Bacteria	*Fervidobacterium nodosum_381764*
Bacteria	*Helicobacter pylori_85963*
Bacteria	*Leptospira interrogans_189518*

Bacteria	*Listeria monocytogenes_1639*
Bacteria	*Myxococcus xanthus_246197*
Bacteria	*Neisseria meningitidis_662598*
Bacteria	*Nostoc sp._103690*
Bacteria	*Paenibacillus sp._481743*
Bacteria	*Pedobacter saltans_762903*
Bacteria	*Rickettsia typhi_257363*
Bacteria	*Rubrobacter xylanophilus_266117*
Bacteria	*Sulfurihydrogenibium azorense_204536*
Bacteria	*Streptomyces coelicolor_100226*
Bacteria	*Synechococcus elongatus_269084*
Bacteria	*Streptococcus pneumoniae_373153*
Bacteria	*Tistrella mobilis_1110502*
Bacteria	*Thermomicrobium roseum_309801*
Bacteria	*Thermus scotoductus_743525*
Bacteria	*Thermoanaerobacter siderophilus_880478*
Bacteria	*Veillonella ratti_883156*
Archaea	*Aciduliprofundum boonei_439481*
Archaea	*Acidianus hospitalis_933801*
Archaea	*Candidatus haloredivivus_1072681*
Archaea	*Candidatus korarchaeum_374847*
Archaea	*Candidatus nitrososphaera_gargensis_ga9*
Archaea	*Candidatus nitrosoarchaeum_1001994*
Archaea	*Cenarchaeum symbiosum_414004*
Archaea	*Ferroglobus placidus_589924*
Archaea	*Haloferax volcanii*
Archaea	*Methanosarcina acetivorans*
Archaea	*Methanocella arvoryzae*
Archaea	*Methanoregula boonei_456442*
Archaea	*Methanopyrus kandleri_190192*
Archaea	*Methanosaeta concilii*
Archaea	*Methanocaldococcus sp.*
Archaea	*Methanocorpusculum labreanum*
Archaea	*Methanothermococcus okinawensis*
Archaea	*Methanosphaera stadtmanae_339860*
Archaea	*Methanothermobacter thermautotrophicus*
Archaea	*Natrialba magadii*
Archaea	*Pyrobaculum calidifontis*
Archaea	*Pyrolobus fumarii*

Archaea	*Pyrococcus horikoshii*
Archaea	*Picrophilus torridus*
Archaea	*Sulfolobus acidocaldarius_330779*
Archaea	*Staphylothermus hellenicus*
Archaea	*Thaumarchaeota archaeon_1198115*
Archaea	*Thermococcus litoralis*
Archaea	*Thermoproteus tenax_768679*
Archaea	*Vulcanisaeta distributa*

Article

Current Biology

A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals

Highlights

- Largest and most internally consistent metazoan-scale superalignment to date
- Sponges (Porifera) are sister-group to all other multicellular animals
- Previous findings of trees with "Ctenophora-sister" were due to artifacts
- Analyses of multigene datasets should employ site-heterogeneous evolutionary models

Authors

Paul Simion, Hervé Philippe, Denis Baurain, ..., Nicole King, Gert Wörheide, Michaël Manuel

Correspondence

herve.philippe@sete.cnrs.fr (H.P.), michael.manuel@upmc.fr (M.M.)

In Brief

Simion et al. demonstrate that sponges (Porifera) are the earliest branching animal lineage, using a combination of 1,719 genes that outperforms in size and quality previous datasets used to address metazoan relationships. Previous findings of comb jellies sister to other animals were likely due to an artifact known as "long branch attraction."

Simion et al., 2017, Current Biology *27*, 958–967
April 3, 2017
http://dx.doi.org/10.1016/j.cub.2017.02.031

CellPress

CellPress

Current Biology
Article

A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals

Paul Simion,[1,17,18] Hervé Philippe,[2,3,17,*] Denis Baurain,[4] Muriel Jager,[1] Daniel J. Richter,[5,6,7] Arnaud Di Franco,[2] Béatrice Roure,[2,3] Nori Satoh,[8] Éric Quéinnec,[1] Alexander Ereskovsky,[9,10] Pascal Lapébie,[11] Erwan Corre,[12,13] Frédéric Delsuc,[14] Nicole King,[5] Gert Wörheide,[15,16] and Michaël Manuel[1,19,*]

[1]Sorbonne Universités, UPMC Univ Paris 06, CNRS, Evolution Paris-Seine UMR7138, Institut de Biologie Paris-Seine, Case 05, 7 quai St Bernard, 75005 Paris, France

[2]Centre de Théorisation et de Modélisation de la Biodiversité, Station d'Ecologie Théorique et Expérimentale, UMR CNRS 5321, Moulis 09200, France

[3]Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, Montréal, QC H3C 3J7, Canada

[4]InBios–Eukaryotic Phylogenomics, Department of Life Sciences and PhytoSYSTEMS, University of Liège, Bât. B22, Quartier Vallée 1, Chemin de la Vallée 4, 4000 Liège, Belgium

[5]Howard Hughes Medical Institute and Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720-3200, USA

[6]CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France

[7]Sorbonne Universités, Université Pierre et Marie Curie (UPMC) Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France

[8]Marine Genomics Unit, Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa 904-0495, Japan

[9]Aix Marseille Univ, Univ Avignon, CNRS, IRD, IMBE, Marseille, France, Chemin de la Batterie des Lions, 13007 Marseille, France

[10]Department of Embryology, Faculty of Biology, Saint-Petersburg State University, Universitetskaya nab. 7/9, Saint-Petersburg 199034, Russia

[11]Sorbonne Universités, UPMC Univ Paris 06, and CNRS, Laboratoire de Biologie du Développement de Villefranche-sur-mer, Observatoire Océanographique, 06230 Villefranche-sur-mer, France

[12]CNRS, FR2424, ABiMS, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France

[13]Sorbonne Universités, Université Pierre et Marie Curie (UPMC) Paris 06, FR2424, ABiMS, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France

[14]Université de Montpellier (UM), Institut des Sciences de l'Évolution (ISEM), UMR 5554 CNRS-IRD-EPHE, Case Courier 64, Place Eugène Bataillon, 34095 Montpellier, France

[15]Department of Earth and Environmental Sciences & GeoBio-Center, Ludwig-Maximilians-Universität München, Richard-Wagner-Str. 10, 80333 München, Germany

[16]SNSB - Bayerische Staatssammlung für Paläontologie und Geologie, Richard-Wagner-Str. 10, 80333 München, Germany

[17]Co-first author

[18]Present address: Université de Montpellier (UM), Institut des Sciences de l'Évolution (ISEM), UMR 5554 CNRS-IRD-EPHE, Case Courier 64, Place Eugène Bataillon, 34095 Montpellier, France

[19]Lead Contact

*Correspondence: herve.philippe@sete.cnrs.fr (H.P.), michael.manuel@upmc.fr (M.M.)

http://dx.doi.org/10.1016/j.cub.2017.02.031

SUMMARY

Resolving the early diversification of animal lineages has proven difficult, even using genome-scale datasets. Several phylogenomic studies have supported the classical scenario in which sponges (Porifera) are the sister group to all other animals ("Porifera-sister" hypothesis), consistent with a single origin of the gut, nerve cells, and muscle cells in the stem lineage of eumetazoans (bilaterians + ctenophores + cnidarians). In contrast, several other studies have recovered an alternative topology in which ctenophores are the sister group to all other animals (including sponges). The "Ctenophora-sister" hypothesis implies that eumetazoan-specific traits, such as neurons and muscle cells, either evolved once along the metazoan stem lineage and were then lost in sponges and placozoans or evolved at least twice independently in Ctenophora and in Cnidaria + Bilateria. Here, we report on our reconstruction of deep metazoan relationships using a 1,719-gene dataset with dense taxonomic sampling of non-bilaterian animals that was assembled using a semi-automated procedure, designed to reduce known error sources. Our dataset outperforms previous metazoan gene superalignments in terms of data quality and quantity. Analyses with a best-fitting site-heterogeneous evolutionary model provide strong statistical support for placing sponges as the sister-group to all other metazoans, with ctenophores emerging as the second-earliest branching animal lineage. Only those methodological settings that exacerbated long-branch attraction artifacts yielded Ctenophora-sister. These results show that methodological issues must be carefully addressed to tackle difficult phylogenetic questions and pave the road to a better understanding of how fundamental features of animal body plans have emerged.

INTRODUCTION

The question of how animal-specific cell types and key animal body plan features first evolved cannot be answered without a clear understanding of the phylogenetic relationships among major animal lineages. Analyses of large-scale molecular superalignments assembled from genomic or transcriptomic data have failed thus far to provide a widely accepted consensus for the branching order among the five major animal lineages, i.e., sponges, placozoans, ctenophores, cnidarians, and bilaterians [1–12]. As a consequence, controversy prevails concerning early animal evolution.

Disagreement mainly hinges on the contradictory phylogenetic placements of ctenophores and sponges in different phylogenomic studies: either with sponges branching first ("Porifera-sister") and ctenophores grouping with cnidarians and bilaterians (consistent with the classical view of a single origin of neurons in the eumetazoan stem lineage) [1–4, 12], or instead with ctenophores as the first offshoot of the animal tree ("Ctenophora-sister") and sponges branching second [5–10]. The Ctenophora-sister hypothesis implies that the nerveless and morphologically much simpler sponges and placozoans are unexpectedly more closely related to cnidarians and bilaterians than are ctenophores, and has noticeably fuelled an intense debate around the possibility of two independent acquisitions for neurons and synapses [13–17]. Comparative analysis of gene content has also been proposed in support of Ctenophora-sister [7], but upon improvement of the inference method, the same data supported Porifera-sister instead [4].

Incongruence between phylogenies can arise from a number of sources, including the use of alignments that are flawed in some way (e.g., contaminated, poorly aligned), the inclusion of sequences that do not faithfully record the organismal phylogeny (e.g., because of lateral gene transfer or paralogy), and the use of inappropriate models of sequence evolution. Ctenophores have a high rate of molecular evolution, making them a priori difficult to place due to their potential for long branch attraction (LBA) artifacts [12]. LBA artifacts can result in the erroneous grouping of unrelated lineages due to their unusually high substitution rates, including the branching of a molecularly highly divergent lineage near the base of the tree due to artifactual attraction by distant outgroups. The conundrum of correctly placing ctenophores and sponges in the animal tree of life is stimulating not only due to its important implications for understanding early animal evolution, but also because it has prompted researchers to re-examine the sources of contradiction between phylogenomic studies.

We assembled an entirely new phylogenomic dataset designed to address relationships between early-diverging metazoan lineages. Our superalignment of 1,719 genes was constructed using a novel multi-step procedure devised to integrate knowledge accumulated in recent years about the various potential causes of artifacts and conflicts in phylogenomics. Analyses of this dataset using the site-heterogenous CAT model provide unambiguous support for the Porifera-sister hypothesis. Ctenophores emerge as the sister group to a clade containing placozoans + cnidarians + bilaterians. Ctenophora-sister was recovered only in analyses containing limited sampling of sponge classes and/or using suboptimal models of sequence evolution, strongly suggesting that it is an artifact of long branch attraction.

RESULTS AND DISCUSSION

A New Pan-metazoan Phylogenomic Dataset of Unprecedented Quality and Size

Phylogenomic datasets previously assembled to address non-bilaterian relationships have often contained a substantial amount of data error (both biological and in silico contamination, alignment errors, and false orthology, see below) and/or the sequences included did not contain enough phylogenetic signal to provide a statistically robust resolution to the problem (see [4, 12]). To tackle these limitations, we developed and implemented a new semi-automated pipeline (i.e., automated procedures supplemented with stringent manual controls of intermediate results) to comprehensively detect and eliminate as many data errors as possible (see Figure 1 for a graphical summary of our pipeline, Supplemental Experimental Procedures for details, and Figure S1 for examples of errors in published datasets constructed with less stringent automated procedures; an example showing trees for a given gene at all successive steps of our filtering procedure is provided at https://github.com/psimion/SuppData_Metazoa_2017). The goal was to simultaneously optimize taxonomic sampling, data quantity (gene number), and data quality. Viewed from this perspective, we therefore reconcile the two principal differing operational philosophies that have thus far competed in the field of phylogenomics: (1) reliance on a limited number of established gene alignments where potential sources of error can be manually curated (e.g., [1, 12]) and (2) entirely automated construction and limited quality control of hundreds of gene alignments (orthology groups) (e.g., [5–10]). The first (manual) approach is not scalable to thousands of genes while the second (automated) approach has not, until now, satisfactorily addressed all sources and types of error (see examples in Figures S1A–S1D).

The resulting dataset comprises 1,719 genes and 97 species (with 39.3% missing data), including 61 non-bilaterian metazoan species. For 21 of these 61 species, we produced new transcriptome assemblies as part of this study (see Supplemental Experimental Procedures for details of taxon sampling). This new dataset outperforms other previously published metazoan phylogenomic supermatrices both in terms of size (total number of amino acid residues) and quality (internal congruence, as estimated from the mean percentage of recovery of clades in single-gene phylogenies that are present in the species tree reconstructed from the supermatrix of concatenated genes) (Figure 2). It contains from two to ten times more information than other datasets and displays a level of congruence (60%) that exceeds that of a manually curated dataset supporting Porifera-sister [1] (57%) and of all automatically assembled datasets that have yielded trees in favor of Ctenophora-sister [5–10] (average of 39%). Our protocol for supermatrix construction therefore appears to be better suited for eliminating data errors than those that have been previously used.

Sponges Are Sister to All Other Metazoans

The supermatrix was analyzed in a Bayesian framework using the site-heterogeneous CAT model, which was originally conceived to minimize LBA artifacts by taking into account the observation that only a limited number of amino acids are functionally acceptable at a given position [18]. Cross-validation experiments, a

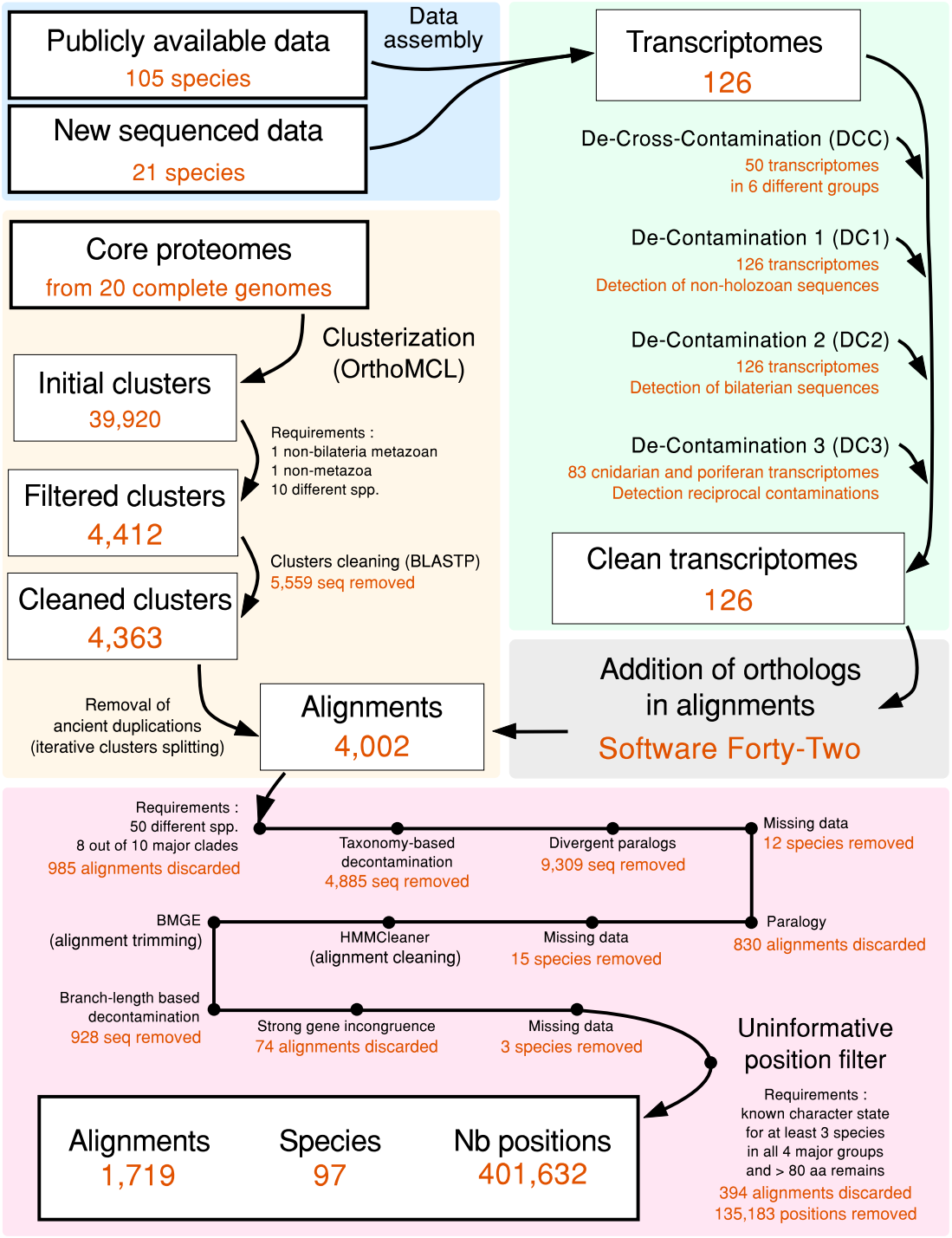


Figure 1. Graphical Summary of the Dataset Construction Protocol Used in This Study
See Experimental Procedures and Supplemental Experimental Procedures for the details of each step.

well-established statistical method to evaluate model fit [19], showed that the fit of the CAT model to the data used in this study is superior to that of site-homogeneous models (ΔlnL = 2,314 ± 164 compared to LG and 1,956 ± 154 compared to GTR), in agreement with previous studies (e.g., [3, 4]). A recent study using simulated data suggested that the CAT model might be less accurate than site-homogeneous models (e.g., LG) under some circumstances [20]. However, the biological relevance of these simulations has not yet been thoroughly explored. In particular, the CAT model appears to fit less well to these simulated data

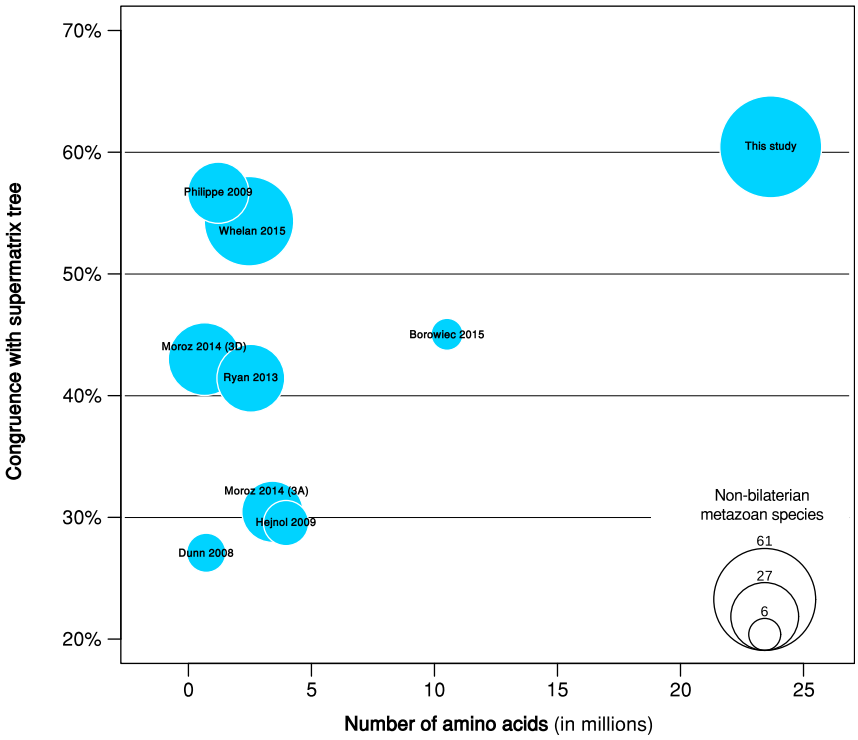


Figure 2. Comparison of Data Quantity and Quality between Nine Recent Phylogenomic Datasets

Quantity of molecular information (x axis) corresponds to the number of amino acids in the supermatrix, and data quality (y axis) corresponds to the percentage of bipartitions (internal tree branches) that are identical in both the single-gene trees and the supermatrix tree (internal congruence); only genes longer than 200 amino acids are represented (see https://github.com/psimion/SuppData_Metazoa_2017 for the corresponding plot with the remaining genes). Circled areas are proportional to the number of non-bilaterian metazoan species present in the datasets, indicated with abbreviated reference information. For each study included here, we analyzed the largest available dataset, except for Moroz et al. [8] where we analyzed both the largest one ("3A" dataset) as well as a smaller one enriched in ctenophores ("3D" dataset). Note that we also analyzed reduced datasets from Whelan et al. [10] (i.e., their datasets 6, 10, and 16, results not shown), which yielded congruence results almost identical to those of their dataset 1. See also Figure S1.

than do site-homogeneous models (H.P., unpublished results), while the opposite is true for most real datasets.

Because of the large size of our supermatrix, it was not computationally feasible to analyze the whole dataset using the CAT model. To circumvent this limitation, we used a gene jackknife strategy based on 100 separate analyses, each involving a random selection of roughly 25% of the genes in our dataset (i.e., about 100,000 amino acid positions per replicate). This strategy allowed us to combine the advantage of reduced computational burden with reliable estimates of the statistical robustness of each clade. Furthermore, jackknifing genes counteracts potentially strongly misleading signals that might be contained in a small number of genes. These analyses (Figure 3) revealed that monophyletic sponges (Jackknife Support [JS] of 100%) emerge as the sister group to all other metazoans (Porifera-sister) with ctenophores as the second diverging animal lineage (JS 95%), followed by placozoans, which are the sister group to a cnidarian + bilaterian clade (JS 100%). Since the inclusion of distant outgroups is known to amplify LBA artifacts [21–24], the tree of Figure 3 was rooted using only choanoflagellates, the closest living relatives of metazoans. In addition, an analysis including a more complete holozoan sampling also supports Porifera-sister; hence, this result does not depend on outgroup sampling (Figure S2A). Sponges were monophyletic in all analyses, and relationships within Porifera, Ctenophora, Cnidaria, and Bilateria were generally strongly supported and consistent with other molecular studies [25–27].

In previous studies, the Porifera-sister hypothesis has tended to be better supported by smaller, manually constructed and curated phylogenomic datasets [1–3], whereas datasets featuring many more genes and assembled using entirely automated procedures have tended to support Ctenophora-sister (e.g., [5–10]). This has fuelled the idea that increasing gene sampling and taxon quantity were the drivers of higher support for Ctenophora-sister [28]. Recent re-analyses of several of these datasets have cautioned, however, that this support vanishes once the effects of outgroup sampling and model choice (site-heterogeneous versus site-homogeneous) are simultaneously taken into account [4]. In this study, a multigene dataset generated using semi-automated procedures, including data quality controls of unparalleled stringency, and containing the largest amount of molecular information and taxonomic representation used to date in metazoan phylogenomics, yielded strong statistical support for sponges rather than ctenophores as the sister-group to all other animals. This result is consistent with previous propositions [4, 12] that Ctenophora-sister stems from an LBA artifact due to the use of poorly fitting evolutionary models that lead to statistical inconsistency (LBA being a form of systematic error) when analyzing large gene numbers.

Drastic Effects of Taxon Sampling and Site-Homogeneous versus Site-Heterogeneous Model Type on the Placement of Long Branches

To assess the impact of using a less well-fitting site-homogeneous model of sequence evolution (LG) versus a better-fitting site-heterogeneous model (CAT) on the placement of long branches in the phylogeny, we examined the behavior of the two metazoan lineages having the highest rate of substitution in our dataset, ctenophores and hexactinellid sponges, when other sponges (the closest relatives of hexactinellids) were either included or excluded from the dataset. With full taxonomic sampling, despite their high substitution rate, hexactinellids were correctly located as the sister group to demosponges [25] by both models (Figures 4A and 4D), probably because the branch that unites these two groups is sufficiently long to overcome

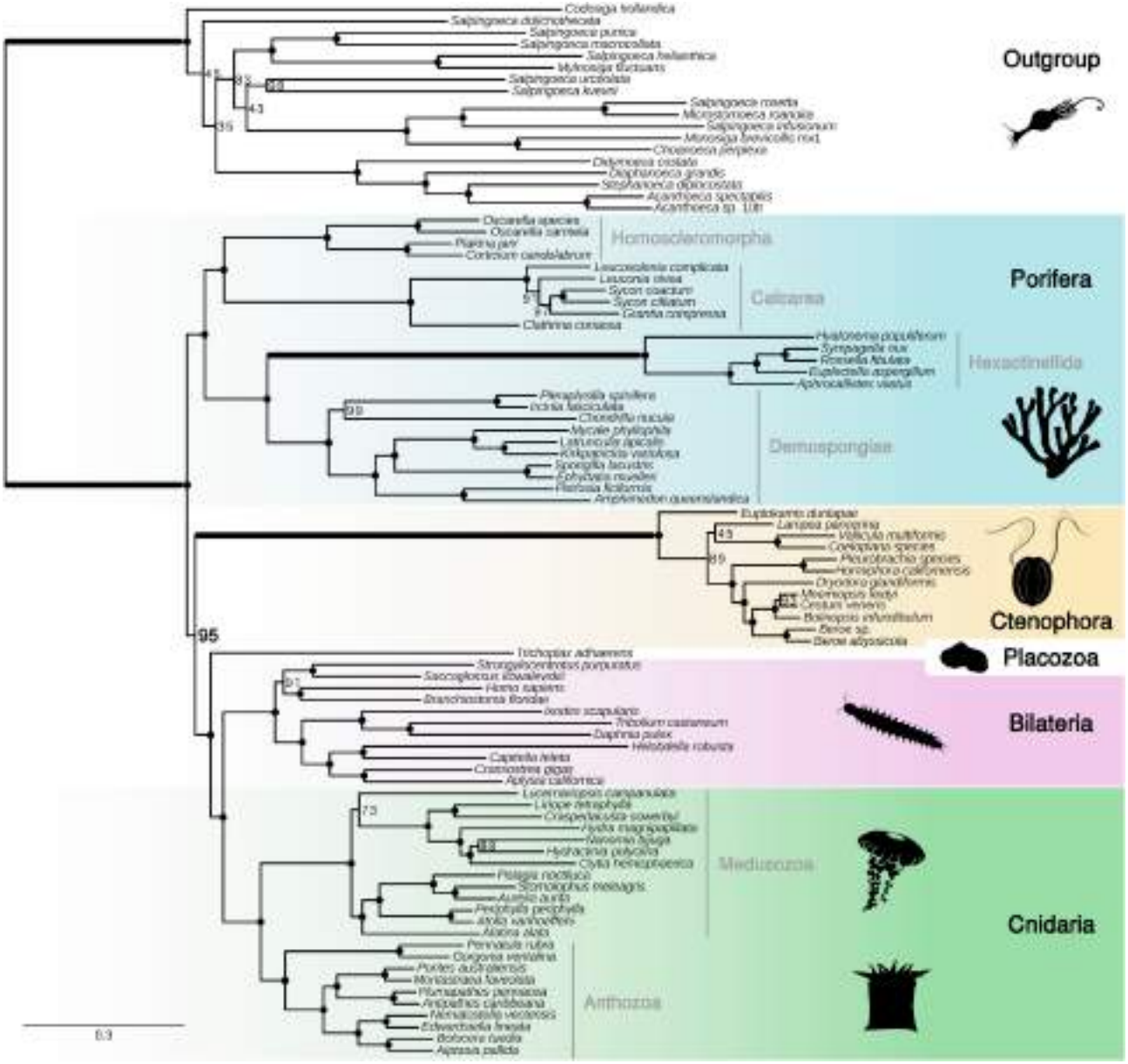


Figure 3. Metazoan Phylogenetic Relationships Inferred from a Supermatrix of 401,632 Amino Acid Positions for 90 Species

Every gene jackknife replicate was composed of ~100,000 positions (each replicate represents a larger dataset than those of any previous phylogenomic study [1, 5–8, 10] except [9]) and was analyzed using PhyloBayes_MPI 1.6j under the site-heterogeneous CAT+ Γ_4 model of sequence evolution. The tree shown here is the consensus of the 100 jackknife replicates and branch support values (JS %) represent the number of analyses in which each branch was recovered; black circles represent nodes with maximal support (100%). The three longest branches in terms of inferred substitutions are highlighted with thicker lines: the branch separating metazoans from outgroups and the terminal branches bearing hexactinellid sponges and ctenophores. The supermatrix had an overall percentage of missing data of 37.3%. Organism drawings were downloaded from the PhyloPic website. See also Figures S2 and S4.

any artifactual signal. We note that the site-homogeneous LG model (Figure 4D) recovered Ctenophora-sister, unlike the site-heterogeneous CAT model (Figure 4A), which recovered Porifera-sister. When demosponges were discarded, hexactinellids remained grouped with the other sponges when the site-heterogeneous CAT model was used (Figure 4B), but with the site-homogeneous LG model they formed a maximally supported clade with ctenophores (bootstrap support [BS] 100%), located at the base of metazoans (Figure 4E). Due to the removal of demosponges, the short internal branch linking hexactinellids to calcareous and homoscleromorph sponges, in combination with the use of a less well-fitting model, was insufficient to counteract the LBA artifact. This represents a quintessential LBA configuration, with the three longest branches (two internal and one external; highlighted with thicker lines in Figure 3) are clustered together. When all other sponges were removed, the

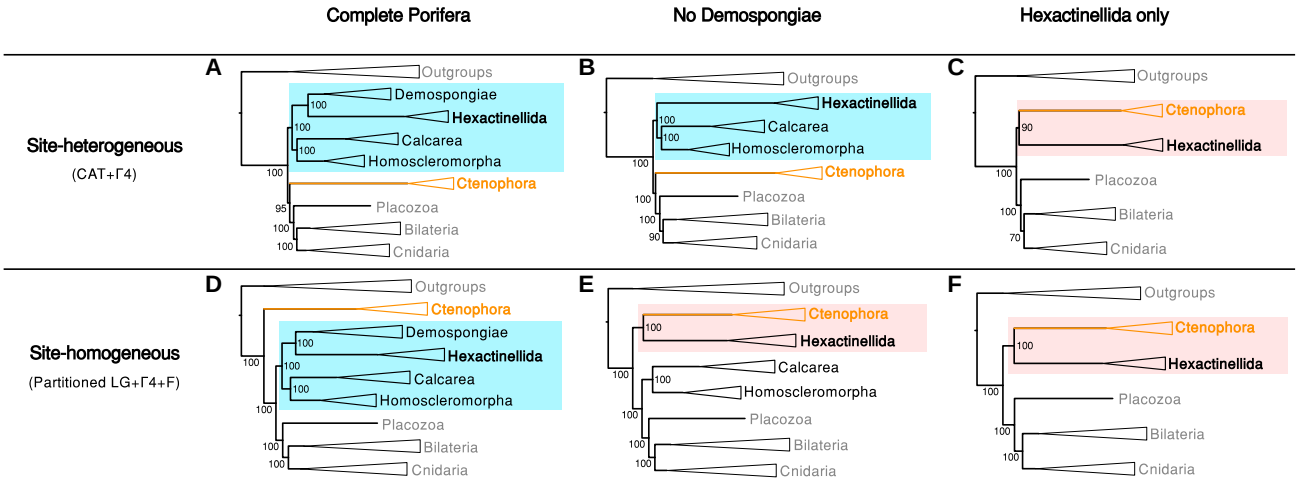


Figure 4. Comparison of the Behavior of Long Branches with Varying Taxon Sampling, under Site-Heterogeneous versus Site-Homogeneous Models

The multigene dataset of this study was analyzed with the CAT+ Γ_4 model (A, B, C) and the LG+ Γ_4+F model (D, E, F) with three different samplings of sponge taxa.
(A and D) Full taxonomic sampling, with all four sponge classes represented.
(B and E) Demosponges removed.
(C and F) Demosponges, calcareous sponges, and homoscleromorphs removed (sponges represented by hexactinellids only).
Node support for (A), (B), and (C) corresponds to jackknife support values (JS %) from 10 jackknife replicates on the complete supermatrix, while node support for (D), (E), and (F) corresponds to bootstrap support values (BS %). Analyses shown in (D) were run with complete and reduced outgroup samplings (see legend of Figure S2B). See also Figure S3 (same comparison performed on another phylogenomic dataset) and Table S1 (same comparison using likelihood computations with additional models of sequence evolution).

CAT analysis also yielded this erroneous placement of hexactinellids (Figure 4C), but with a slightly lower support than LG (JS 90% versus BS 100%, Figures 4C and 4F). This experiment confirms the higher sensitivity of site-homogeneous models to LBA [29] and indicates that the CAT model can also be impacted, albeit to a lesser extent. The same procedure applied to a previously published dataset [10] led to the same results, suggesting that the choice of the model of evolution, and not the dataset per se, is responsible for the effect observed (see Figure S3).

To test whether the CAT model and its Bayesian implementation are necessary to handle the heterogeneity of the evolutionary process across sites, we compared the likelihoods of the topologies of Figure 4 using various models, from purely site-homogeneous ones (LG, WAG) to site-heterogeneous models that are implemented in a maximum likelihood framework but are more restricted than CAT in the levels of pattern heterogeneity they can accommodate (see results in Table S1). Even with the simple removal of only demosponges, all these models yielded an erroneous clade of ctenophores and hexactinellids, although the likelihood difference (ΔlnL) in comparison to the correct topology (hexactinellids grouped with other sponges) was higher for site-homogeneous models and decreased as progressively more complex site-heterogeneous models were used (C20, C40, C60). These observations suggest that the levels of pattern heterogeneity present in empirical sequence alignments cannot be appropriately modeled by empirical finite mixtures and require instead the use of Dirichlet process models, such as CAT.

Implications for Body Plan Evolution

The principal outcome of this study is a rejection of the Ctenophora-sister hypothesis of animal evolution. With the artifactual basal long branch attraction of ctenophores now reduced, Porifera is strongly supported as the sister group of other metazoans. Therefore, the absence in sponges of features such as a gut, neurons, synapses, and muscles is more parsimoniously interpreted as ancestral in metazoans (i.e., plesiomorphic), in line with classical views.

The topology obtained in our analyses (Figure 3) does not fully clarify, however, the pattern of emergence of these features (shared by ctenophores, cnidarians, and bilaterians) within the non-sponge clade, because they are also lacking in placozoans, one of the most simply organized metazoan phyla. Thus, according to our phylogeny, either nervous systems and other advanced features originated independently in ctenophores and in the lineage leading to cnidarians + bilaterians, or placozoans represent the evolutionary loss of features that existed in a more complex ancestor, as has been previously suggested [14, 30], in consistence with the eumetazoan-like gene content of the placozoan genome [31]. Furthermore, "Ctenophora-second" as obtained here (Figure 3) could either represent the correct topology or reflect a trade-off between their attraction toward the base (attenuated but not suppressed with the CAT model) and a more internal true position. This is suggested by analyses in which most heteropecillous sites (i.e., violating CAT model assumptions [32]) were removed: in these trees, ctenophores are sister-group to other eumetazoans (Figure S4A) or to cnidarians (Figure S4B) [1], consistent with a single origin of eumetazoan features including nerve cells.

The anatomical features of ctenophores are of little help for understanding their relationships with other early-diverging animal lineages, because they include numerous unique traits, such as distinctive biradial symmetry, extraordinary macrocilia forming

swimming paddles or combs (arranged in eight rows around the body), a sophisticated aboral neuro-sensory complex without any counterpart in other phyla, and adhesive cells with unparalleled cytological features (colloblasts) (see [33, 34]). The problem of homology or convergence of their nervous system and muscles with those of cnidarians and bilaterians remains open to debate. Furthermore, regardless of its precise phylogenetic position, the ctenophore lineage has evolved a dramatic increase in anatomical complexity, which is paralleled only in bilaterians, and whose genomic and molecular developmental bases remain obscure.

Conclusions

The recent debate about deep metazoan relationships stems from methodological issues that call for improvements in terms of both data curation and inference. Our novel semi-automated protocol yielded a pan-metazoan phylogenomic dataset of greatly increased quality and size relative to other datasets used for reconstructing metazoan phylogenies. Using a stringent gene jackknifing strategy, we obtained strong support for placing sponges as the sister group to all other metazoans. Nonetheless, our observations with different sub-samplings of the sponge classes indicate that the site-heterogeneous CAT model, despite outcompeting any site-homogeneous model, is unsurprisingly not immune to reconstruction artifacts [23, 32].

These considerations call for further improvements in phylogenomics procedures, in order to better handle the tremendous level of data complexity that characterizes supermatrices composed of thousands of genes sampled across many different phyla. In-depth data curation in order to remove contaminants and paralogs is currently a necessity, as they are a major source of inter-gene incongruence, although a promising alternative approach could be to model them in a probabilistic framework (e.g., [35]). In parallel, the pervasive role of epistasis in protein evolution makes modeling of sequence evolution in a site-independent framework (currently a condition for models to remain computationally tractable) problematic. Therefore, determining which violations of model assumptions are the most detrimental to phylogenetic reconstruction accuracy is of primary importance [24]. These violations include heterotachy, heteropecilly, global compositional bias, and relationships among sites due to 3D structural features, among others. Integrating these various properties into models will significantly improve phylogenetic reconstruction and may resolve current controversies regarding deep metazoan relationships, in addition to those in other parts of the tree of life.

EXPERIMENTAL PROCEDURES

A graphical summary of the whole pipeline for dataset construction is provided in Figure 1. Supporting material such as gene alignments, programs, supermatrices and additional phylogenetic trees can be found at https://github.com/psimion/SuppData_Metazoa_2017.

Genome Sampling and Dataset Assembly

We used protein sequences predicted from the complete genomes of 20 selected "core species" (17 metazoans, 2 choanoflagellates, and 1 filasterean) (see Table S4) to create clusters of putative orthologous genes. For this, we used the programs USEARCH [36] (e-value = 1×10^{-5}; accel = 1.0) and OrthoMCL [37] with the default inflation parameter (I = 1.5), which resulted in the creation of a set of 39,920 clusters of putative orthologs. We then selected the 4,412 clusters that contained $\geq$10 different species, including at least one non-bilaterian metazoan and one non-metazoan species (outgroup). Because of the limitations resulting from the use of similarity scores and of single-linkage clustering, OrthoMCL clusters may contain non-homologous sequences. Thus, in a two-step BLAST procedure, we discarded sequences that did not match (e-value $\leq 1 \times 10^{-10}$): (1) $\geq$30% of other sequences in their cluster (2,990 sequences removed), (2) $\geq$50% other sequences in their cluster (2,520 sequences removed). Thinned clusters were aligned using MAFFT [38] (localpair, maxiterate = 5,000) then cleaned of non-homologous stretches via HMMCleaner [39] and of ambiguously aligned positions via BMGE [40]. The resulting alignments were analyzed with RAxML [41] (LG+Γ_4+F model) to yield single-gene trees. To determine whether clusters contained anciently duplicated genes, trees were split on branches (1) separating two subtrees with $\geq$10 different species in each of the subtrees and (2) within the top 10% longest branches of the tree. Clusters that could be split were iteratively reduced into smaller clusters. Finally, we applied the same taxonomic filter as above, which resulted in the generation of 4,002 core orthologous clusters.

RNA Preparation and Sequencing

Biological samples (see Table S2) were carefully cleaned to remove biological contaminants, then powdered in liquid nitrogen. RNA extraction was performed using either the QIAGEN RNeasy Kit (according to the manufacturer's instructions) or a TRIzol-based protocol. In the latter case, frozen sample powder was incubated for 5 min in TRIzol solution, before addition of chloroform for another 15 min of incubation. The solution was then centrifuged for 15 min at 4°C (12,000 × *g*) in order to retain the upper aqueous phase only, which was subsequently incubated for 10 min in isopropanol. Samples were then centrifuged for 10 min at 4°C (12,000 × *g*) and the supernatant eliminated. The pellet was vortexed and centrifuged for 5 min at 4°C (7,500 × *g*) in ethanol 75%. After supernatant elimination, the dried pellet was finally resuspended in RNase-free water. Construction of cDNA libraries and their sequencing using either 454 pyrosequencing or Illumina technology (see Table S2) was carried out at GATC Biotech. 6 of the 22 newly sequenced species were pooled for a single 454 run (group E in Table S3), while 14 others were pooled in two Illumina lanes of the same run (group F in Table S3).

Transcriptome Sampling, Assembly, and Decontamination

We used 126 non-bilaterian species for which sequence data were either publicly available or provided by Daniel Richter, Nicole King, and Nori Satoh (see third column in Table S3). 454 reads were assembled using MIRA alone [42] or a combination of MIRA and CAP3 [43], whereas Illumina reads were assembled using either Trinity [44] or SOAPdenovo-trans [45]. To reduce subsequent computational time, transcripts that did not match any of the 4,002 orthologous clusters (BLAST e-value $\leq 1 \times 10^{-10}$) were discarded, which reduced the total number of transcripts from 12,247,929 to 1,787,422. We then designed a new procedure to detect and remove cross-contaminating sequences between transcriptomic datasets obtained in the same lab, belonging to the same sequencing project or for which cross-contamination issues were observed in preliminary analyses (see details in Supplemental Experimental Procedures). Species transcriptomes were processed in six groups (see "group" column in Table S3), which allowed us to estimate the level of cross-contamination of each species, ranging from 0.16% (*Pleurobrachia pileus*, this study) to 70.64% (ctenophora sp3 A [8]) of the reads. Transcriptomes were further screened for additional contamination sources using a three-step procedure aiming at detecting contaminations at different taxonomical scales: by non-holozoans (DC1), by bilaterians (DC2), and reciprocal contamination between sponges and cnidarians (DC3). Details about decontamination procedures are given in Supplemental Experimental Procedures, and an illustration of their behavior, in the case of the highly expressed ribosomal protein rpl2 (an extreme case of contamination) is provided at https://github.com/psimion/SuppData_Metazoa_2017.

Transcriptomic Data Integration into Orthologous Clusters

Decontaminated transcriptomic data were then incorporated into the 4,002 previously assembled core orthologous clusters using a multiple Best Reciprocal Hit approach implemented in the newly designed Forty-Two software (see details in Supplemental Experimental Procedures). We then discarded

clusters with ≤50 species or ≤8 out of 10 major taxonomic groups (Bilateria, Anthozoa, Medusozoa, Ctenophora, Demospongiae, Hexactinellida, Calcarea, Homoscleromorpha, Placozoa, outgroup), thus retaining only 3,414 enriched orthologous clusters.

Paralogy Treatment and Removal of Contaminants

At this stage, despite considerable efforts to remove ancient paralogs and contaminants, some contaminating sequences or recent paralogs were still present in our alignments. That is why we applied several additional filters, based on BLAST similarity searches or on single-gene phylogenetic trees, to identify and remove them.

1) A genuine sequence from one of the ten major clades defined above should be more similar to other sequences of the same clade than to sequences of any other clade because of the long internal branch defining each of these clades. Each sequence was thus BLASTed against the other sequences of the same cluster (only if they were ≥90% complete on the overlapping part and after discarding positions containing ≤ 10% known character states) and sequences were removed when their best hit belonged to a clade other than the expected one. This step eliminated 4,885 sequences.
2) When multiple sequences from the same species are present in a given cluster, the one(s) that is(are) most similar to sequences from the other species is(are) more likely to be orthologs. Hence, for each species having multiple sequences, each sequence was BLASTed against the rest of the alignment and the best hit identified; a sequence was removed if it overlapped with the best hit sequence by ≥95% and if its BLAST score was below the best hit score by a given threshold. Using first a threshold of 25% and then a threshold of 10%, 21,444 and 4,668 sequences were removed, respectively. The resulting clusters were cleaned using HMMCleaner and the same process was repeated, this time removing 7,030 and then 2,279 additional sequences. Most of these sequences were variants of the same transcripts (due to sequencing errors or to in vivo transcript degradation), whereas the others corresponded to distant paralogs, and very few to previously undetected contaminants.
3) Based on a preliminary supermatrix tree built with RAxML using the LG+Γ_4+F model, 12 cnidarian and poriferan species that were incomplete and very closely related to more complete species were discarded, thereby reducing the number of species to 115. Subsequently, all alignments in which ctenophores were no longer represented were discarded. Finally, all alignments that did not contain ≥50 species in ≥8 out of the 10 major clades (see above) were discarded, leaving 3,176 clusters.
4) Paralogous genes in these 3,176 putatively orthologous clusters were discarded in two steps (see details in Supplemental Experimental Procedures): (1) only the 2,424 alignments with at most two of the previously defined major taxonomic groups affected by paralogy were conserved and (2) for each major taxonomic group affected by paralogy in these remaining alignments, we selected the largest set of species without out-paralogy.
5) We further eliminated 15 additional species that were incomplete and very closely related to more complete species, thereby reducing the number of species to 100. All clusters were re-aligned with MAFFT (same parameters as above) and applying the same taxonomic filter as above led us to retain 2,187 clusters.
6) Our last quality check was based on the rationale that non-orthologous sequences (being either a contaminant or a paralog) usually display very long branches when constrained on the species tree because they are misplaced. First, alignments were cleaned with HMMCleaner and BMGE and concatenated using SCaFoS [46]. The phylogeny inferred using RAxML from the supermatrix under the LG+Γ_4+F model was considered as a proxy of the species tree (note that ctenophores were sister to all other metazoans in this tree). Then, for each alignment, the reference topology was pruned of the species missing in that alignment, and branch lengths on this constrained topology were estimated using RAxML (LG+Γ_4+F model). This allowed us to compare terminal branch lengths observed in the single-gene tree to those observed in the pruned supermatrix tree and to remove sequences for which the branch-length ratio was >5, thereby eliminating 928 individual sequences. We repeated the same protocol, now computing the Pearson correlation coefficient R^2 between branch lengths in each single-gene tree and the corresponding pruned supermatrix tree. We obtained a mean R^2 of 0.797 and a standard deviation (SD) of 0.090, which led us to remove 74 clusters showing a R^2 outside the interval [0.6209, 0.9730] (i.e., the mean ± 1.96 SD). These included, for instance, a cluster in which a gene from a bacterium used for choanoflagellate culture was present in the transcriptomes of two closely related choanoflagellates.
7) Since missing data increases computational time and LBA artifacts [47], we removed three species that had >85% missing data. More importantly, to retain only genes that potentially bear phylogenetic information on the relative position of Ctenophora and Porifera, we now defined four major groups (Bilateria+Cnidaria+Placozoa, Ctenophora, Porifera, and outgroups) and removed positions that did not have a determined amino acid for ≥3 species in each of these four groups. Last, an alignment was discarded if its length was below 80 amino acid positions, leading to a final set of 1,719 orthologous gene clusters.

Phylogenetic Analyses

Supermatrix Construction

We used SCaFoS [46] to assemble the supermatrix, build chimeras of closely related species (Table S3), and retain only the slowest-evolving sequence when multiple copies were available for a given species (using Tree-Puzzle and the WAG+F model [48] to compute distances). This produced a supermatrix containing 401,632 amino acid positions for 97 species, with an overall amount of 39.3% missing data. A reduced sampling in which distant outgroups were removed resulted in a supermatrix with 90 species and an overall amount of 37.3% missing data.

Evaluation of Congruence

Nine phylogenomic datasets (our dataset and [1, 5–10]) were evaluated for their internal congruence. Single-gene trees were inferred with RAxML [49] under the LG+Γ_4 model, after discarding species with >50% missing data. The corresponding supermatrix trees were either retrieved from the original publications when possible or computed as for single-gene trees. For each gene, missing species were removed from the supermatrix tree, and the percent of bipartitions in agreement between each single-gene tree and its pruned supermatrix was computed. Lastly, we computed the mean percent of bipartition in agreement across single-gene trees for each dataset.

Model Testing

Bayesian cross-validation [50] implemented in PhyloBayes 3.3 [51] was used to compare the fit of the site-homogeneous LG and GTR models and of the site-heterogeneous CAT model. Ten replicates were considered, each one consisting of a random subsample of 10,000 sites for training the model and 2,000 sites for computing the cross-validation likelihood score.

Site-Homogeneous Model

Maximum likelihood analyses were run on the full dataset (i.e., 401,632 amino acid positions) using RAxML [49]. A partition was attributed to each gene and each partition was given an independent LG+Γ_4+F model. Such a partitioned analysis allows the alpha parameter of the gamma distribution and stationary frequencies of amino acids to vary across genes. See Supplemental Experimental Procedures for details on differences between site-homogeneous and site-heterogeneous models.

Site-Heterogeneous Model

Since the better fitting site-heterogeneous CAT model is very time consuming, we analyzed the dataset using a jackknifing strategy with 100 replicates. Each replicate was built by randomly sampling genes until >100,000 positions were obtained (equivalent to ~430 genes per replicate). Jackknife replicates (cleared of constant sites) were analyzed with PhyloBayes MPI 1.6j [52] under the CAT+Γ_4 model, until 6,000 cycles were obtained. Convergence of the parameters was assessed using criteria given in the PhyloBayes manual and a conservative burn-in of 3,000 cycles was used for all replicates.

Computation of Likelihoods for Various Models and Topologies

The relative fits of various available sequence evolution models, including several that aim to model site heterogeneity, were computed on four topologies (see Table S1) with iqtree [53]. This was done for three different taxon

samplings in order to observe the impact of progressive removal of poriferan clades on the position of hexactinellids and ctenophores. The complete taxon sampling corresponds to our 90 species supermatrix.

Removal of Heteropecillous Positions

Since heteropecillous positions (i.e., sites with a substitution process that is heterogeneous in time) violate the assumptions of the CAT model, they may lead to systematic error [32]. In order to account for this, we used the protocol of Roure and Philippe [32] to compute the level of heteropecilly of each position using five pre-defined clades (Choanoflagellatea, Porifera, Ctenophora, Cnidaria, and Bilateria). We then used the CAT+Γ_4 model to analyze the datasets obtained after removal of 60% and 70% of the most heteropecillous positions (136,618 and 102,464 remaining variable positions, respectively; Figure S4).

Testing the Impact of Compositional Bias

In order to reduce potential impact of saturation and compositional bias, we recoded our supermatrix using the Dayhoff 6-states alphabet corresponding to amino acid groups [54, 55], which we then analyzed with the CAT+Γ_4 and CAT+Γ_4+GTR models. This recoding did not affect in any way the deep metazoan relationships inferred in this study, as sponges were always recovered with maximal support (PP = 1) as sister group to all other metazoans, although slight incongruences within choanoflagellates hampered topological convergence of our replicates (data not shown).

Example Analyses of Data Errors in Previous Phylogenomic Datasets

Trees inferred with RAxML [49] under the LG+Γ_4 model from the single genes of ref [7–10] were manually scanned, and one for each study was arbitrarily selected to illustrate the occurrence of erroneous groupings. The original alignments were enriched with data from GenBank (nr) or transcriptomic datasets (retrieved from the NCBI portal) to improve taxonomic sampling and therefore reveal contaminations and/or paralogy. The same positions as in the original studies were selected and trees were inferred with RAxML [49] under the LG+Γ_4 model (Figure S1).

ACCESSION NUMBERS

The data are available under BioProject PRJNA316185 at SRA (accession number SRP072932).

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures, four tables, and Supplemental Experimental Procedures and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2017.02.031.

AUTHOR CONTRIBUTIONS

P.S., H.P., and M.M. designed the study. P.S., M.M., E.Q., A.E., P.L., D.J.R., N.S., and N.K. collected samples. P.S., M.J., and D.J.R. prepared RNA for sequencing. P.S., D.B., F.D., and D.J.R. assembled the transcriptomes. H.P. and P.S. conceived the protocol of data supermatrix assembly including data quality controls, with contribution from M.M.; D.B. wrote the Forty-Two software and H.P. and A.D. debugged it. H.P., P.S., B.R., and D.B. wrote the scripts for the various data quality controls; H.P. and P.S. built the supermatrix. H.P., P.S., G.W., and E.C. performed the phylogenetic analyses. P.S. made the figures and tables. M.M. drafted the manuscript main text and P.S. the Experimental Procedures; all other authors amended the manuscript and approved the final version.

ACKNOWLEDGMENTS

We thank Alain Goyeau's diving team for their help collecting antipatharians in Guadeloupe. We thank the UPMC biological stations of Banyuls-sur-Mer and Villefranche-sur-Mer, as well as Olivier Gros (Université des Antilles, Pointe-à-Pitre) for providing lab facilities. Funding was mainly from the Institut Universitaire de France (M.M. junior membership 2009–2014) and from the TULIP Laboratory of Excellence (ANR-10-LABX-41) to H.P. Computations were made on the supercomputers Mp2 and Ms2 from the Université de Sheerbrooke, managed by Calcul Québec and Compute Canada. The operation of this supercomputer is funded by the Canada Foundation for Innovation (CFI), the ministère de l'Économie, de la science et de l'innovation du Québec (MESI), and the Fonds de recherche du Québec - Nature et technologies (FRQ-NT). The Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities also provided some computational resources. D.J.R. was supported by a National Defense Science and Engineering Graduate fellowship from the United States Department of Defense, a National Science Foundation Central Europe Summer Research Institute Fellowship, a Chang-Lin Tien Fellowship in Environmental Sciences and Biodiversity, a postdoctoral fellowship from the Conseil Régional de Bretagne, and the French Government "Investissements d'Avenir" program OCEANOMICS (ANR-11-BTBR-0008). G.W. was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft [DFG]) and the Ludwig-Maximilians-Universität München LMUexcellent program (Project MODELSPONGE) through the German Excellence Initiative. This is publication ISEM 2017-043 of the Institut des Sciences de l'Evolution de Montpellier.

Received: December 2, 2016
Revised: February 7, 2017
Accepted: February 13, 2017
Published: March 16, 2017

REFERENCES

1. Philippe, H., Derelle, R., Lopez, P., Pick, K., Borchiellini, C., Boury-Esnault, N., Vacelet, J., Renard, E., Houliston, E., Quéinnec, E., et al. (2009). Phylogenomics revives traditional views on deep animal relationships. Curr. Biol. *19*, 706–712.
2. Pick, K.S., Philippe, H., Schreiber, F., Erpenbeck, D., Jackson, D.J., Wrede, P., Wiens, M., Alié, A., Morgenstern, B., Manuel, M., and Wörheide, G. (2010). Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. Mol. Biol. Evol. *27*, 1983–1987.
3. Nosenko, T., Schreiber, F., Adamska, M., Adamski, M., Eitel, M., Hammel, J., Maldonado, M., Müller, W.E., Nickel, M., Schierwater, B., et al. (2013). Deep metazoan phylogeny: when different genes tell different stories. Mol. Phylogenet. Evol. *67*, 223–233.
4. Pisani, D., Pett, W., Dohrmann, M., Feuda, R., Rota-Stabelli, O., Philippe, H., Lartillot, N., and Wörheide, G. (2015). Genomic data do not support comb jellies as the sister group to all other animals. Proc. Natl. Acad. Sci. USA *112*, 15402–15407.
5. Dunn, C.W., Hejnol, A., Matus, D.Q., Pang, K., Browne, W.E., Smith, S.A., Seaver, E., Rouse, G.W., Obst, M., Edgecombe, G.D., et al. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. Nature *452*, 745–749.
6. Hejnol, A., Obst, M., Stamatakis, A., Ott, M., Rouse, G.W., Edgecombe, G.D., Martinez, P., Baguñà, J., Bailly, X., Jondelius, U., et al. (2009). Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc. Biol. Sci. *276*, 4261–4270.
7. Ryan, J.F., Pang, K., Schnitzler, C.E., Nguyen, A.D., Moreland, R.T., Simmons, D.K., Koch, B.J., Francis, W.R., Havlak, P., Smith, S.A., et al.; NISC Comparative Sequencing Program (2013). The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. Science *342*, 1242592.
8. Moroz, L.L., Kocot, K.M., Citarella, M.R., Dosung, S., Norekian, T.P., Povolotskaya, I.S., Grigorenko, A.P., Dailey, C., Berezikov, E., Buckley, K.M., et al. (2014). The ctenophore genome and the evolutionary origins of neural systems. Nature *510*, 109–114.
9. Borowiec, M.L., Lee, E.K., Chiu, J.C., and Plachetzki, D.C. (2015). Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. BMC Genomics *16*, 987.
10. Whelan, N.V., Kocot, K.M., Moroz, L.L., and Halanych, K.M. (2015). Error, signal, and the placement of Ctenophora sister to all other animals. Proc. Natl. Acad. Sci. USA *112*, 5773–5778.
11. Chang, E.S., Neuhof, M., Rubinstein, N.D., Diamant, A., Philippe, H., Huchon, D., and Cartwright, P. (2015). Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. Proc. Natl. Acad. Sci. USA *112*, 14912–14917.

12. Philippe, H., Brinkmann, H., Lavrov, D.V., Littlewood, D.T., Manuel, M., Wörheide, G., and Baurain, D. (2011). Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. *9*, e1000602.
13. Moroz, L.L. (2015). The genealogy of genealogy of neurons. Commun. Integr. Biol. *7*, e993269.
14. Ryan, J.F., and Chiodin, M. (2015). Where is my mind? How sponges and placozoans may have lost neural cell types. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*, 370.
15. Marlow, H., and Arendt, D. (2014). Evolution: ctenophore genomes and the origin of neurons. Curr. Biol. *24*, R757–R761.
16. Jékely, G., Paps, J., and Nielsen, C. (2015). The phylogenetic position of ctenophores and the origin(s) of nervous systems. Evodevo *6*, 1.
17. Leys, S.P. (2015). Elements of a 'nervous system' in sponges. J. Exp. Biol. *218*, 581–591.
18. Lartillot, N., and Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. *21*, 1095–1109.
19. Smyth, P. (2000). Model selection for probabilistic clustering using cross-validated likelihood. Stat. Comput. *10*, 63–72.
20. Whelan, N.V., and Halanych, K.M. (2016). Who let the CAT out of the bag? Accurately dealing with substitutional heterogeneity in phylogenomic analyses. Syst. Biol. Published online September 14, 2016. http://dx.doi.org/10.1093/sysbio/syw084.
21. Philippe, H., Lartillot, N., and Brinkmann, H. (2005). Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol. Biol. Evol. *22*, 1246–1253.
22. Schneider, A., and Cannarozzi, G.M. (2009). Support patterns from different outgroups provide a strong phylogenetic signal. Mol. Biol. Evol. *26*, 1259–1272.
23. Gouy, R., Baurain, D., and Philippe, H. (2015). Rooting the tree of life: the phylogenetic jury is still out. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*, 20140329.
24. Philippe, H., and Roure, B. (2011). Difficult phylogenetic questions: more data, maybe; better methods, certainly. BMC Biol. *9*, 91.
25. Wörheide, G., Dohrmann, M., Erpenbeck, D., Larroux, C., Maldonado, M., Voigt, O., Borchiellini, C., and Lavrov, D.V. (2012). Deep phylogeny and evolution of sponges (phylum Porifera). Adv. Mar. Biol. *61*, 1–78.
26. Simion, P., Bekkouche, N., Jager, M., Quéinnec, E., and Manuel, M. (2015). Exploring the potential of small RNA subunit and ITS sequences for resolving phylogenetic relationships within the phylum Ctenophora. Zoology (Jena) *118*, 102–114.
27. Zapata, F., Goetz, F.E., Smith, S.A., Howison, M., Siebert, S., Church, S.H., Sanders, S.M., Ames, C.L., McFadden, C.S., France, S.C., et al. (2015). Phylogenomic analyses support traditional relationships within Cnidaria. PLoS ONE *10*, e0139068.
28. Halanych, K.M. (2015). The ctenophore lineage is older than sponges? That cannot be right! Or can it? J. Exp. Biol. *218*, 592–597.
29. Lartillot, N., Brinkmann, H., and Philippe, H. (2007). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol. Biol. *7* (*Suppl 1*), S4.
30. Collins, A.G. (1998). Evaluating multiple alternative hypotheses for the origin of Bilateria: an analysis of 18S rRNA molecular evidence. Proc. Natl. Acad. Sci. USA *95*, 15458–15463.
31. Srivastava, M., Begovic, E., Chapman, J., Putnam, N.H., Hellsten, U., Kawashima, T., Kuo, A., Mitros, T., Salamov, A., Carpenter, M.L., et al. (2008). The Trichoplax genome and the nature of placozoans. Nature *454*, 955–960.
32. Roure, B., and Philippe, H. (2011). Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. BMC Evol. Biol. *11*, 17.
33. Dunn, C.W., Leys, S.P., and Haddock, S.H. (2015). The hidden biology of sponges and ctenophores. Trends Ecol. Evol. *30*, 282–291.
34. Jager, M., and Manuel, M. (2016). Ctenophores: an evolutionary-developmental perspective. Curr. Opin. Genet. Dev. *39*, 85–92.
35. Boussau, B., Szöllosi, G.J., Duret, L., Gouy, M., Tannier, E., and Daubin, V. (2013). Genome-scale coestimation of species and gene trees. Genome Res. *23*, 323–330.
36. Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460–2461.
37. Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. *13*, 2178–2189.
38. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.
39. Amemiya, C.T., Alföldi, J., Lee, A.P., Fan, S., Philippe, H., Maccallum, I., Braasch, I., Manousaki, T., Schneider, I., Rohner, N., et al. (2013). The African coelacanth genome provides insights into tetrapod evolution. Nature *496*, 311–316.
40. Criscuolo, A., and 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.
41. Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics *22*, 2688–2690.
42. Chevreux, B. (2005). MIRA: An Automated Genome and EST Assembler (Ruprecht-Karls-University).
43. Huang, X., and Madan, A. (1999). CAP3: A DNA sequence assembly program. Genome Res. *9*, 868–877.
44. Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. *8*, 1494–1512.
45. Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., et al. (2014). SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics *30*, 1660–1666.
46. Roure, B., Rodriguez-Ezpeleta, N., and Philippe, H. (2007). SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. BMC Evol. Biol. *7* (*Suppl 1*), S2.
47. Roure, B., Baurain, D., and Philippe, H. (2013). Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol. Biol. Evol. *30*, 197–214.
48. Schmidt, H.A., Strimmer, K., Vingron, M., and von Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics *18*, 502–504.
49. Stamatakis, A., and Ott, M. (2008). Efficient computation of the phylogenetic likelihood function on multi-gene alignments and multi-core architectures. Philos. Trans. R. Soc. Lond. B Biol. Sci. *363*, 3977–3984.
50. Stone, M. (1974). Cross validatory choice and assessments of statistical predictions. J. R. Stat. Soc. B *36*, 111–117.
51. Lartillot, N., Lepage, T., and Blanquart, S. (2009). PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics *25*, 2286–2288.
52. Lartillot, N., Rodrigue, N., Stubbs, D., and Richer, J. (2013). PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst. Biol. *62*, 611–615.
53. Nguyen, L.T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268–274.
54. Rodríguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B.F., and Philippe, H. (2007). Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol. *56*, 389–399.
55. Hrdy, I., Hirt, R.P., Dolezal, P., Bardonová, L., Foster, P.G., Tachezy, J., and Embley, T.M. (2004). *Trichomonas* hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. Nature *432*, 618–622.

Current Biology, Volume *27*

Supplemental Information

A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals

Paul Simion, Hervé Philippe, Denis Baurain, Muriel Jager, Daniel J. Richter, Arnaud Di Franco, Béatrice Roure, Nori Satoh, Éric Quéinnec, Alexander Ereskovsky, Pascal Lapébie, Erwan Corre, Frédéric Delsuc, Nicole King, Gert Wörheide, and Michaël Manuel

Supplemental Figures

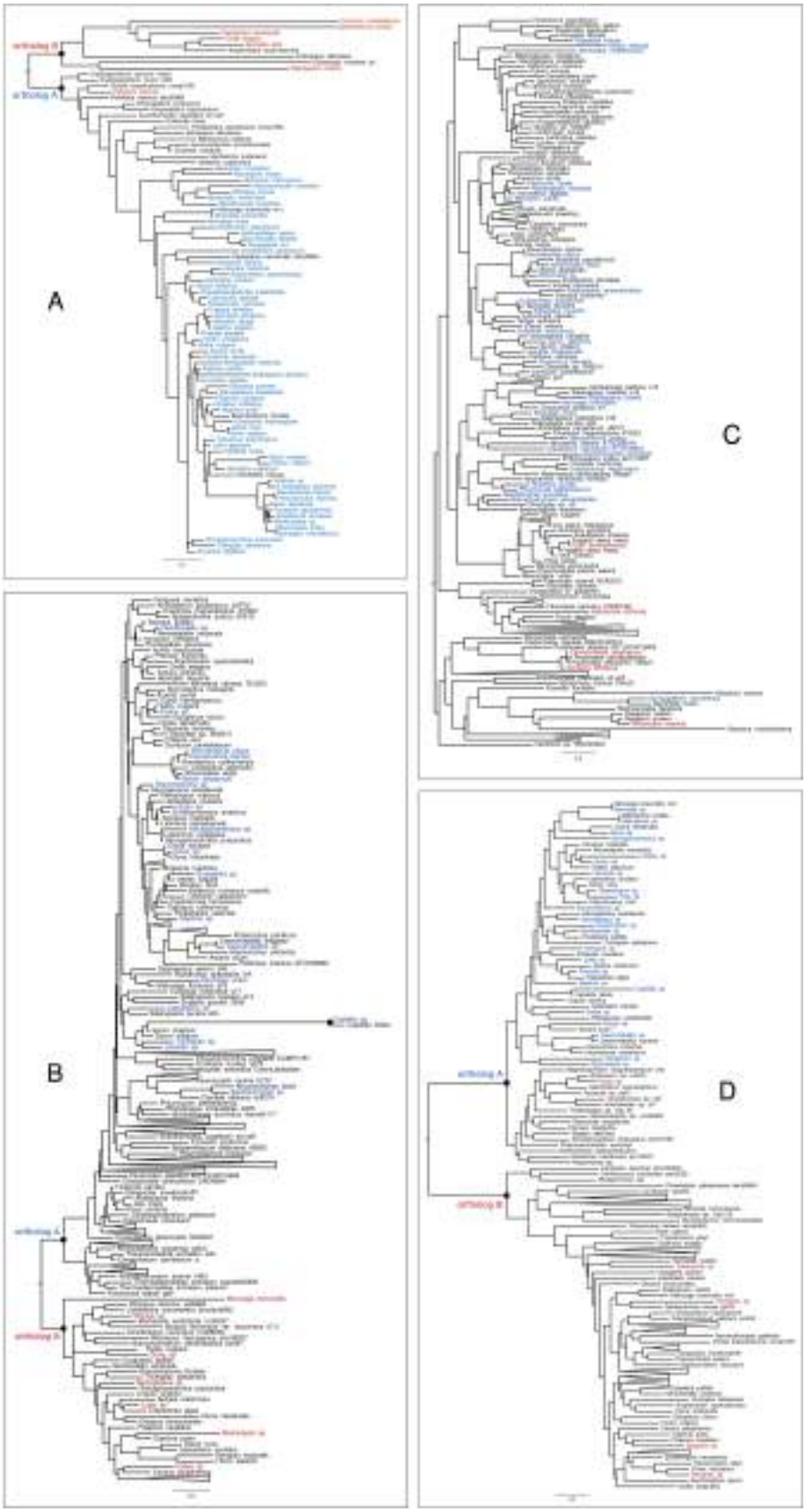


Figure S1. Examples of paralogy and contaminations in previous phylogenomic datasets. Related to Figure 2.

Sequences included in the original dataset are coloured. Correct sequences are coloured in blue whereas contaminations or deep paralogues are coloured in red. Sequences in black were added in the context of this study in order to detect problematic sequences.

(A) Gene 90 from [S1].

(B**)** Gene 953 in dataset 3A from [S2].

(C) Gene 120 from [S3].

(D) Gene 1160 from [S4]. The single-gene tree inference procedure is presented in Experimental Procedures.

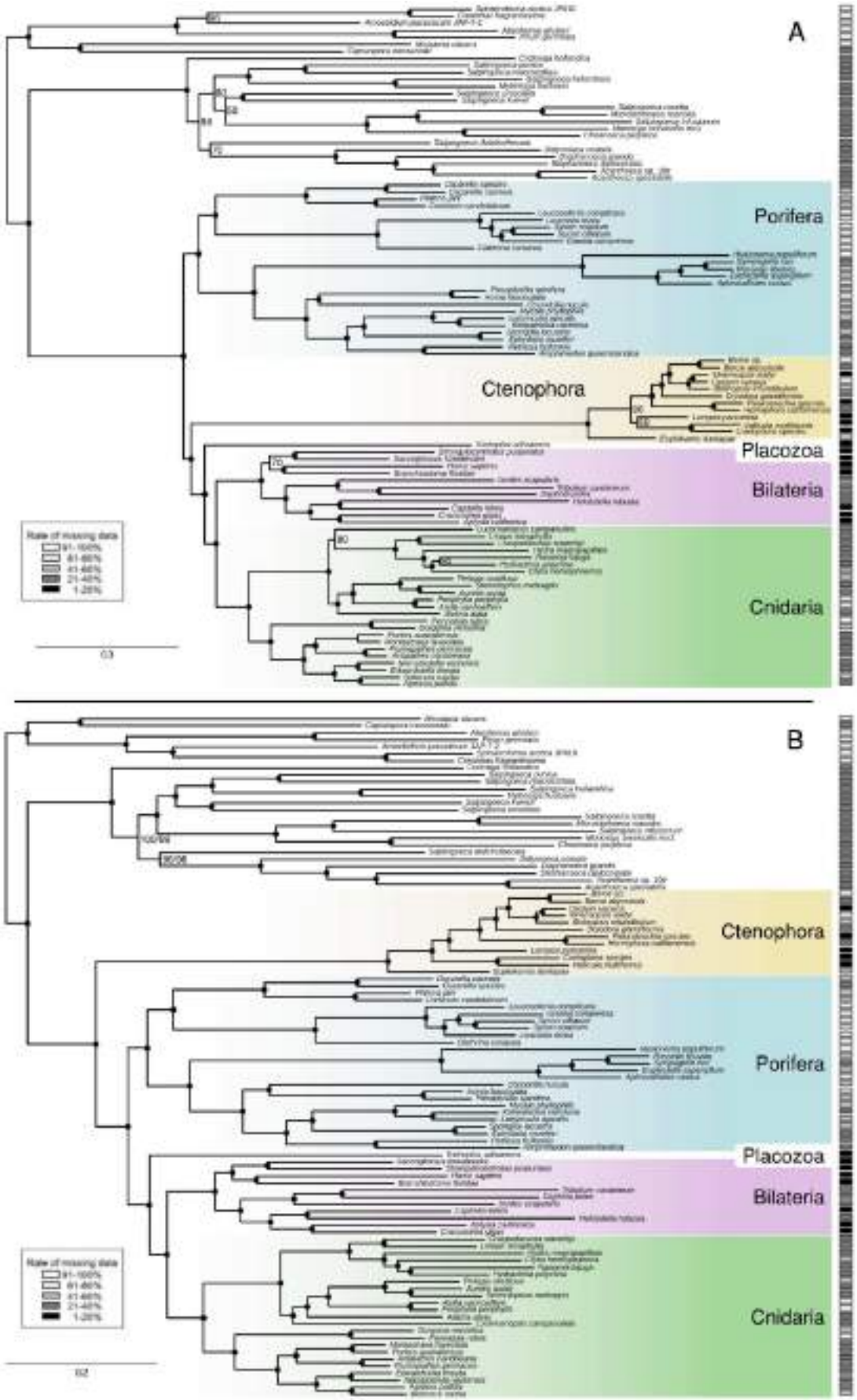

A
Porifera
Ctenophora
Placozoa
Bilateria
Cnidaria
Rate of missing data
91-100%
61-80%
41-60%
21-40%
1-20%
0.3
B
Ctenophora
Porifera
Placozoa
Bilateria
Cnidaria
Rate of missing data
91-100%
61-80%
41-60%
21-40%
1-20%
0.2

Figure S2. Metazoan phylogenetic relationships inferred when including holozoan outgroups more distant than choanoflagellates. Related to Figure 3.

Proportions of missing data indicated for each species in the trees.

(A) Results from 10 jackknife replicates on our complete supermatrix (*i.e.* 401,632 amino acids positions for 97 species with an overall percentage of missing data of 39,3%). Every jackknife replicate was composed of randomly chosen 100,000 positions and has been analysed using PhyloBayes 1.6j under the site-heterogeneous CAT+ Γ_4 model of sequence evolution. The tree shown is the consensus of the 10 jackknife replicates; node support represents the number of analyses in which the branch was found; black circles represent maximal support.

(B) Results from analyses under a site-homogeneous partitioned LG+ Γ_4+F model (one partition per gene) with RAxML on the 401,632 amino acids supermatrix, with the full taxonomic sampling of 97 species (first support value) and with a reduced taxonomic sampling after removal of the most distant outgroups (90 species; second support value). Black circles indicate maximal bootstrap support with both taxonomic samplings.

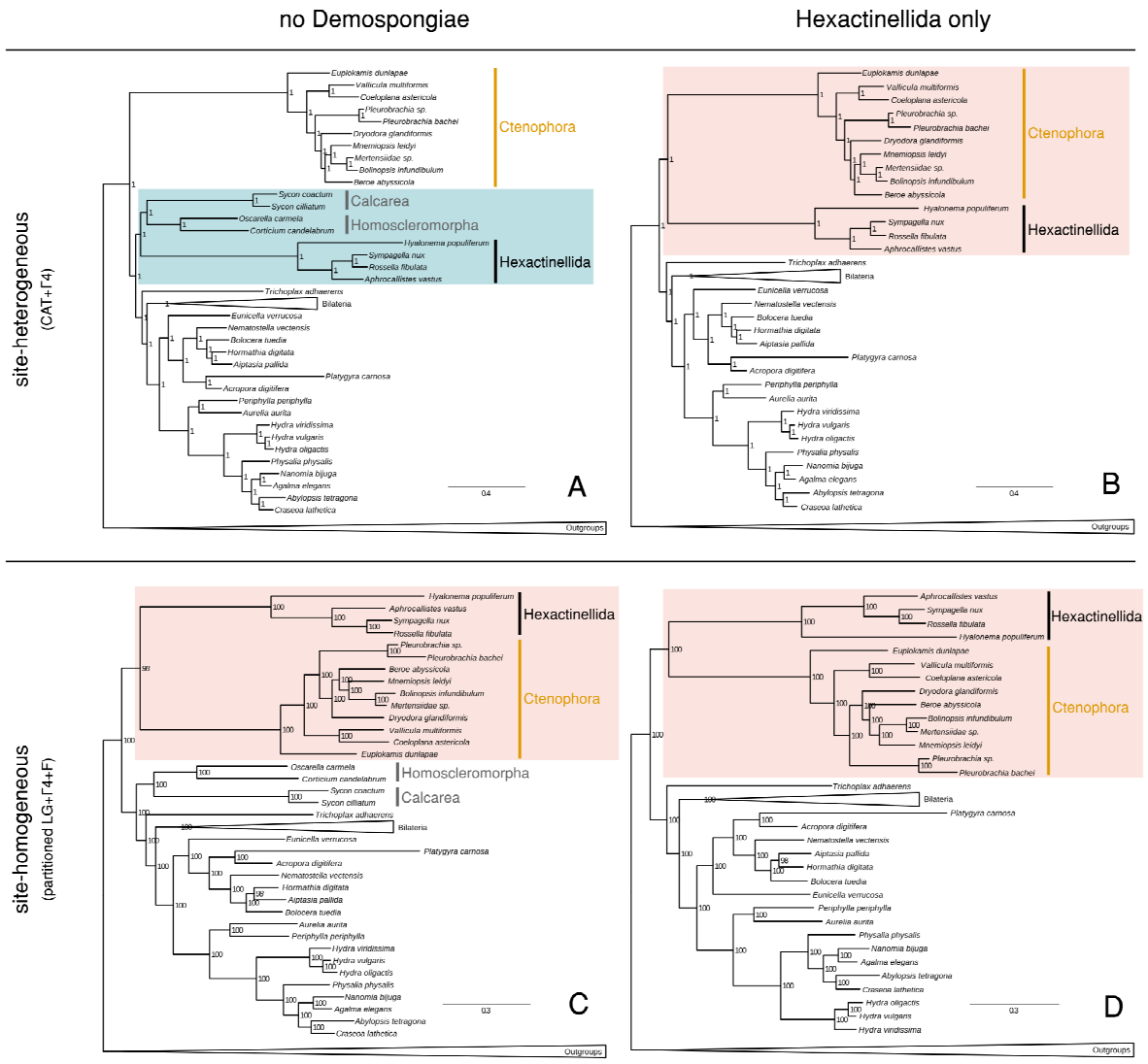


Figure S3. Comparison of the effect of using the site-heterogeneous CAT model vs. the site-homogeneous LG model on metazoan relationships when progressively removing hexactinellid sister-clades from the dataset of [S1]. Related to Figure 4.

(A) Removal of demosponges and Bayesian inference with PhyloBayes under the CAT+ Γ_4 evolution model.

(B) Removal of all sponges except hexactinellids and Bayesian inference with PhyloBayes under the CAT+ Γ_4 evolution model.

(C) Removal of demosponges and maximum likelihood inference with RAxML under the LG+ Γ_4+F evolution model.

(D) Removal of all sponges except hexactinellids and maximum likelihood inference with RAxML under the LG+ Γ_4+F evolution model. Node support for (A) and (B) are posterior probabilities, and node support for (C) and (D) are bootstrap values. Details of phylogenetic relationships within bilaterians and outgroups are not shown for clarity of presentation. Full trees are available at https://github.com/psimion/SuppData_Metazoa_2017.

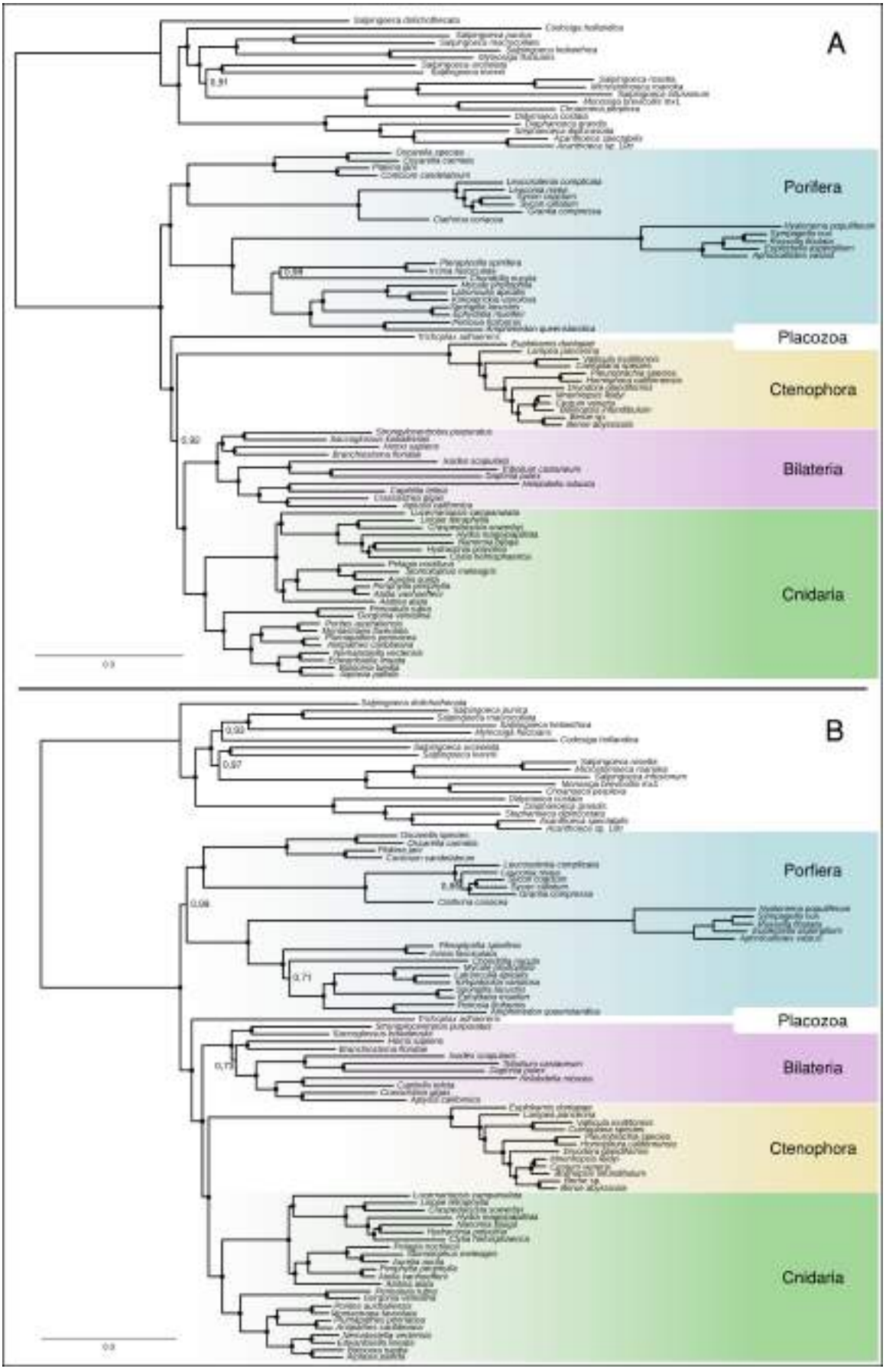
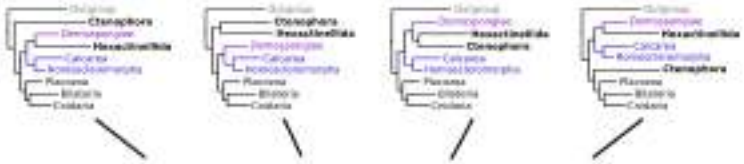

A
Porifera
Placozoa
Ctenophora
Bilateria
Cnidaria
B
Porifera
Placozoa
Ctenophora
Bilateria
Cnidaria

Figure S4. Metazoan phylogenetic relationships inferred under the CAT+ Γ_4 evolution model with PhyloBayes when progressively removing heteropecilleous sites. **Related to Figure 3.**
(A) Removal of 60% of the most heteropecilleous sites (*i.e.* 136,618 amino acids positions for 90 species remained).
(B) Removal of 70% of the most heteropecilleous sites (*i.e.* 102,464 amino acids positions for 90 species remained). The tree shown for each analysis is a consensus of two MCMC chains after a 50% burn-in and corresponding posterior probabilities are indicated at each node.

Supplemental Tables

Table S1. Impact of model and taxon sampling on the position of hexactinellids, evaluated using computation of likelihoods for various models. Related to Figure 4.

For each of the four topologies presented on top and for each sequence evolution model listed on the left, likelihood was computed and the corresponding ΔlnL are shown in the coloured cells. This was repeated for the three taxon samplings presented on the taxon removal column, which correspond to progressive removal of the close relatives of hexactinellids (Demo., Calc. and Homo. respectively stand for Demospongiae, Calcarea and Homoscleromorpha). When taxa are removed, the tested topology corresponds to the illustrated tree pruned from the removed taxa. For clarity, ΔlnL values are associated with colours (green for the most favoured topology to yellow, orange and lastly red). For each model, Bayesian Information Criterion (BIC) and corrected Akaike Information Criterion (AICc) values are also shown (left side).



	Model	ΔlnL	ΔlnL	ΔlnL	ΔlnL	Taxon removal
	CF4+Γ	0.0	3810.4	5086.8	2787.8	
BIC	40450416.1	876.3	0.0	2100.1	3142.5	Demospongiae
AICc	40448224.7	592.2	0.0	0.0	509.1	Demo.,Calc.,Homo.
	WAG+Γ	0.0	4818.6	6077.9	3431.6	
BIC	39256336.9	911.3	0.0	2357.6	3807.8	Demospongiae
AICc	39254189.2	538.5	0.0	0.0	588.3	Demo.,Calc.,Homo.
	LG+Γ	0.0	4814.5	5988.5	3280.5	
BIC	38851736.6	803.8	0.0	2237.8	3580.1	Demospongiae
AICc	38849588.8	431.9	0.0	0.0	482.2	Demo.,Calc.,Homo.
	UL3+Γ	0.0	4528.9	5451.1	2634.6	
BIC	38708552.2	543.6	0.0	1890.1	2830.9	Demospongiae
AICc	38706382.7	270.6	0.0	0.0	333.5	Demo.,Calc.,Homo.
	EX_EHO+Γ	0.0	4713.4	5774.1	2989.6	
BIC	38643674.4	578.6	0.0	2053.9	3147.8	Demospongiae
AICc	38641472.2	279.9	0.0	0.0	367.5	Demo.,Calc.,Homo.
	C20+Γ	0.0	2875.6	3146.6	964.1	
BIC	37795816.3	255.9	0.0	876.7	1049.0	Demospongiae
AICc	37793875.7	135.3	0.0	0.0	89.7	Demo.,Calc.,Homo.
	LG4X	0.0	4510.9	5420.7	2705.5	
BIC	37701045.6	642.5	0.0	1924.7	2968.6	Demospongiae
AICc	37699050.5	372.3	0.0	0.0	419.8	Demo.,Calc.,Homo.
	C40+Γ	0.0	2742.6	2831.1	722.6	
BIC	37572612.9	151.5	0.0	669.7	753.8	Demospongiae
AICc	37570672.2	109.5	0.0	0.0	78.2	Demo.,Calc.,Homo.
	C60+Γ	0.0	2653.9	2679.4	607.6	
BIC	37474544.5	139.4	0.0	593.4	634.7	Demospongiae
AICc	37472603.9	98.8	0.0	0.0	67.8	Demo.,Calc.,Homo.

Table S2. Summary of the 21 non-bilaterian species newly sequenced for this study. Related to Experimental Procedures.

Species, sampling location, sequencing method, species pooling for sequencing, number of reads, mean read length and SRA accession number.

	Taxonomy	Species	Location	sequencing technology	read number	mean read size (bp)	Accession number
PORIFERA	Calcarea	*Leucosolenia variabilis*	laboratory tanks	454	132,287	226	SRR3405426
		Leuconia nivea	Roscoff, France	Illumina	27,051,000	2 * 100	SRR3417190
		Clathrina coriacea	Roscoff, France	Illumina	24,118,600	2 * 100	SRR3417192
		Granta compressa	Roscoff, France	Illumina	32,598,500	2 * 100	SRR3417193
	Homoscleromorpha	*Plakina jani*	Marseille, France	Illumina	28,462,400	2 * 100	SRR3417194
	Demospongiae	*Plevaplysilla spatifera*	Marseille, France	Illumina	25,884,600	2 * 100	SRR3417588
CTENOPHORA	Cestida	*Cestum veneris*	Villefranche-sur-Mer, France	454	186,537	282	SRR3341364
	Platyctenida	*Coeloplana* cf. *meteoris*	Inland Aquatics, USA	Illumina	73,657,000	2 * 100	SRR3407215
		Vallicula multiformis	Inland Aquatics, USA	Illumina	34,999,800	2 * 100	SRR3407164
	« Cydippida »	*Lampea pancerina*	Villefranche-sur-Mer, France	Illumina	28,732,300	2 * 100	SRR3407163
		Pleurobrachia pileus	Roscoff, France	454	1,260,758	326	SRR3417589
CNIDARIA	Octocorallia	*Pennatula rubra*	Banyuls-sur-Mer, France	454	188,407	256	SRR3405415
		Alcyonium palmatum	Banyuls-sur-Mer, France	Illumina	28,441,600	2 * 100	SRR3407218
	Hexacorallia	*Parazoanthus axinellae*	Villefranche-sur-Mer, France	454	155,636	282	SRR3405413
		Pachycerianthus solitarius	Etang de Berre, France	454	206,817	244	SRR3405399
		Antipathes caribbeana	Pori-Louis, Guadeloupe	Illumina	49,997,500	2 * 100	SRR3407160
		Plumapathes pennacea	Pori-Louis, Guadeloupe	Illumina	59,779,500	2 * 100	SRR3407161
	Medusozoa	*Carybdea xaymacana*	Pointe-à-Pitre, Guadeloupe	454	161,900	244	SRR3405427
		Lucernariopsis campanulata	Île Callot, Morlaix, France	Illumina	28,539,600	2 * 100	SRR3407219
		Pelagia noctiluca	Villefranche-sur-Mer, France	Illumina	37,484,100	2 * 100	SRR3407257
		Linuche unguiculata	Sainte-Rose, Guadeloupe	Illumina	30,934,800	2 * 100	SRR3407335

Table S3. Information about sequence data for the taxon sampling of non-bilaterian species used in this study. Related to Experimental Procedures.

Simplified classification, species name, data source, sequencing groups, initial number of transcripts, number of transcripts remaining after the successive steps of decontamination (DCC, DC1, DC2 and DC3), final number of transcripts used to enrich core orthologous clusters, data status (*i.e.* discarded from the dataset, merged or kept for further analyses). For justification of data removals, see Experimental Procedures.

CLASSIFICATION	SPECIES	SOURCE	GROUP	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	FINAL [illegible]	FINAL STATUS
[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	A	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	A	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	A	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	fixed
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]	C	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]	C	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	C	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	C	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	C	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	C	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]	C	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]	D	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	C	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	C	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]	D	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	D	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	D	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]	C	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	C	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	fixed
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
	[illegible]	[illegible]	D	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	fixed
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
	[illegible]	[illegible]	D	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	D	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	fixed
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
	[illegible]	[illegible]	D	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	fixed
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]	D	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	D	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	fixed
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	C	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	D	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	B	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	B	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]	B	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	D	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	fixed
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	discarded
	[illegible]	[illegible]	D	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok
	[illegible]	[illegible]	D	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	ok

Table S4. List of the 20 "core species", corresponding sources, and sizes of proteomes used to build similarity clusters with OrthoMCL. Related to Experimental Procedures.

The number of proteins indicated for each species corresponds to non-redundant proteins after removal of sequences smaller than 100 amino-acids or with more than 10% ambiguous or missing characters.

phylum	species	source	proteins
Filasterea	*Capsaspora owczarzaki*	NCBI	16878
Choanoflagellata	*Monosiga brevicollis*	NCBI	18141
Choanoflagellata	*Salpingoeca rosetta*	NCBI	22780
Porifera	*Amphimedon queenslandica*	Ensembl	25858
Placozoa	*Trichoplax adhaerens*	Ensembl	11217
Ctenophora	*Mnemiopsis leidyi*	NHGR	16030
Cnidaria	*Nematostella vectensis*	Ensembl	22084
Cnidaria	*Acropora digitifera*	Compagen	21433
Cnidaria	*Hydra magnipapillata*	Compagen	29647
Echinodermata	*Strongylocentrotus purpuratus*	Ensembl	27147
Hemichordata	*Saccoglossus kowalevskii*	JGI	28734
Cephalochordata	*Branchiostoma floridae*	UniProt	27293
Chordata	*Homo sapiens*	UniProt	24884
Mollusca	*Crassostrea gigas*	Ensembl	23542
Mollusca	*Aplysia californica*	NCBI	24436
Annelida	*Capitella teleta*	Ensembl	30014
Annelida	*Helobdella robusta*	Ensembl	21952
Arthropoda	*Daphnia pulex*	Ensembl	26539
Arthropoda	*Ixodes scapularis*	Ensembl	16659
Arthropoda	*Tribolium castaneum*	Ensembl	14942

Supplemental Experimental Procedures

Decontamination of transcriptomes

De-Cross-Contamination (DCC) of the transcriptomes

Multiple deep sequencing of different samples prepared by the same lab and/or sequenced at the same facility easily leads to cross-contamination between samples [S5], which can be extremely deleterious to phylogenetic inference if not eliminated [S6]. Cross-contaminating sequences can readily be detected by building and inspecting single-gene trees inferred from alignments with a rich taxonomic diversity. We performed this check for a number of ribosomal protein alignments after addition of sponge, cnidarian and ctenophore transcripts sequenced by us. It was not rare to see a transcript supposed to belong to a given species (e.g., a cnidarian) positioned in the tree at an obvious "wrong" place (e.g., closer to a poriferan transcript than to another cnidarian) or with 100% nucleotide identity between two transcripts from unrelated species (in this example, the cnidarian transcriptome was cross-contaminated by poriferan transcripts). Such cross-contaminations can occur despite the care taken during preparation of the RNA samples at the bench. Although their causes are not entirely clear, contamination might have been introduced at different steps: RNA preparation, library construction, addition of oligonucleotide tags, and demultiplexing of samples after sequencing (here, all steps following RNA preparation were carried out at the company carrying out the sequencing, GATC). In particular, sequencing using multiplexing strategies with sample-specific tags are known to suffer from a proportion of incorrect sample identification: as high as 0.3% of the reads can be assigned to an incorrect sample [S5].

We designed a "De-Cross-Contamination" (DCC) protocol that detects contaminant transcripts within a series of several species sequenced by the same lab and removes them from the transcriptomes. This protocol is based on the idea that a cross-contaminant should be quantitatively under-represented among the sequencing reads, with respect to the genuine transcript of the sequenced species. Indeed, Illumina RNA-seq is considered reliable to provide quantitative estimates of transcript abundance and, as such, is widely used to evaluate differential gene expression, the expression level of a gene being reflected by the relative numbers of reads mapping to this gene in the various sequenced samples. Hence, a contaminant RNA should behave like a lowly expressed RNA, and the genuine RNA should have many more reads (under the assumption that the expression level of orthologues is similar in the different species). Our DCC protocol is composed of three steps:

1) All transcripts from several species produced by the same lab or sequencing facility were compared to each other (all-vs-all nucleotide BLAST).
2) The number of reads mapping to each transcript was extracted from assembly statistics for each transcriptome.
3) When the same transcript (sequence identity at the nucleotide level > 98%) was ascribed to several species, we compared the number of reads. If a sequence of one species had 5 times more reads than all others, we considered that species as the genuine source of this transcript and removed the transcript from the transcriptomes of the other species. Otherwise, we eliminated the transcript from all transcriptomes, since its origin could not be determined with confidence. For a few transcriptome datasets (i.e. *Aphrocallistes vastus*, *Chondrilla nucula*, *Ephydatia muelleri*, *Ircinia fasciculata*, *Oscarella* sp. SN2011, *Spongilla lacustris*, *Pleurobrachia pileus* A, *Madracis auretenra*), read counts were not known, in which case any nucleotide identity > 98% across species resulted in the removal of the corresponding transcript for all concerned species. When a transcript had no hit, even to itself, it was also removed from its transcriptome (this only happens for very low-quality transcripts mostly containing low-complexity regions).

This method posits that samples used for RNA extraction are biologically comparable (i.e., mixtures of adult tissues, as those used for all species sequenced in this study). Our protocol was validated by comparing the disappearance of contaminants after applying our DCC protocol, as determined in single-gene trees for the 80 ribosomal protein markers manually curated and maintained in Hervé Philippe's lab (RIBO-80 dataset, see below list of numbers of sequences at each step of the cleaning procedure for this marker). Cross-contaminating sequences are in most cases easily identified in these single-gene trees thanks to the extensive taxonomic representation. Visual inspection of these gene trees (see example for rpl2 provided at https://github.com/psimion/SuppData_Metazoa_2017) showed that the DCC procedure indeed removed cross-contaminants, and that the transcripts retained by our program was correctly positioned with respect to the taxonomy of the species. The major drawback of our protocol is the possible loss of correct sequences that were not sufficiently "overexpressed" in the right sample, or that were naturally identical in nucleotides to another (very closely-related) species sequenced and equally expressed (in our sampling of sequenced species, this situation might have potentially occurred only for the two *Antipatharia* species).

Based on data source, 6 different groups of transcriptomes were defined and each one was processed through the DCC protocol (see "group" column in Table S3). Note that the DCC for the "DX1" group used a nucleotide identity threshold of 95% instead of 98%. Indeed, this DCC group contained only distantly related species, which enabled us to reduce the identity threshold with a lower risk of removing transcripts that would not be true cross-contaminants. On average, the DCC process resulted in the removal of 8.9% transcripts per species, but the amount of removed

transcripts was heterogeneous, ranging from 0.16% (*Pleurobrachia pileus*, this study) to 70.64% (ctenophore sp3 A, [S2]).
This automated approach to detect cross contamination is currently the object of a dedicated manuscript under preparation.

Removal of contaminants from other (natural) sources

When sequencing animal tissues at deep coverage, contaminating life forms, such as epibiotic organisms, symbionts, parasites (all these either small multicellular or unicellular organisms) are almost inevitable, regardless of the efforts deployed in "cleaning" the samples. However, this type of "natural" contamination bears a strong contradictory non-phylogenetic signal that, like cross-contamination, hampers a correct reconstruction of the evolutionary history of the groups of interest [S7]. Hence, contaminations across metazoan phyla are the most deleterious because we specifically aim to resolve relationships between these phyla. To solve this problem, we designed a "De-Contamination" (DC) protocol, which is composed of three successive steps:

Suppression of non-holozoan contaminations (DC1) – A dataset of proteomes representing the molecular diversity of Eukaryota, Eubacteria and Archaea was compiled (see list below). We extracted the open reading frames (ORFs) of several transcriptomes, for which we had no established proteomes available, using the findorfs option of USEARCH (xlat, orfstyle = 7, mincodons = 60; for very large transcriptomes, the mincodon parameter was increased up to 140 to reduce final database size). To represent the molecular diversity of holozoans, we added the 20 core proteomes used to determine the orthologous groups. The ORFs of every transcript from every holozoan transcriptome (DCC-processed, if required) were "BLASTed" against this database using diamond (blastx, sensitive, evalue = $1e^{-10}$). When the best bit score corresponded to a transcript that did not belong to the 20 core proteomes, we removed the transcript from the corresponding transcriptome, resulting in the removal of 291,933 transcripts (see example for rpl2 provided at https://github.com/psimion/SuppData_Metazoa_2017).

Suppression of bilaterian contaminations (DC2) – A dataset of various proteomes and transcriptomes from taxa representing the molecular diversity of Bilateria down to the phylum level was compiled (see list below). We extracted the ORFs of bilaterian transcriptomes for which we had no proteomes available using the findorfs option of USEARCH (xlat, orfstyle = 7, mincodons = 100). We also extracted the ORFs of some of the cleanest non-bilaterian transcriptomes to be decontaminated, which we incorporated as references in the database to complete the molecular diversity representation of holozoans. The ORFs of every transcript from non-bilaterian

transcriptomes (including outgroups) were “BLASTed” against this database using diamond, as above (ignoring hits against themselves). When the best bit score corresponded to a bilaterian transcript or when the best non-bilaterian bit score was strictly equal to the best bilaterian bit score, we considered the transcript as a bilaterian contamination and subsequently removed from the corresponding transcriptome, resulting in the removal of 272,913 additional transcripts (see example for rpl2 provided at https://github.com/psimion/SuppData_Metazoa_2017).

Suppression of contaminations between Porifera and Cnidaria (DC3) – Preliminary observations of single-gene trees (mostly for ribosomal proteins) indicated a non-negligible level of contaminating sequences of cnidarian origin in sponge transcriptomes, and reciprocally. These marine animals (particularly sponges and benthic cnidarians) are very difficult to clean from natural contaminants (e.g., small hydrozoans present on the surface and/or in the cavities of a sponge; sponge larvae or young adults on a cnidarian sample). It was therefore necessary to specifically remove sponge contaminations from cnidarian transcriptomes and reciprocally. We extracted the ORFs of poriferan and cnidarian transcriptomes to be decontaminated as in DC2. We used the resulting ORFs as well as the core proteomes from *Amphimedon* and cnidarians to create a database of protein sequences. The ORFs of each transcript were then “BLASTed” against this database using diamond (as in DC2), again ignoring self hits. When the best bit score for a transcript of a given species did not belong to the expected phylum (Porifera or Cnidaria), we removed it from its transcriptome, resulting in the removal of 43,234 additional transcripts (see example for rpl2 provided at https://github.com/psimion/SuppData_Metazoa_2017). Note that contaminations between non-bilaterian phyla involving ctenophores or placozoans are much less likely, and were never observed during our inspection of single-gene trees.

All these decontamination steps were evaluated against the RIBO-80 datasets (see below list of numbers of sequences at each step of the cleaning procedure for this marker). The number of correct sequences that were removed is generally small (but see e.g. *Abeoforma*), indicating that the false positives are rare. Yet, the number of remaining contaminants can be high (e.g. 340 for *Ectopleura* or 2,304 for *Sycon ciliatum*), indicating an important level of false negatives. Several attempts to reduce the number of false negatives drastically increased the number of false positives. It should be noticed that the highly expressed genes (here ribosomal proteins) are extreme cases for contaminations, so the impact of these remaining erroneous sequences should be negligible, all the more since additional cleaning procedures have been used downstream in the protocol (see Experimental Procedures). On the decontaminated tree of rpl2 (https://github.com/psimion/SuppData_Metazoa_2017), one can see that the sequences finally used

in the super-matrix are not contaminants despite the fact that the decontaminated transcriptomes still contained a large number of erroneous sequences.

List of organisms used as references during the first step of transcriptome decontamination (DC1), with species names and data sources.

Non-Holozoa	
species	**sources**
Thalassiosira pseudonana	PRJNA34119
Naegleria gruberi strain NEG-M	PRJNA43883
Chlamydomonas reinhardtii	PRJNA21081
Arabidopsis lyrata lyrata	PRJNA48645
Chondrus crispus	PRJNA193762
Acanthamoeba castellanii	PRJNA193615
Dictyostelium purpureum	PRJNA63631
Bigelowiella natans	PRJNA27559
Toxoplasma gondii ME49	PRJNA32719
Oxytricha trifallax	ARYC00000000.1
Tetrahymena thermophila SB210	PRJNA16792
Perkinsus marinus ATCC 50983	PRJNA46451
Symbiodinium	nucleotide sequences from nr
Physcomitrella patens	PRJNA49677
Emiliania huxleyi	PRJNA222932
Trypanosoma cruzi	PRJNA15540
Batrachochytrium dendrobatidis JAM81	PRJNA225932
Rhizopus delemar	PRJNA13068
Phanerochaete carnosa	PRJNA38425
Aspergillus oryzae RIB40	PRJNA28175
Debaryomyces hansenii CBS767	PRJNA12410
Nostoc punctiforme PCC 73102	PRJNA57767
Escherichia coli	PRJNA176127
Rhizobium etli by mimosae	PRJNA213896
Sinkania negevensis	PRJNA88451
Bacillus firmus DS1	PRJNA193594
Streptomyces avermitacus JA 4570	PRJNA209659
Prevotella bergensis DSM 17361	PRJNA55805
Marinobacter adhaerens HP1S	PRJNA162099
Marinomonas mediterranea	PRJNA64763
Idiomarina loihiensis L2TR	PRJNA205256
Pyrococcus abyssi GE5	PRJNA62902
Sulfolobus solfataricus P2	PRJNA57721
Candidatus Nitrosopumilus	PRJNA176130
Chlorella uncinata	PRJNA161697
Cyanophyche gloeocystis SAG467	MMETSP
Protocruzia adherens Boccale	MMETSP
Strombidinopsis acuminatum SPMC142	MMETSP
Alexandrium andersoni CCMP2222	MMETSP
Symbiodinium sp cladeA	MMETSP
Diplonema sp.	Gertraud Burger
Pyrodinium bahamense pbaha01	MMETSP
Entamoeba dispar	PRJNA28615, PRJNA12914
Guillardia theta	PRJNA223935
Phaeodactylum tricornutum	PRJNA13251
Reticulomyxa filosa	PRJNA28155
Tetraselmis sp.	MMETSP
Trypanosoma grayi	PRJNA258390, PRJNA244695
Ammonia sp.	MMETSP
Euglena gracilis 3039	Denis Baurain
Filamoeba nolandi NCA5231	MMETSP
Gloeochaete wittrockiana SAG46B4	MMETSP
Litonotus pesus P1	MMETSP
Mikrocytos mackini	PRJNA197256
Paramikrylella glossopodia RCC365	MMETSP
Pessonella sp.	MMETSP
Roscilna sp.	MMETSP
Vannella sp. DNA1617612	MMETSP
Miamiensis avidus	DR981915-DR980992
Amphyroides haemophila	GT633512-GT640124
Paramoeba atlantica	SRR1296884

Holozoa	
species	**sources**
Oscarella sp. SN2011	SRR1044413
Pleurobrachia bachei	GeneModels
Aphrocallistes vastus	ERA archives
Clione varians	PRJNA214560
Latrunculia apicalis	SRR1915758
Acanthoeca sp. 10b	SRR1296844, SRR1294413
Savillea parva%14	King N. & Richter D.
Didymoeca costata%10	King N. & Richter D.
Salpingoeca urceolata%04	King N. & Richter D.
Salpingoeca infusionum%12	King N. & Richter D.
Mylnosiga fluctuans%19	King N. & Richter D.
Capsaspora owczarzaki	PRJNA193611, PRJNA20041
Monosiga brevicollis	PRJNA28133, PRJNA19045
Salpingoeca rosetta	PRJNA193541, PRJNA37927
Amphimedon queenslandica	ENSEMBL
Trichoplax adhaerens	ENSEMBL
Mnemiopsis leidyi	NHGR
Nematostella vectensis	ENSEMBL
Acropora digitifera	COMPAGEN
Hydra magnipapillata	COMPAGEN
Daphnia pulex	ENSEMBL
Ixodes scapularis	ENSEMBL
Tribolium castaneum	ENSEMBL
Crassostrea gigas	ENSEMBL
Aplysia californica	PRJNA209509
Capitella teleta	ENSEMBL
Helobdella robusta	ENSEMBL
Strongylocentrotus purpur.	ENSEMBL
Saccoglossus kowalevskii	JGI
Branchiostoma floridae	UNIPROT
Homo sapiens	UNIPROT

List of organisms used as references during the second step of transcriptome decontamination (DC2). Species names and data sources.

Non-Bilateria	
species	**sources**
Capsaspora owczarzaki	PRJNA193613, PRJNA20341
Monosiga brevicollis	PRJNA28133, PRJNA19045
Salpingoeca rosetta	PRJNA193541, PRJNA37927
Amphimedon queenslandica	ENSEMBL
Trichoplax adhaerens	ENSEMBL
Mnemiopsis leidyi	NHGR
Nematostella vectensis	ENSEMBL
Acropora digitifera	COMPAGEN
Hydra magnipapillata	COMPAGEN
Acanthoeca spectabilis VA%02	Previous step (DC1)
Acropora cervicornis	Previous step (DC1)
Aphrocallistes vastus	Previous step (DC1)
Beroe abyssicola	Previous step (DC1)
Bolocera tuediа	Previous step (DC1)
Choanoeca perplexa	Previous step (DC1)
Ciona varians	Previous step (DC1)
Clytia hemisphaerica	Previous step (DC1)
Codosiga hollandica%17	Previous step (DC1)
Cyanea capillata	Previous step (DC1)
Diaphanoeca grandis RI%01	Previous step (DC1)
Didymoeca costata%10	Previous step (DC1)
Helgoeca nana%08	Previous step (DC1)
Oscarella sp	Previous step (DC1)
Oscarella sp SN2011	Previous step (DC1)
Pachycerianthus solitarius	Previous step (DC1)
Parazoanthus axinellae	Previous step (DC1)
Pelagia noctiluca	Previous step (DC1)
Pennatula rubra	Previous step (DC1)
Pleurobrachia pileus	Previous step (DC1)
Pseudospongosorites suberitoides	Previous step (DC1)
Salpingoeca qvevri%09	Previous step (DC1)
Salpingoeca roanoka%13	Previous step (DC1)
Stephanoeca diplocostata FR	Previous step (DC1)
Sycon raphanus	Previous step (DC1)
Tripedalia cystophora	Previous step (DC1)
Vallicula multiformis	Previous step (DC1)
Pleurobrachia bachei	Previous step (DC1)
Latrunculia apicalis	Previous step (DC1)
Acanthoeca sp 10tr	Previous step (DC1)
Savillea parva%14	Previous step (DC1)
Salpingoeca urceolata%04	Previous step (DC1)
Salpingoeca infusionum%12	Previous step (DC1)
Mylnosiga fluctuans%19	Previous step (DC1)

Bilateria	
species	**sources**
Daphnia pulex	ENSEMBL
Ixodes scapularis	ENSEMBL
Tribolium castaneum	ENSEMBL
Crassostrea gigas	ENSEMBL
Aplysia californica	PRJNA209509
Capitella teleta	ENSEMBL
Helobdella robusta	ENSEMBL
Strongylocentrotus purpuratus	ENSEMBL
Saccoglossus kowalevskii	JGI
Branchiostoma floridae	UNIPROT
Homo sapiens	UNIPROT
Hymenolepis microstoma	NCBI
Clonorchis sinensis	NCBI
Ascaris suum	NCBI
Necator americanus	NCBI
Trichinella spiralis	NCBI
Oikopleura dioica	NCBI
Ciona intestinalis	NCBI
Spadella cephaloptera	NCBI
Bugula neritina	SRA
Sipunculus nudus	CLC (curtesy of Anna Riesgo)
Lepeophtheirus Salmonis	NCBI
Botryllus schlosseri	NCBI
Brachionus plicatilis	NCBI
Tubulipora sp.	NCBI
Platynereis dumerilii	NCBI
Patria pectinifera	NCBI

List of numbers of sequences for the RIBO-80 dataset, at the various steps of the cleaning procedures (example of validation of our procedure).

For each species, the number of correct sequences as well as contaminants/paralogs incorporated in the alignements with the Forty-Two software is shown at various cleaning steps (*i.e.* initial transcriptomes, after the filter focus step, after the de-cross-contamination step, after decontamination 1, after decontamination 2 and after decontamination 3). “OK” indicates the total number of correct sequences added, “Contam” indicates the total number of contaminated sequences added, and “contigs” is the number of remaining transcript per transcriptomes.

	INITIAL			FIFO			DCC			DC1			DC2			DC3		
Species	OK	Contam	contigs	OK	Contam	contigs	OK	Contam	contigs	OK	Contam	contigs	OK	Contam	contigs	OK	Contam	contigs
Abeoforma whisleri	88	118	57555	85	111	7771	82	6	7105	69	5	5021	43	0	3231	43	0	3231
Acanthoeca sp._10tr	88	2	91237	88	2	33225	88	2	33225	88	2	33082	88	2	31826	88	2	31826
Acanthoeca spectabilis_VA%02	79	3	44902	79	3	10943	79	3	10943	79	3	10254	79	3	10015	79	3	10015
Actinosa cervicornis	84	226	95681	84	205	20409	84	205	20409	83	7	14725	77	4	13386	77	2	13344
Aegina citrea	113	45	138972	113	45	14281	113	45	14281	111	34	12831	102	3	10327	93	3	10072
Aigtasia pallida	109	599	147665	104	527	25438	104	527	25438	103	72	15978	97	33	14029	95	23	13509
Alatina alata	88	24	401703	87	25	29180	87	25	29180	88	17	25918	82	4	19632	90	0	18553
Alcyonium palmatum	85	409	102069	80	326	12963	76	141	12168	73	34	11465	66	12	9110	66	10	9009
Amoebidium parasiticum_JAP72	111	137	105237	109	130	18147	109	130	18147	87	33	13040	65	18	8238	65	18	8238
Anemonia viridis	215	21	10575	214	17	3085	214	17	3085	214	15	2852	212	15	2753	179	12	2533
Antipathes caribbeana	85	153	126470	81	113	14122	79	63	13948	78	43	13459	74	9	12061	74	7	12003
Aphrocallistes vastus	80	2	46987	80	2	21615	80	2	21608	79	2	21574	54	2	9290	54	2	8660
Asbestopluma hypogea	160	17	17930	155	12	4208	155	12	4208	154	6	3616	153	4	3272	143	3	2919
Atella vanhoeffeni	72	103	34090	72	100	8396	72	100	8396	72	92	7773	69	13	6152	69	13	6089
Aurelia aurita_B	90	32	196771	89	32	29395	89	32	29395	87	29	26161	84	26	22306	84	26	21945
Aurelia aurita	1049	301	112401	982	247	19855	982	247	19855	890	177	18554	818	85	16549	804	79	16419
Berae abyssicola	92	20	44289	92	19	8691	90	8	8636	88	8	8356	84	7	7929	84	7	7829
Berae sp.	254	52	22352	248	50	4525	248	50	4525	242	27	4214	233	11	4000	233	11	4000
Bolinopsis infundibulum	93	343	199328	91	339	24188	69	19	13757	69	18	12715	63	13	10220	63	13	10220
Bolocera tuedia	72	6	50694	72	6	12379	72	6	12379	71	6	11914	71	6	10960	71	6	10922
Bugula neritina	108	110	103166	106	100	16789	106	100	16789	100	45	13457	0	3	131	0	3	131
Carteriospongia foliascens	87	5	3556	84	4	1130	84	4	1130	83	4	1043	53	1	650	53	0	601
Carybdea xaymacana	75	10	6281	75	10	2749	75	10	2749	73	9	2482	66	6	2114	65	6	2075
Cestum veneris	138	28	15003	138	26	7711	137	26	7344	137	25	7120	136	20	6863	136	20	6863
Choanoeca perplexa%11	80	0	25369	80	0	9411	80	0	9411	79	0	9808	77	0	8673	77	0	8673
Chondrilla nucula	80	225	41571	80	223	12351	77	168	12104	77	162	11159	64	59	7942	60	22	6720
Clathrina coriacea	86	213	118488	86	187	14461	86	178	14426	83	98	12927	69	3	6289	67	2	4746
Cliona varians	92	4	39481	92	4	10358	83	4	10205	80	3	9753	79	4	8591	77	4	8314
Clytia hemisphaerica	93	49	36542	92	47	11269	92	47	11269	92	7	10570	86	3	8673	33	0	6443
Codosiga hollandica%17	89	6	70401	89	6	17260	89	6	17260	80	1	14659	69	1	10537	69	1	10537
Coeloplana astericola	92	478	142939	91	467	12995	89	34	9879	89	30	9236	82	14	8294	82	14	8294
Coeloplana cf._meteoris	98	457	110525	94	352	11391	86	119	10322	85	43	9846	80	24	9510	80	24	9510
Corallium rubrum	570	1055	504635	344	139	38010	344	139	38010	205	58	32004	143	17	21820	136	16	21490
Corticium candelabrum	93	381	560461	87	276	20062	87	90	18464	81	56	13921	66	18	9141	61	12	7852
Craspedacusta sowerbyi	149	272	350418	147	262	16752	138	113	15849	128	19	14263	110	13	11838	103	12	11696
Crateromorpha meyeri	15	6	1498	15	6	770	15	6	770	15	6	764	15	3	736	15	3	730
Crella elegans	97	472	288363	91	431	18352	23	343	17402	21	211	15244	19	33	12964	19	15	12393
Creolimax fragrantissima	92	50	38323	92	50	5416	92	36	5379	60	30	3965	51	1	2578	51	1	2573
Ctenophora sp3_A	385	10	171855	384	10	22316	31	2	6293	28	2	5303	13	0	3500	13	0	3500
Ctenophora sp3	163	3	1434	163	2	548	163	2	548	156	1	518	154	1	498	154	1	498
Cyanea capillata	134	6	2472	134	6	853	134	6	853	133	5	815	121	5	644	121	5	635
Diaphanoeca grandis_R%01	79	2	64955	79	2	9982	79	2	9982	78	1	9155	78	1	8841	78	1	8841
Didymoeca costata%10	85	1	36575	83	1	10414	83	1	10414	82	1	10322	79	1	9493	79	1	9493
Dryodora glandiformis	91	104	55048	91	104	10367	88	11	8748	86	9	8140	79	8	7390	79	8	7390
Ectopleura larynx	101	3796	329663	100	3360	34428	75	3158	33098	72	1034	21071	66	340	10658	60	208	10063
Edwardsiella lineata	91	25	90440	90	22	10774	90	22	10774	90	22	10662	86	5	10414	85	4	10328
Ephydatia muelleri	108	163	167630	105	144	19880	105	144	17148	9	13	16103	8	6	13542	8	4	12847
Euplectella aspergillum	122	63	20332	121	60	8020	121	60	8020	119	52	7816	114	43	7424	114	12	7097
Euplokamis dunlapae	101	104	170355	100	102	14126	99	14	13939	95	12	11792	84	4	9971	84	4	9971
Fungia scutaria	91	37	90930	91	37	24068	91	20	24030	90	18	23394	88	16	22282	88	16	22244
Gorgonia ventalina	107	434	169692	106	419	25529	106	419	25529	99	174	13992	88	86	10627	86	75	10280
Grantia compressa	89	619	124273	88	533	13196	88	452	13026	85	73	11073	80	34	5949	78	23	5454
Hartaetosiga balthica%15	81	1	19524	81	1	6937	81	1	6937	79	1	6686	77	1	6599	77	1	6599
Hartaetosiga gracilis%05	81	0	24580	81	0	8791	81	0	8791	79	0	8495	78	0	8354	78	0	8354
Helgoeca nana%08	80	1	58606	80	1	10372	80	1	10372	80	1	9744	79	1	9537	79	1	9537
Heterochone calyx	12	6	4790	12	6	2265	12	6	2265	12	6	2262	12	6	2224	12	6	2162
Hormathia digitata	79	33	40375	79	33	12445	79	33	12445	78	27	11969	77	14	11089	76	11	11055
Hormiphora californensis	102	39	100290	102	37	26125	102	37	26125	95	36	25165	92	16	23284	92	16	23284
Hyalonema papilliferum	78	174	34559	76	155	5194	76	155	5194	74	124	4313	71	52	3388	70	7	2988
Hydractinia echinata	635	88	9464	635	84	3212	635	84	3212	624	80	2743	607	78	2517	568	69	2411
Hydractinia polyclina	85	2384	407627	85	2102	28522	85	1899	26914	82	414	15412	78	135	11893	75	71	11276
Ircinia fasciculata	74	128	34868	74	127	11152	73	59	10920	72	58	9928	65	49	7245	60	18	6738
Kirkpatrickia variolosa	77	30	41768	77	25	9210	77	25	9210	76	12	8840	76	7	7893	75	7	7837
Lampea pancerina	79	85	80639	79	80	9796	79	9	9648	77	9	9330	75	5	9001	75	5	9001
Latrunculia apicalis	81	2	32806	80	2	8791	80	2	8751	80	1	8538	76	1	6565	73	1	6404
Leucetta chagosensis	314	9	5703	313	9	2424	313	9	2424	306	6	2250	278	6	1368	266	6	1194
Leuconia nivea	92	925	172384	92	868	15972	92	757	14864	88	499	12835	75	35	6356	74	18	5761
Leuconoides complicata	66	3250	92106	66	3227	37423	65	3227	37423	65	1798	29436	63	729	14081	63	433	10272
Leuconoides variabilis	303	27	7331	302	24	2175	301	24	2172	265	20	1873	265	18	1282	259	16	1205
Liriope tetraphylla	152	382	184056	147	375	23646	143	64	21779	126	33	16731	110	18	15071	100	15	14831
Lubomirskia baicalensis	154	12	2188	154	12	1053	154	12	1053	150	10	1006	142	8	880	141	8	839
Lucernariopsis campanulata	93	149	73120	91	140	11097	90	62	10723	81	21	9997	75	14	7925	73	10	7749
Madracis auretenra	74	707	182599	74	690	34858	61	685	34707	60	361	30514	58	173	27351	58	117	26529
Metridium senile	218	18	6284	217	17	2736	217	17	2736	217	8	2618	212	8	2466	144	8	2238
Ministeria vibrans	68	198	13531	67	162	5302	67	162	5302	55	30	4133	30	13	2974	30	13	2974
Mnemiopsis leidyi_A	89	225	229899	99	224	22720	77	28	19689	77	27	19293	72	15	14142	72	15	14142

Monosiga ovata	981	15	76608	981	14	47271	981	14	47271	975	11	47122	969	11	46107	969	11	46107
Montastraea faveolata_2	62	351	158677	62	344	31422	62	344	31422	62	105	21296	60	96	19724	60	93	19306
Montastraea faveolata	137	83	13498	137	82	3855	137	82	3955	135	63	3744	131	45	3488	126	38	3423
Mycale phyllophila	93	9	111910	93	9	17264	93	9	17264	90	8	16556	89	7	14806	89	6	14190
Mytilosiga fluctuans%19	83	2	31671	83	2	9097	83	2	9097	80	1	9062	67	1	7797	67	1	7797
Nanomia bijuga	225	169	976015	209	127	16197	203	127	16197	196	40	14859	98	10	12620	89	10	12497
Opisetes minuta	24	0	1443	24	0	701	24	0	701	24	0	693	23	0	630	22	0	621
Oscarella carmela	84	177	84729	83	166	11409	9	83	9708	7	10	9085	7	2	6881	6	2	6152
Oscarella tuberculata	17	3	1502	17	2	673	17	2	673	17	2	636	17	2	560	17	2	535
Oscarella sp.	89	8	172354	87	1	18178	87	1	18178	85	0	16927	81	0	14588	75	0	14019
Oscarella sp._SN2011	90	1	64883	90	0	13675	90	0	13675	90	0	12690	88	0	11168	85	0	10751
Pachycerianthus solitarius	170	6	13375	169	6	4097	163	5	3884	158	4	3585	145	3	2892	140	2	2769
Pachydictyum globosum	79	7	1719	79	7	690	79	7	690	75	6	645	67	3	556	67	3	548
Parazoanthus axinellae	132	19	10190	132	19	3903	132	5	3825	131	4	3618	128	2	3081	127	2	2982
Pelagia noctiluca	93	75	111636	82	70	13475	77	7	13329	74	4	12111	66	2	9184	65	3	9038
Pennatula rubra	118	8	12768	118	8	4979	117	6	4914	113	6	4559	92	4	3586	91	4	3511
Periphylla periphylla	80	23	171078	79	20	9878	79	20	9878	76	19	8802	67	5	7068	67	4	7000
Petrosia ficiformis	107	587	270687	95	525	17107	93	471	16797	83	269	13257	79	146	11134	77	32	9755
Physalia physalis	180	1227	404234	178	984	26353	171	916	26104	163	411	18957	154	76	13906	130	49	13577
Pirum gemmata	93	106	70985	83	99	7232	81	59	7166	63	50	5076	47	2	3188	47	2	3188
Plakina jani	92	31	128893	82	28	12564	82	14	12537	82	11	11899	64	2	7759	54	0	6659
Platygyra carnosus	60	275	162149	60	268	27648	60	268	27648	59	34	17823	59	21	15727	59	20	16592
Pleraplysilla spinifera	120	479	114816	116	452	12383	111	235	11625	106	93	10151	97	18	7294	93	15	6435
Pleurobrachia bachei	66	1	19523	66	1	6787	66	1	6787	65	1	6772	64	1	6606	64	1	6606
Pleurobrachia pileus_A	98	242	242292	98	229	21617	98	69	20437	96	12	18033	79	8	13635	79	8	13635
Pleurobrachia pileus	103	6	48892	103	6	16265	103	6	16283	103	1	15931	103	1	15552	103	1	15552
Plumapathes pennacea	85	143	153927	93	126	16255	85	92	15941	78	33	15146	78	9	13458	78	8	13377
Pocillopora damicornis	86	173	190816	85	157	24153	85	157	24153	83	19	19192	78	15	17273	78	15	17206
Podocoryna carnea	462	45	3486	463	43	1849	463	43	1849	454	41	1768	447	39	1656	435	39	1590
Porites astreoides	54	6	7245	54	6	1487	54	6	1487	53	5	1388	52	5	1328	52	3	1274
Porites australiensis	60	236	74998	59	233	18252	58	233	18252	58	20	11340	55	18	10423	54	17	10349
Pseudospongosorites suberitoides	76	15	19278	76	15	6887	77	8	6852	74	9	6458	73	9	5775	73	8	5619
Rossella fibulata	71	26	25240	71	24	6832	71	24	6832	71	15	6676	71	4	6253	71	4	6253
Salpingoeca dolichothecata%16	92	2	52403	81	1	12523	91	1	12523	80	0	11172	68	0	9860	68	0	9860
Salpingoeca helianthica%18	80	0	32829	80	0	9147	80	0	9147	79	0	8703	77	0	8472	77	0	8472
Salpingoeca infusionum%12	84	0	45247	84	0	8994	84	0	8994	83	0	8958	82	0	8427	82	0	8427
Salpingoeca macrocollata%06	90	0	44734	80	0	11088	90	0	11088	77	0	10055	73	0	9302	73	0	9302
Salpingoeca punica%03	90	4	40846	80	4	8989	80	4	8989	78	0	8126	69	0	7393	69	0	7393
Salpingoeca qwerti%09	90	0	25813	80	0	9698	90	0	9698	76	0	8906	65	0	7760	65	0	7760
Salpingoeca rosetta%13	84	2	43523	84	2	9388	84	2	9388	84	1	8803	82	1	8633	82	1	8633
Salpingoeca urceolata%04	95	3	57706	93	3	11972	93	3	11972	90	3	11913	79	3	10080	79	3	10080
Savillea parva%14	90	4	47054	80	4	9926	80	4	9926	79	1	9451	79	1	9041	79	1	9041
Sphaeroforma arctica_JP610	78	43	47232	78	39	9264	78	39	9264	60	4	5244	45	3	3393	45	3	3393
Spongilla lacustris	91	334	56696	81	325	13579	81	268	13192	76	228	12222	67	36	8578	65	23	8074
Stephanoeca diplocostata_AU	83	1	41143	82	1	12349	82	1	12349	81	0	11690	81	0	11649	81	0	11649
Stephanoeca diplocostata_FR	93	0	80355	82	0	14702	92	0	14702	81	0	13933	80	0	12311	80	0	12311
Stephanoeca diplocostata	93	3	76911	93	3	13147	93	3	13147	92	0	12434	90	0	12413	90	0	12413
Stomolophus meleagris	61	230	116942	61	229	29064	61	229	29064	60	137	28168	77	8	24690	77	7	18920
Suberites domuncula	289	16	13304	288	16	7277	288	16	7277	287	16	7150	282	13	6744	218	6	5522
Sycon ciliatum	121	16996	776523	118	16544	177332	118	16544	177332	104	5721	84797	95	2304	28467	89	977	17925
Sycon coactum	74	770	70220	73	758	13759	25	645	13057	25	289	11157	23	65	5692	23	38	5079
Sycon raphanus	163	3	1629	163	3	772	163	3	772	159	1	720	109	1	425	109	1	415
Sympagella nux	81	31	47128	81	29	8686	81	29	8686	79	24	8456	77	13	8060	76	2	7101
Tethya wilhelma	90	45	26057	89	44	6523	89	44	6523	87	15	5688	83	9	5295	82	6	5101
Tripedalia cystophora	82	11	10256	82	9	3657	82	9	3657	76	6	3360	65	5	2767	64	5	2728
Urticina eques	129	89	38965	129	86	7495	129	86	7495	127	25	6788	123	12	6334	122	8	6302
Vallicula multiformis_A	90	435	194558	90	432	23904	90	68	11793	88	60	10236	80	10	8822	80	10	8822
Vallicula multiformis	93	238	79493	90	233	9699	96	15	9739	77	10	8526	75	6	8170	75	6	8170

Integration of transcriptomic data into core orthologous clusters

We completed the 4,002 core clusters with the corresponding homologous sequences from the decontaminated transcriptomes. In order to only add orthologous sequences and facilitate future treatment of paralogy within the clusters, this procedure includes a first preliminary assessment of orthology using a multiple Best Reciprocal Hit (multi-BRH) strategy against reference proteomes. The procedure allows the importation of several sequences per species, to permit the subsequent detection and removal of paralogs. The procedure is implemented in the software Forty-Two available as a Bitbucket repository (https://bitbucket.org/dbaurain/42/).

For each of the 4,002 core clusters:

1. Sequences from selected core species ("query species") present in the core cluster were BLASTed against the transcriptome to be integrated, and matching sequences with an e-value threshold of $1e^{-10}$ were retained as "homologous transcripts". The query species were: *Amphimedon queenslandica*, *Monosiga brevicollis*, *Lottia gigantea* and *Nematostella vectensis*.

2. The same query sequences were BLASTed against specified reference proteomes leading to the extraction of one (or several nearly) best hit(s) from each reference proteome. Indeed, when several best hits from the same reference proteome had nearly equal bit scores (within 99% of the bit score of each preceding best hit), all of them were retained. This feature is particularly important as such equal hits often correspond to in-paralogs or to close out-paralogs. Retaining all those paralogs will later facilitate their detection and, if necessary, subsequent suppression. The reference proteomes for this step were those of *Amphimedon queenslandica*, *Monosiga brevicollis*, *Lottia gigantea*, *Homo sapiens*, *Nematostella vectensis* and *Batrachochytrium dendrobatidis*.

3. Each of the "homologous transcripts" identified in step 1 was BLASTed against the same reference proteomes, resulting in the creation of a list of "homologous best hits". Whenever the best hit for each reference proteome was also found in the list of best hits from step 2, the homologous transcript was considered to be an orthologue.

4. Orthologues were BLASTed against the entire cluster in order to detect their closest relative sequences. Out of the five closest sequences, the one that aligned with the longest part of the orthologue was chosen as the most appropriate template to be used for aligning the newly identified orthologue with the cluster sequences.

Because *S. ciliatum* and *L. complicata* transcriptomes remain seriously contaminated (see above list of numbers of sequences at each step of the cleaning procedure for RIBO-80 as an example), these species were added after the other transcriptomes and using a very stringent taxonomic filter: we

only added an orthologous transcript when its closest relative already present in the orthologous cluster was a calcarean sponge.

Removal of paralogs

Removal of orthologous clusters containing too many paralogues

Paralogy detection and pruning is a complex phylogenetic problem that we tackled by categorizing paralogous sequences into three types, based on the 10 previously defined major clades:

- In-paralogues correspond to paralogous sequences that group together (= multiple sequences from the same species forming a monophyletic group in the single-gene tree). These paralogues are adequately handled by automated software, such as SCaFoS [S8] and are not problematic for phylogenetic inference.
- Out-in-paralogues correspond to paralogous sequences that disrupt the monophyly of a given species in the single-gene tree, but do not disrupt the clade (e.g., 2 sequences of *Acropora* that belong to a clade containing only Anthozoa) and therefore should not hamper the correct inference of phylogenetic relationships between the 10 clades.
- Out-out-paralogues correspond to paralogous sequences that disrupt the monophyly of an entire clade in the single-gene tree and therefore have a deleterious impact on the inference of phylogenetic relationships between the 10 major clades.

We first cleaned each alignment using HMMCleaner and eliminated all the positions having less than 5% of known character states. To reduce stochastic error, only sequences with at least 50 parsimony-informative positions were used to infer the single-gene trees with RAxML and the LG+Γ_4+F model. Single-gene trees were analysed to detect out-in-paralogues and out-out-paralogues. We discarded 380 alignments for having > 1 clade disrupted by out-out-paralogues, and 411 alignments for having ≥ 2 clades affected by out-in-paralogues. We therefore retained a total of 2,424 alignments, of which 931 were totally devoid of any out-in- or out-out-paralogous sequences.

Removal of paralogous sequences

The 1,493 alignments affected by out-paralogy (either out-in or out-out) were specifically processed. Each of them was divided into up to 10 sub-alignments corresponding to the 10 pre-defined clades. Every sub-alignment with an out-paralogy issue (2,121 among 14,930) and with ≥ 3 species was cleaned with HMMCleaner prior to building a tree using RAxML (LG+Γ_4+F model). For each tree, we automatically retained the largest (in term of species number) possible clade (a partition of the unrooted tree) that did not contain any out-paralogues (i.e., each species represented only once or only by in-paralogs). All selected sequences from all sub-alignments were then re-assembled into a new alignment. However, because of tree reconstruction errors, it is possible that

some out-paralogy remained undetected. Therefore, phylogenies were anew inferred from the re-assembled alignments (HMMCleaner, BMGE, RAxML LG+Γ_4+F), and 78 alignments displaying ≥ 2 clades affected by out-paralogy were discarded, resulting in 2,346 remaining alignments.

Phylogenetic analyses: site-homogeneous vs. site-heterogeneous sequence evolution models

The LG+Γ_4+F sequence evolution model is referred to as *site-homogeneous*, whereas the CAT+Γ_4 model is referred to as *site-heterogeneous*. Indeed, a LG+Γ_4+F model applies the same exchangeability rates (i.e. the LG component) and the same stationary frequencies (i.e. the +F component) to the entire supermatrix. All sites are therefore analysed with the same parameters (although they can evolve more or less rapidly thanks to the +Γ_4 component).

It is possible to partition a supermatrix (e.g. by gene, codon position, or using software such as PartitionFinder) and to apply different site-homogenous models such as JTT, WAG or LG to partitions (see ref. [S1, S9]). Such partitioning allows to modify exchangeability rates and/or stationary frequencies across partitions. However, such a mixture of site-homogeneous models can not be regarded as site-heterogeneous. Indeed, all sites belonging to a given partition are still analysed under the same set of parameters. This method relies on the inadequate assumption that positions partitioned together evolve in the same way. In fact, two given positions of the same gene (e.g. in the same partition defined by PartitionFinder) often show *extremely different* stationary frequencies and/or exchangeability rates. Such site-specificity can not be modelled in analyses using site-homogeneous models, even when partitioned as above.

The CAT+Γ_4 model is referred to as site-heterogeneous because it takes into account that two sites – even when belonging to the same partition – will not necessarily evolve according to the same exchangeability rates and/or stationary frequencies. This is done by detecting positions that *effectively* evolve the same way, and attributing them to a dedicated *profile* that will best describe their common specificity (e.g. “this position only accepts hydrophobic amino-acids”). This can not be done using general WAG or LG models (conversely to implicit claims of a recent study [S10]) because such homogeneous replacement matrices allow any amino-acid to potentially occur at any position.

By taking into account the *biological fact* that often only a reduced diversity of amino-acids occurs at a given site (see Figure 3 in [S11]), the CAT model will more easily infer reverse replacement

between these amino-acids and will therefore more accurately estimate saturation in supermatrices. Since detecting saturation is the key to correctly handle fast-evolving lineages in phylogenetic inference, it is not surprising that discrepancies regarding the phylogenetic position of long-branches (e.g. ctenophores, acoelomorphs) are commonly observed when comparing topologies obtained using site-homogeneous vs. site-heterogeneous models.

Supplemental References

S1. Whelan, N.V., Kocot, K.M., Moroz, L.L., and Halanych, K.M. (2015). Error, signal, and the placement of Ctenophora sister to all other animals. Proc. Natl. Acad. Sci. USA *112*, 5773-5778.

S2. Moroz, L.L., Kocot, K.M., Citarella, M.R., Dosung, S., Norekian, T.P., Povolotskaya, I.S., Grigorenko, A.P., Dailey, C., Berezikov, E., Buckley, K.M., et al. (2014). The ctenophore genome and the evolutionary origins of neural systems. Nature *510*, 109-114.

S3. Ryan, J.F., Pang, K., Schnitzler, C.E., Nguyen, A.D., Moreland, R.T., Simmons, D.K., Koch, B.J., Francis, W.R., Havlak, P., Program, N.C.S., et al. (2013). The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. Science *342*, 1242592.

S4. Borowiec, M.L., Lee, E.K., Chiu, J.C., and Plachetzki, D.C. (2015). Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. BMC Genomics *16*, 987.

S5. Kircher, M. (2012). Analysis of high-throughput ancient DNA sequencing data. Methods Mol. Biol. *840*, 197-228.

S6. Laurin-Lemay, S., Brinkmann, H., and Philippe, H. (2012). Origin of land plants revisited in the light of sequence contamination and missing data. Curr. Biol. *22*, R593-594.

S7. Philippe, H., Brinkmann, H., Lavrov, D.V., Littlewood, D.T., Manuel, M., Worheide, G., and Baurain, D. (2011). Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. *9*, e1000602.

S8. Roure, B., Rodriguez-Ezpeleta, N., and Philippe, H. (2007). SCaFoS: a tool for Selection, Concatenation and Fusion of Sequences for phylogenomics. BMC Evol. Biol. *7 Suppl 1*, S2.

S9. Halanych, K.M., Whelan, N.V., Kocot, K.M., Kohn, A.B., and Moroz, L.L. (2016). Miscues misplace sponges. Proc. Natl. Acad. Sci. USA *113*, E946-947.

S10. Whelan, N.V., and Halanych, K.M. (2016). Who Let the CAT Out of the Bag? Accurately Dealing with Substitutional Heterogeneity in Phylogenomic Analyses. Syst. Biol. [e-pub ahead of print]

S11. Lartillot, N., and Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. *21*, 1095-1109.

Table des matières