Étude des processus introgressifs en évolution par des méthodes de réseaux

Raphaël Méheust

Université Pierre et Marie Curie

École doctorale Complexité du Vivant

Évolution Paris Seine - UMR 7138

Équipe Adaptation, Intégration, Réticulation et Évolution

Étude des processus introgressifs en évolution par des méthodes de réseaux

Par Raphaël MÉHEUST

Thèse de doctorat en Biologie Évolutive

Dirigée par Éric BAPTESTE et Philippe LOPEZ

Présentée et soutenue publiquement le 09 décembre 2016

Devant un jury composé de :

Dr Éric Bapteste (DR CNRS, Université Pierre et Marie Curie)	Encadrant
Dr Chris Bowler (DR CNRS, ENS Paris)	Examinateur
Pr Abdelaziz Heddi (PR, INSA Lyon)	Rapporteur
Pr Eugene Koonin (Senior Investigator, NCBI)	Examinateur
Pr Philippe Lopez (PR, Université Pierre et Marie Curie)	Encadrant
Dr Claudine Médigue (DR CNRS, Génoscope)	Rapporteur
Pr Michel Morange (PR, Université Pierre et Marie Curie)	Examinateur
Dr Fabrice Not (DR CNRS, Université Pierre et Marie Curie)	Examinateur

Remerciements

Même si à l'heure où j'écris ces lignes mes envies oscillent entre partir élever des moutons dans les Alpes et devenir gérant d'une micro-brasserie, j'ai passé trois très bonnes années à Paris. Et je le dois en grande partie à Éric et Philippe. Merci à vous deux, et désolé pour mes sauts d'humeur!

Merci aux collègues de l'équipe. Merci Eduardo (et Philippe), j'ai fait d'énorme progrès en contrepèteries grâce à toi, d'ailleurs la mairie de Paris s'y essaye... Merci aux anciens doctorants de l'équipe, Cédric pour ta gentillesse, Pierre-Alain et Guifre pour les nombreuses discussions. Jananan, Chloé et Mathis, je vous dis merci pour tous les moments partagés ces deux dernières années.

Je remercie l'ensemble des doctorants du labo anciens et actuel pour les bons moments passés ensembles dont nous tairont certains! Merci donc à Kader, Marie, Joao. Merci à Thomas d'avoir eu la bonne idée d'organiser des SCEP. Merci Thomas, Gabriel, Juliette, Camille, Marie-Pierre, Arnaud, Anne-Sophie.

Merci à l'ensemble des visiteurs du laboratoire, qu'ils soient passés quelques heures, quelques jours ou plusieurs semaines : ce fut un plaisir d'apprendre, d'échanger ou de collaborer avec vous. Merci à Debashish Bhattacharya, James McInerney, Mary O'Connell, Fabini, Thane Papke, Étienne, Richard, Chris Lane, Lucie, Laurent Viennot, Michel Habib, Martin Embley, et bien d'autres.

Je remercie également l'ED515 pour la fluidité de nos échanges par email. Ça fait chaud au cœur de se sentir soutenu par les instances de l'UPMC!

Je remercie ma concierge, ma boulangère ainsi que Stéphanie qui est une fille vraiment sympa et pétrie de qualité!

Je remercie ma famille et mes amis de toujours.

Je remercie les membres de mon comité de thèse pour leur aide.

Un grand merci à Danielle!

Merci aux membres de mon jury de thèse de prendre de leurs temps précieux pour évaluer mon travail.

Enfin last but not least, je remercie également le ou la référé(e) anonyme 1 de l'article sur l'eucaryogenèse pour ces remarques constructives qui ne pourront qu'améliorer sensiblement

la qualité de l'article! Un grand merci à vous! C'est grâce à des personnes comme vous que le métier de chercheur est l'un des plus beaux du monde!

Et voilà "g ksa ka dire"®.

Bonne lecture aux courageux.

Sommaire

Liste des figures

I. Les processus introgressifs en évolution et l'utilisation de réseaux pour les étudier

Les évolutionnistes sont confrontés à plusieurs défis : comprendre comment la diversité biologique observée aujourd'hui est structurée (l'*explanandum*) et en expliquer les causes (les *explanans*).

Le processus classique pour expliquer cette structuration découle des travaux de Charles Darwin : les êtres vivants évoluent depuis des formes antérieures par modification et divergence selon un processus arborescent [1]. La notion de descendance avec modification appelée aussi "descendance verticale" décrit ce processus durant lequel le matériel génétique d'un objet biologique modifié par des mutations va se propager par réplication vers sa descendance. Au-delà de la perspective processuelle, le modèle arborescent est aussi un formidable outil pour classifier les différents groupes d'organismes entre eux. Or le modèle arborescent n'est pas suffisant pour expliquer les 3,7 milliards d'années d'évolution [2]. D'autres processus existent. Contrairement au processus arborescent, dans les processus introgressifs, le matériel génétique d'un hôte est transmis vers un nouvel hôte puis se réplique dans ces structures [3]. Les processus introgressifs sont importants en évolution car ils affectent différents niveaux d'organisation biologiques (Figure 1): les séquences, les génomes ou encore les organismes. Un gène peut être le résultat de la combinaison de plusieurs gènes, un génome celui de la combinaison de gènes d'origines distinctes, un organisme le résultat de la combinaison de plusieurs génomes etc. Comment étudier toutes ces réticulations au niveau génomique et dans plusieurs génomes à la fois ? Les méthodes de réseaux [4–6] semblent un bon moyen pour détecter, analyser et de visualiser beaucoup de ces phénomènes.

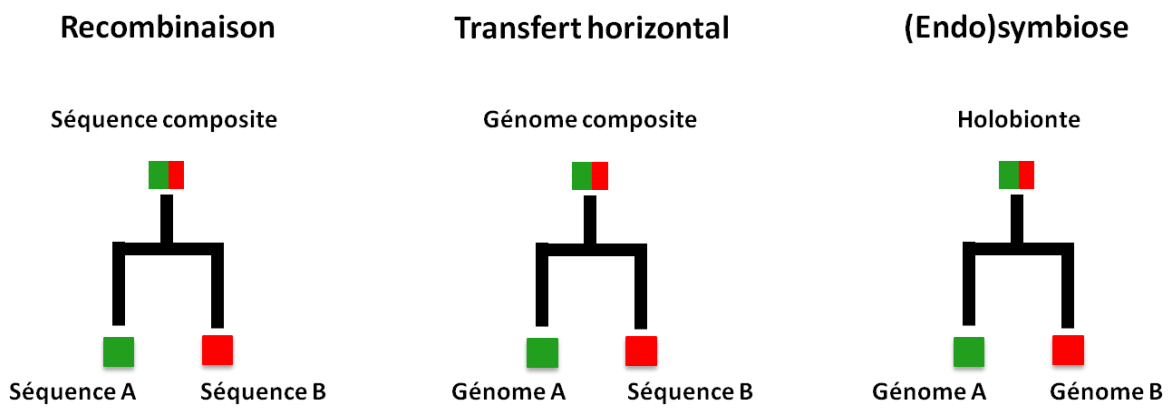


Figure 1 : Création d'objets chimériques à différents niveaux.

II. Les endosymbioses et leurs impacts en évolution

A. La théorie endosymbiotique, de sa réfutation à son acceptation.

Aujourd'hui il est bien accepté par la communauté scientifique que la plupart des organismes vivants, qu'ils soient microscopiques ou macroscopiques, vivent en associations avec d'autres organismes. Ces associations intimes et durables entre organismes hétérospécifiques sont appelés symbioses et les découvertes faites ces dernières années montrent à quel point la symbiose n'a rien d'un phénomène exceptionnel. C'est même plutôt la règle dans le vivant [7]. Les symbioses sont partout et impliquent tous types d'organismes et de virus [4,7–38]. Elles peuvent être tellement intriquées que les partenaires symbiotiques forment un tout indissociable [39]. Enfin, certaines d'entre elles ont été essentielles dans certaines transitions évolutives majeures [40,41]. Malgré tout, l'intérêt pour l'étude des associations entre organismes est assez récent et fut longtemps une discipline secondaire en biologie. Cette première section a pour but de faire un rappel historique sur la symbiose en se concentrant sur l'histoire de la théorie endosymbiotique, qui est un des axes principaux de cette thèse.

Le mot symbiose (du Grec ensemble et vivre) apparait pour la première fois en 1877 dans un texte du biologiste allemand Albert Bernhard Franck sous le terme Symbiotismus : "*we must bring all the cases where two different species live on or in one another under a comprehensive concept which does not consider the role which the two individuals play but is based on the mere coexistence and for which the term Symbiosis [Symbiotismus] is to be recommended*". Ce terme sera peu à peu accepté par la communauté scientifique, suite notamment à la redéfinition faite par Anton De Bary en 1878 dans "The phenomena of Symbiosis" : "*the living together of unlike organisms.*". L'apparition du terme symbiose à la fin du XIX^{ème} siècle n'est évidemment pas dû au hasard et a largement émergé suite à la découverte de la nature mosaïque des lichens. En 1867, le suisse Simon Schwendener fait l'hypothèse que les lichens résultent de l'association entre un champignon et un organisme photosynthétique. Bien que rejetée par de nombreux biologistes, sa nature composite sera démontrée plus tard notamment par le russe Andrei Sergeyevich Famintsyn qui réussira à séparer et à cultiver les deux composantes de cette association. En 1875, le biologiste Pierre Joseph Van Beneden dans son livre "les commensaux et les parasites" définit les différents types d'associations symbiotiques (parasites, mutualistes et commensales). A la fin du XIX^{ème} siècle, les découvertes d'associations durables et profitables chez les lichens, les

anémones de mer, les radiolaires ou encore chez les légumineuses et plus généralement chez les plantes mettent fin à la vision du "tout parasitisme".

La mise en évidence d'associations mutuellement bénéfiques ouvre la voie à de nouvelles recherches notamment concernant l'interprétation de l'origine des constituants de la cellule eucaryote, les organites. Le botaniste allemand Andreas Schimper suggère, dans une note en bas de page, pour la première fois en 1883, une origine bactérienne des chloroplastes chez les eucaryotes photosynthétiques. Au même moment, un autre allemand, Altmann étudie de petits granules présents dans les cellules eucaryotes et qu'il appellera "bioblast" avant que Theodor Boveri ne les renomme mitochondries. Ces deux événements sont les prémices de la future théorie endosymbiotique qui stipule que la mitochondrie et le chloroplaste sont d'anciennes bactéries symbiotiques de la cellule eucaryote. Entre ces premières découvertes et l'acceptation de la théorie marquée par le célèbre papier de Ford Doolittle and Michael Gray [42], il s'écoulera 100 ans. Les raisons sont nombreuses pour expliquer cette durée mais d'après le livre de Jan Sapp, "Evolution by association: a history of symbiosis" [43], d'où est tirée en grande partie de cette introduction, les raisons techniques et idéologiques semblent avoir été deux obstacles importants à la progression de cette théorie. Par exemple, la vision nucléocentrique du début du XX^{ème} siècle (i.e. toutes les structures nucléoplasmiques sont formées grâce au noyau et par conséquent en dérivent) ou encore la vision "tout pathogène" des microbes. Une autre raison est le manque d'intérêt pour ces études : la symbiose reste pour la majeure partie des scientifiques de l'époque un phénomène rare et suscite peu d'intérêt en comparaison avec l'essor de la génétique à la fin du XX^{ème} siècle, suite à la redécouverte des travaux de Mendel. Bien que minoritaire lors de la première moitié du XX^{ème} siècle, l'étude de l'évolution de la cellule ou des symbioses trouve quelques adeptes comme le russe Constantin Merezhkowsky et le français Paul Portier qui vont contribuer à l'étude de l'origine des organites.

Constantin Merezhkowsky est un biologiste russe qui étudie la structure cellulaire des diatomées, il s'intéresse particulièrement à leurs chloroplastes. En combinant ses connaissances en biologie cellulaire avec les idées développées à l'époque sur la symbiose, il rassemble dans son célèbre papier de 1905 les preuves en faveur de l'origine symbiotique des chloroplastes. Il est notamment le premier à montrer les similarités remarquables entre les chloroplastes et un groupe de bactéries libres, les *cyanophyceaea*, groupe qui sera renommé plus tard les cyanobactéries. Pour Merezhkowsky, les liens de parenté entre ces deux entités sont évidents, les cyanobactéries et les chloroplastes sont petits, de forme ronde, ayant une

couleur oscillant entre le vert et le bleu, ils ne possèdent pas de noyau, prolifèrent par division et sont tous les deux capables d'assimiler le carbone. Pour le scientifique russe, les cyanobactéries sont des formes libres de chloroplastes. Aujourd'hui Merezhkowsky est considéré comme le père fondateur de la théorie endosymbiotique mais il ne faut pas oublier que d'autres chercheurs ont contribué (dans une moindre mesure) à ces recherches comme Andreas Schimper ou encore d'Andrei Sergeyevich Famintsyn. Merezhkowsky ne suggéra jamais la possibilité que la mitochondrie soit elle aussi d'origine bactérienne. Cette proposition fut l'œuvre du français Paul Portier.

Paul Portier travaillait à l'institut océanographique de Monaco. C'est un scientifique reconnu qui a contribué à la découverte de l'anaphylaxie, qui vaudra un prix Nobel à son mentor Charles Richet. Ses centres d'intérêt sont très divers. En 1918, il publie "les symbiotes" où il propose que la symbiose soit une caractéristique primordiale de la cellule eucaryote. Pour Portier, les cellules eucaryotes fonctionnent à l'aide de symbiotes! Les mitochondries sont pour lui des bactéries vivant à l'intérieur de la cellule eucaryote, elles peuvent donc être séparées et cultivées comme pour les lichens, ce qu'il affirme avoir réalisé chez différents organismes. Inutile de dire que son livre provoqua de vives contestations en France dont l'illustration la plus marquante fut la publication, un an plus tard, par Auguste Lumière, l'inventeur du cinématographe, du "Mythe des Symbiotes". Dans son livre, Auguste Lumière critique la vision de Portier arguant que les bactéries sont avant tout pathogènes. Portier sera mis sous pression par les chercheurs de l'Institut Pasteur pour qu'il confirme ses résultats avec eux, ce qu'il n'arrivera jamais à faire; les mitochondries qu'il pensait avoir réussi à cultiver séparément des cellules eucaryotes étaient en fait des contaminations. Quelques mois après la publication des symbiotes, la réputation de Paul Portier est défaite et sa théorie ignorée. Les travaux de Portier, bien que très décriés, ont permis de populariser l'étude des symbioses dans les pays anglo-saxons à travers les travaux d'Ivan Wallin notamment. Ivan Wallin reprend les travaux de Portier, il pense que les mutations ponctuelles ne peuvent pas être à l'origine d'espèces nouvelles; pour lui, seule la symbiose peut provoquer de grands changements. Pour cela, il s'appuie sur les travaux du biologiste allemand Paul Buchner, considéré comme le fondateur de la "systematic symbiosis research". Paul Buchner a montré chez les insectes que les symbioses pouvaient créer des variations importantes. Par exemple, de nombreux insectes possèdent des cellules spécialisées pour garder les endosymbiontes. Comme pour Portier, le livre d'Ivan Wallin, " Symbionticism and the origin of species " (1927) fut en grande partie ignoré par les chercheurs qui s'intéressaient aux

nouvelles découvertes en génétique de son époque. En 1930, le généticien américain et futur prix Nobel Hermann Muller montre le lien entre hérédité et gènes, les gènes se trouvent dans le noyau et non dans le cytoplasme. De plus, les travaux sur la radioactivité montrent que les rayons X peuvent induire des mutations et des changements phénotypiques aussi puissants que ceux attendus par la génération spontanée. Le généticien américain Thomas Hunt Morgan conclura que "*the cytoplasm may be ignored genetically*". Dans la même période, la théorie synthétique de l'évolution réunit les travaux de Darwin et ceux de Mendel. L'idée de la symbiose comme source d'innovation apparait fantaisiste pour la plupart des biologistes de l'époque. Depuis ces prémices jusqu'aux années 1950, la théorie endosymbiotique et plus généralement la symbiose sont restées très peu étudiées. Il faudra attendre la repopularisation de la théorie endosymbiotique par la biologiste américaine Lynn Margulis et l'essor de la biologie moléculaire pour les remettre au gout du jour.

A la différence de ses prédécesseurs, Lynn Margulis est arrivée avec ses grandes idées au bon moment, enfin presque. Les débuts scientifiques de Lynn Margulis sont modestes, ses premières publications décrivent la présence d'ADN dans le cytoplasme de différents eucaryotes unicellulaires. Un peu plus tard, Hans Ris et Walter Plant fournissent la preuve de la présence d'ADN dans le chloroplaste de *Chlamidomonas reinhardtii*. En 1963, Sylvan Nass et Margit Nass feront de même pour la mitochondrie. Hans Ris et Walter Plant étaient clairement au courant des travaux de Merezhkowsky et de Wallin mais comme le duo Nass, ils restèrent très prudents concernant l'interprétation de ces résultats pour le moins étonnants. Lynn Margulis n'a que faire de la prudence, c'est une "think big person" qui a très bien compris ce que signifiaient ces résultats. En 1967, elle publie sous le nom de Lynn Sagan, dans le Journal of Theorical Biology, un article de 56 pages "On the origin of mitosing cells". Alliant des informations de disciplines diverses incluant la génétique, la bactériologie, la biologie cellulaire, l'écologie et la paléontologie, elle formula que la mitochondrie, le chloroplaste et le corps basal du flagelle sont le résultat d'anciennes symbioses. La publication en 1970 du livre de vulgarisation "Origin of Eukaryotic Cells" sur la théorie endosymbiotique permettra de populariser sa théorie chez les scientifiques de l'époque.

Malgré le travail de Lynn Margulis, les réticences demeurent. De nombreux scientifiques restent en faveur d'une origine autogène de la mitochondrie et du chloroplaste plutôt qu'exogène. Les avancées réalisées en biologie moléculaire vont clore le débat. Au milieu des années 1970, Carl Woese réalise les premières phylogénies moléculaires incluant des procaryotes et des eucaryotes en utilisant l'ARN ribosomal (ARNr 16S) comme

marqueur, découvrant par cette occasion que les procaryotes sont en fait constitués de deux groupes: les bactéries et un nouveau domaine du vivant, les Archaea. Cette découverte est essentielle car elle montre que l'on peut inférer l'origine d'une séquence nucléique. Les équipes de Ford Doolittle et de Mickael Gray, toutes deux basées à l'université de Dalhousie au Canada, utilisent la technique développée par Carl Woese pour définitivement prouver que l'ADN présent dans le chloroplaste et la mitochondrie est d'origine bactérienne. Un an plus tard, il sera démontré que l'ADN chloroplastique est proche de celui des cyanobactéries comme l'avait suggéré Merezhkowsky au début du siècle. A la fin des années 1970, l'équipe de Woese identifiera que l'ADN mitochondrial se rapproche de celui des alpha-protéobactéries.

Entre la première suggestion de l'origine bactérienne d'un des organites cellulaires par Andreas Schimper et l'acceptation de la théorie endosymbiotique par la communauté des biologistes, marquée par le papier célèbre de Ford Doolitle and Michael Gray, il se sera écoulé 100 ans. Aujourd'hui cette théorie est totalement acceptée et les découvertes faites ces dernières années montrent que les symbioses ont un rôle majeur en évolution.

B. Les origines de la cellule eucaryote

L'origine bactérienne de la mitochondrie et du chloroplaste est d'autant plus intéressante que l'acquisition de ces deux endosymbiontes est liée à l'origine des eucaryotes dans le cas de la mitochondrie et à l'acquisition de la photosynthèse chez les eucaryotes dans le cas du chloroplaste [44].

1. Singularités des eucaryotes par rapport aux procaryotes.

L'eucaryogénèse, les processus conduisant aux premières cellules eucaryotes, est un des évènements les plus difficiles à résoudre en évolution [45–48]. Comment une cellule aussi singulière a-t-elle pu émerger d'organismes procaryotes il y a plus de 2 milliards d'années? A la différence des Archaea et des Bactéries, les cellules eucaryotes sont 1000 à 10000 fois plus volumineuses et très compartimentées [45]. Cette différence de taille affecte le fonctionnement de la cellule eucaryote : alors que chez la plupart des Archaea et des Bactéries, les molécules se déplacent par diffusion, les cellules eucaryotes utilisent de nombreux systèmes de transport afin d'adresser les molécules dans les nombreux compartiments qu'elles possèdent. Cette singularité est d'autant plus surprenante que les études de génomique comparative ont montré que le dernier ancêtre commun aux eucaryotes (LECA pour Last Common Eukaryotic Ancestor) n'avait rien d'une cellule primitive et

possédait déjà toutes les caractéristiques des cellules eucaryotes modernes [49]. Comprendre les processus évolutifs ayant aboutit aux eucaryotes est donc un défi majeur.

2. Théories sur l'eucaryogenèse

De nombreux scénarios ont été proposés pour expliquer l'origine des eucaryotes [47]. L'ensemble de ces théories s'accorde pour dire que les premières cellules eucaryotes possédaient déjà un noyau, une mitochondrie, un système endomembranaire, la méiose et la mitose ainsi qu'un cytosquelette complexe qui leur permettait de faire de la phagocytose. Ce qui les différencie, c'est l'ordre et l'importance de ces évènements dans la construction de la cellule eucaryote [44]. L'acquisition de la mitochondrie est un point central du débat et les scénarios peuvent être séparés en deux types : les modèles qui placent l'acquisition de la mitochondrie comme l'évènement final de l'eucaryogénèse ("mitochondrion-late") et ceux qui affirment que l'eucaryogenèse résulte de l'association de deux procaryotes, l'hôte et l'endosymbionte, plaçant ainsi l'endosymbiose de l'alpha-protéobactérie comme l'événement fondateur de l'eucaryogénèse ("mitochondrion-early") [44].

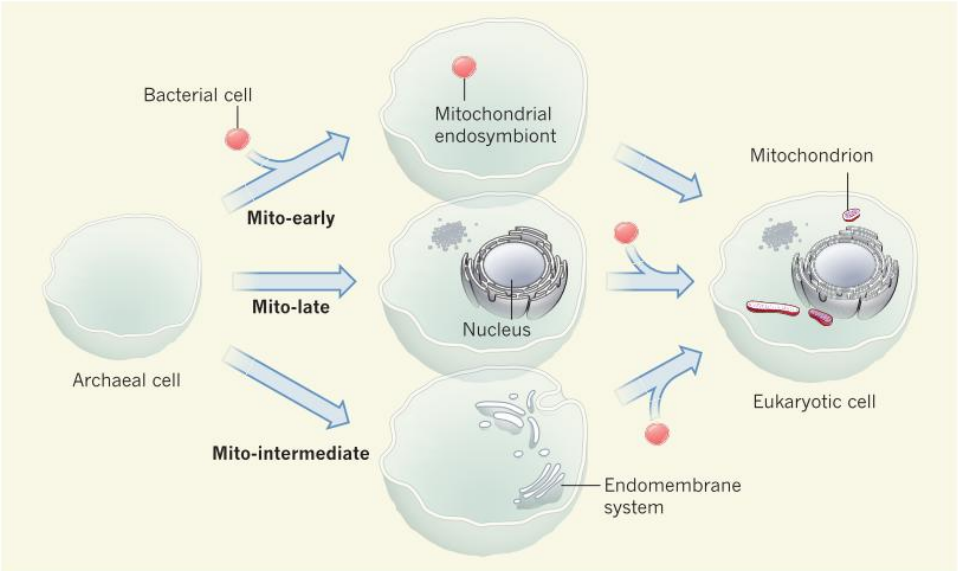


Figure 2. Origine de la cellule eucaryote et de la mitochondrie (source : [50]**)**

a) « Mitochondrion-late »

Ce qui différencie ces deux modèles c'est finalement la nature de l'hôte qui a acquis l'endosymbionte. Était-il un procaryote, un eucaryote ou un organisme intermédiaire (proto-eucaryote) ? Dans les années 1980, Thomas Cavalier-Smith proposa que l'hôte était déjà un eucaryote en se basant sur deux observations. Plusieurs groupes d'eucaryotes

unicellaires comme les microsporidies, le groupe des entamoeba, les diplomonades ou encore les parabasalides ne possèdent pas de mitochondries et les phylogénies réalisées à l'époque plaçaient ces groupes à la racine des autres eucaryotes suggérant ainsi que ces organismes sans mitochondries, regroupés sous le nom d'Archezoa, avaient divergé avant l'acquisition de la mitochondrie. Par conséquent, les Archezoa ont été considérés comme les descendants d'un proto-eukaryote phagotrophe possédant un noyau. L'accord entre données moléculaires et structurales semblait se montrer clairement en faveur d'une acquisition tardive de la mitochondrie et l'étude des Archezoa promettait d'importantes avancées dans la compréhension de l'eucaryogénèse. Cependant, l'hypothèse Archezoa de Thomas Cavalier-Smith fut réfutée à la fin des années 1990 pour deux raisons. D'une part, les avancées en phylogénétique ainsi que l'amélioration de l'échantillonnage des eucaryotes montra que le groupe des Archezoa était en fait polyphylétique et que le placement des organismes sans mitochondrie à la racine des eucaryotes était en fait dû à des attractions de longues branches [48]. D'autre part et de manière plus importante, les organismes sans mitochondries sont en fait des lignées eucaryotes ayant perdu secondairement leurs mitochondries [51,52]. Ce résultat est essentiel car il dit que tous les eucaryotes connus actuellement ont évolué à partir d'un ancêtre possédant une mitochondrie, replaçant l'acquisition de la mitochondrie au centre des débats.

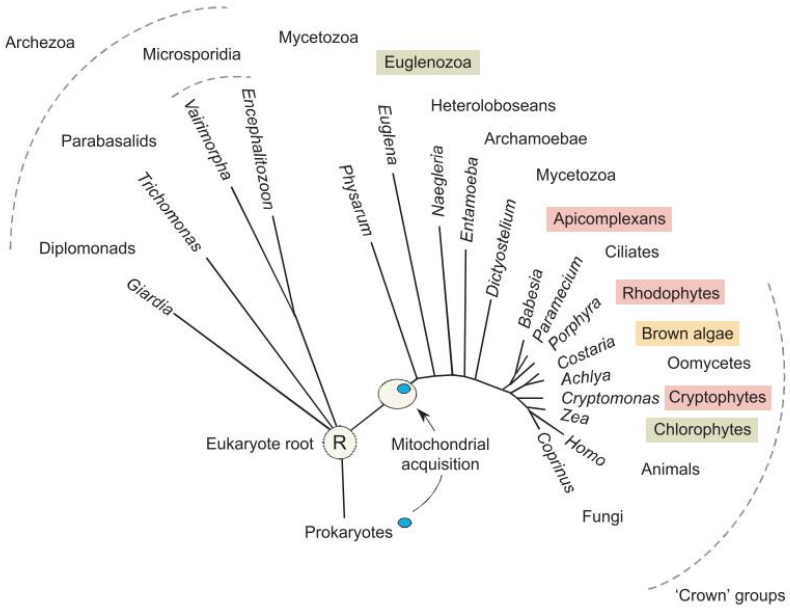


Figure 3. La théorie "Archezoa" de Thomas Cavalier-Smith (source : [51]**)**

b) « Mitochondrion-early »

L'absence d'organismes sans mitochondrie ayant divergé avant l'émergence des premiers eucaryotes a poussé les chercheurs à s'interroger sur la nature des processus évolutifs responsables de cette transition évolutive majeure et à repenser leurs modèles. La grande majorité des modèles actuels suppose que l'hôte était une Archaea.

(1) De trois à deux domaines primaires du vivant

La représentation la plus populaire de l'arbre du vivant est l'hypothèse des trois domaines [53]. Selon ce scénario, les organismes cellulaires se divisent en trois grands groupes monophylétiques, les Bactéries, les Archaea et les Eucaryotes [54]. Cette division remonte à la première phylogénie moléculaire réalisée par Carl Woese et son équipe en 1977, sur la base de marqueurs universels d'ARNr 16S [54] puis confirmée par d'autres marqueurs moléculaires conservés chez tous les organismes vivants, essentiellement des protéines impliquées dans la traduction. Elle montre pour la première fois que les procaryotes sont en fait divisés en deux groupes, les Archaea et les Bactéries, et, bien que structuralement les Archaea et les Bactéries semblent proches, elle place les Eucaryotes en groupe-frère des Archaea (Figure 4).

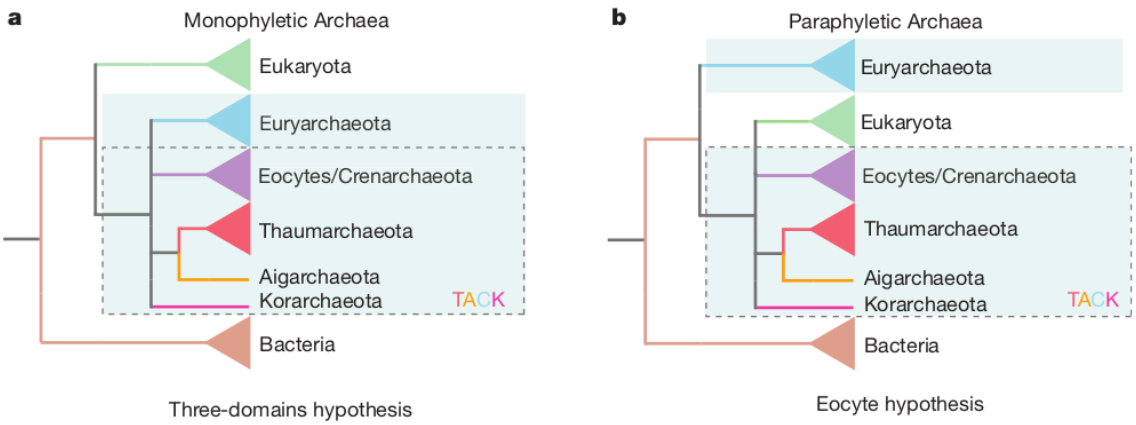


Figure 4. Deux topologies pour l'origine de la cellule eucaryote (source : [53]**)**

Une topologie alternative à celle des trois domaines fut proposée par James Lake en 1983 en se basant sur la structure des ribosomes [55]. Elle place les eucaryotes non plus en groupe-frère des Archaea mais l'intérieur de ce domaine, proche des éocytes qui correspondent à l'actuel groupe des Crenarchaeota (Figure 4) [53]. L'hypothèse Eocyte de James Lake fut longtemps occultée par l'hypothèse des trois domaines de Carl Woese. Cependant, l'amélioration ces dix dernières années des méthodes de reconstruction

phylogénétique couplée à un meilleur échantillonnage des Archaea proches des Crenarchaeota, notamment grâce au séquençage d'organismes non cultivables, soutient l'hypothèse de James Lake comme la topologie la plus probable [53]. En accord avec les reconstructions phylogénétiques basées sur les gènes codant pour des fonctions informationnelles (c'est à dire des gènes impliqués dans la réplication, la transcription et la traduction), de nombreux membres du groupe des TACK (*Thaumarchaeota, Aigarchaeota, Crenarchaeota* et *Korarchaeota*) possèdent des gènes homologues à des gènes que l'on pensait spécifiques des eucaryotes comme l'actine, la tubuline ou des gènes appartenant au système ubiquitine [56]. Cependant, la distribution clairsemée de ces signatures eucaryotes ne permettait pas de préciser davantage le groupe le plus proche des eucaryotes [56]. Récemment, le séquençage d'échantillons de sédiments de hauts fonds marins a permis la découverte d'un nouveau phylum Archaea appelé *Lokiarchaeota* [57]. La reconstruction phylogénétique place les eucaryotes avec le groupe des *Lokiarchaeota* et groupe-frère du groupe des TACK suggérant ainsi que *Lokiarchaeota* est le procaryote le plus proche des eucaryotes découvert à ce jour (Figure 5).

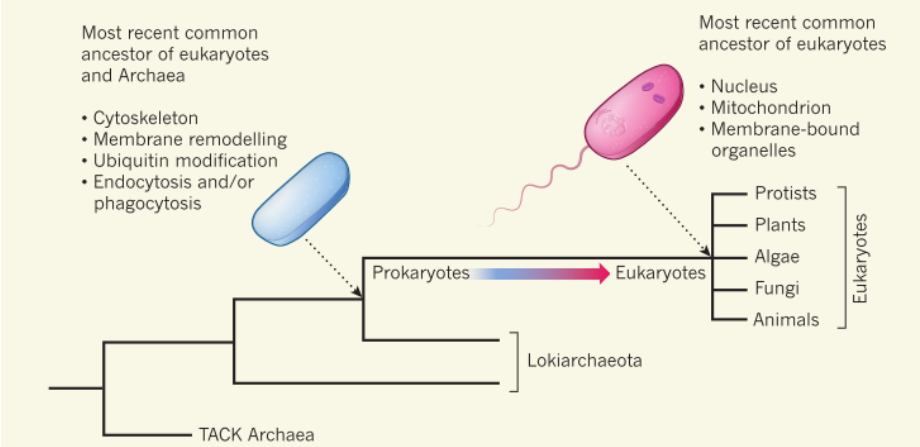


Figure 5. L'origine "Archaea" des eucaryotes (source : [58]**)**

Comme pour les génomes du groupe TACK, le génome de Lokiarchaeum possède de nombreux gènes préalablement considérés comme eucaryotes. Les résultats des dernières années soutiennent tous le fait que les organismes cellulaires sont composés de deux domaines primaires, les Bactéries et les Archaea, et d'un domaine secondaire, les eucaryotes, plus récent et ayant émergé à partir des Archaea et des Bactéries [59].

(2) Hypothèse bioénergétique

Il existe plusieurs scénarios en faveur d'une acquisition précoce de la mitochondrie [47]. Le but de cette introduction n'étant pas de les énumérer, je me propose plutôt de décrire l'hypothèse qui, à mon sens, est la plus forte. Celle-ci s'appuie sur l'apport énergétique de la mitochondrie [60]. Comme expliqué précédemment, les organismes complexes sont tous des eucaryotes. Les procaryotes, bien que possédant une énorme diversité métabolique, n'ont jamais évolué une organisation cellulaire aussi complexe que celle observée chez les eucaryotes en 4 milliards d'années d'existence. Pourquoi ? Lane et Martin proposent que la cause de cette complexité réside dans le gain énergétique apporté par la mitochondrie : la mitochondrie aurait permis une augmentation drastique de l'énergie disponible par gène [60]. Alors que la taille du protéome des procaryotes est contrainte énergétiquement, les cellules eucaryotes disposent de plus d'énergie et peuvent donc produire beaucoup plus de protéines (de l'ordre de 200 000 fois plus). Par coût énergétique par gène, les auteurs veulent parler du coût énergétique nécessaire pour l'expression du gène en protéine et non de sa réplication. Répliquer un gène ne coûte par cher pour la cellule ce qui n'est pas le cas de son expression. En effet, la majeure partie de l'énergie d'un organisme est allouée à la synthèse des protéines. Chez *Escherichia coli,* elle occupe 75% du budget énergétique de la cellule et les ribosomes sont de loin les protéines les plus nombreuses dans le cytoplasme [60]. Lane et Martin estiment qu'une cellule eucaryote peut allouer 5000 fois plus d'énergie par gène qu'une cellule procaryote, lui permettant ainsi d'augmenter le nombre de protéines dans la cellule.

Récemment l'hypothèse bioénergétique a été remise en cause par la découverte de gènes codant pour des protéines impliquées dans le remodelage membranaire (protéines du complexe ESCRT pour « endosomal sorting complexes required for transport ») et le cytosquelette (gelsoline et actine) chez Lokiarchaea. L'existence de tels gènes suggère que Lokiarchaea pourrait être capable de faire de la phagocytose [45,57]. Michael Lynch quant à lui explique que la mitochondrie n'est pas nécessaire à la complexité génomique des eucaryotes; pour lui, c'est la petite taille des populations qui en est responsable [61–65]. Plus important encore, le résultat d'une étude phylogénomique montre que les gènes hérités de la mitochondrie sont les gènes qui possèdent les plus courtes distances phylogénétiques avec les eucaryotes et par extension semblent être donc plus récents que les autres gènes procaryotiques [66]. Ces découvertes sont récentes et demandent de la maturation pour falsifier l'hypothèse énergétique qui reste l'explication la plus consiliente sur le rôle de la mitochondrie dans l'émergence des eucaryotes.

3. La nature hybride des eucaryotes

Bien que les modalités de la fusion restent très débattues [67], il y a un consensus pour dire que les eucaryotes sont le produit d'une symbiose entre deux procaryotes, une Archaea et une alpha-protéobactérie qui deviendra la mitochondrie. Les cellules eucaryotes ont une nature hybride, elles possèdent deux génomes (un mitochondrial et un nucléaire), voire trois dans le cas des eucaryotes photosynthétiques (génome chloroplastique).

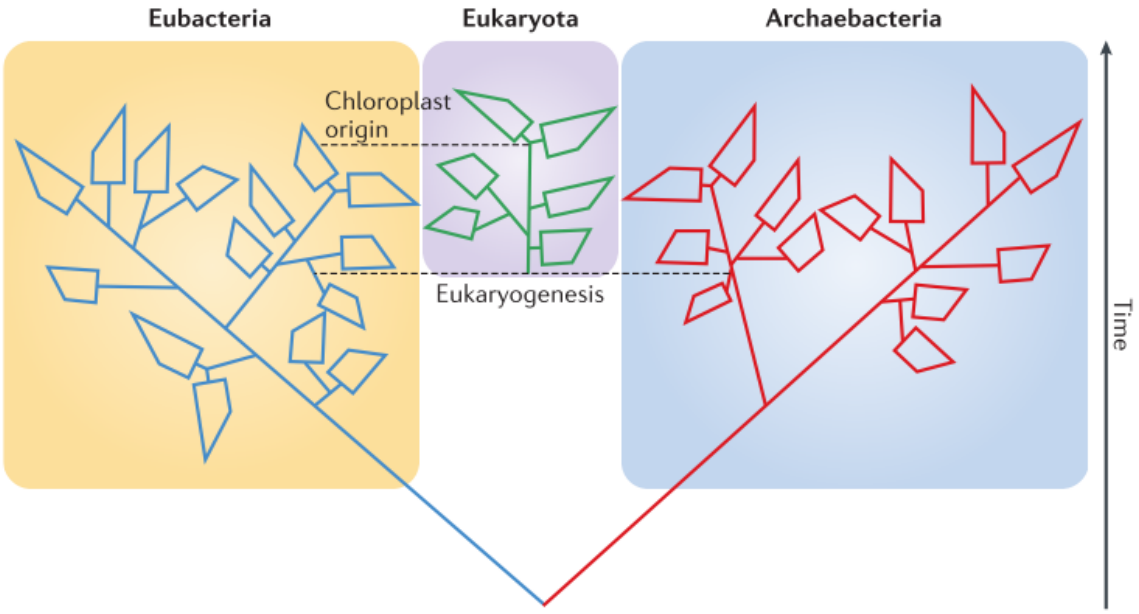


Figure 6. La nature hybride des eucaryotes (source : [59]**)**

Le génome mitochondrial de la majorité des organismes contemporains ne possède qu'une soixantaine de gènes ce qui est beaucoup moins que les alpha-protéobactéries libres actuelles, qui possèdent souvent plus de 4000 gènes. Ce nombre limité chez les mitochondries s'explique par le processus d'intégration de la mitochondrie dans l'hôte car de nombreux gènes mitochondriaux ont été perdus mais pas tous. Une large partie a été transférée du génome de l'endosymbionte vers le génome nucléaire de l'hôte par transferts de gènes endosymbiotiques (EGT), un cas particulier de transferts de gènes horizontaux [68]. La conséquence de ces transferts massifs de gènes bactériens est que le mosaicisme se retrouve au sein du génome nucléaire des eucaryotes. Les génomes eucaryotes sont composés de gènes appartenant aux différents partenaires symbiotiques et l'origine de ces gènes détermine leurs fonctions dans la cellule eucaryote. Dans un papier paru dans le journal PNAS en 1998, Rivera et Lake ont montré que les gènes hérités de l'ancêtre Archaea sont statistiquement enrichis en fonctions dites informationnelles alors que les gènes d'origine bactérienne, plus

nombreux, sont généralement liés à des fonctions dites opérationnelles telles que le métabolisme ou la signalisation cellulaire [69]. Une analyse récente basée sur la distribution de ces gènes chez les eucaryotes confirme que la grande majorité des gènes qui possèdent un homologue procaryote a été acquise lors des deux évènements d'endosymbiose chez les eucaryotes [70].

La nature hybride des eucaryotes pose une question épistémique intéressante : pourquoi les eucaryotes sont vus comme émergents d'une Archaea et pas d'une alpha-protéobactérie ? La contribution du partenaire bactérien n'a rien à envier au partenaire Archaea, surtout si l'acquisition de la mitochondrie est ce qui rend la cellule eucaryote si spéciale comme l'affirme Martin et Lane [60,71]. Pourtant la majorité des articles scientifiques mettent en avant la contribution de l'Archaea plus que la contribution bactérienne. Une réponse suggérée par Eugene Koonin [45] et William Martin [72] est que les phylogénies se font sur les gènes conservés informationnels qui sont d'origine Archaea or l'histoire verticale des gènes informationnels ne concerne que quelques dizaines de familles de gènes [45] et n'est en aucun cas suffisant pour décrire l'histoire des organismes vivants. L'histoire des eucaryotes en est un excellent exemple. Une autre réponse vient de James McInerney qui remet en cause la notion d'hôte et d'endosymbionte qui donne plus d'importance à l'hôte [59].

C. L'acquisition de la photosynthèse chez les eucaryotes

Comme pour la mitochondrie, l'évolution des chloroplastes est une singularité et la distribution phylogénétique des organismes photosynthétiques, entremêlés avec des organismes non photosynthétiques, montre à quel point cette histoire est complexe (Figure 7). Cependant, et bien que de nombreux points restent à éclaircir, les nombreuses découvertes faites ces dernières années permettent de la démêler en grande partie [73].

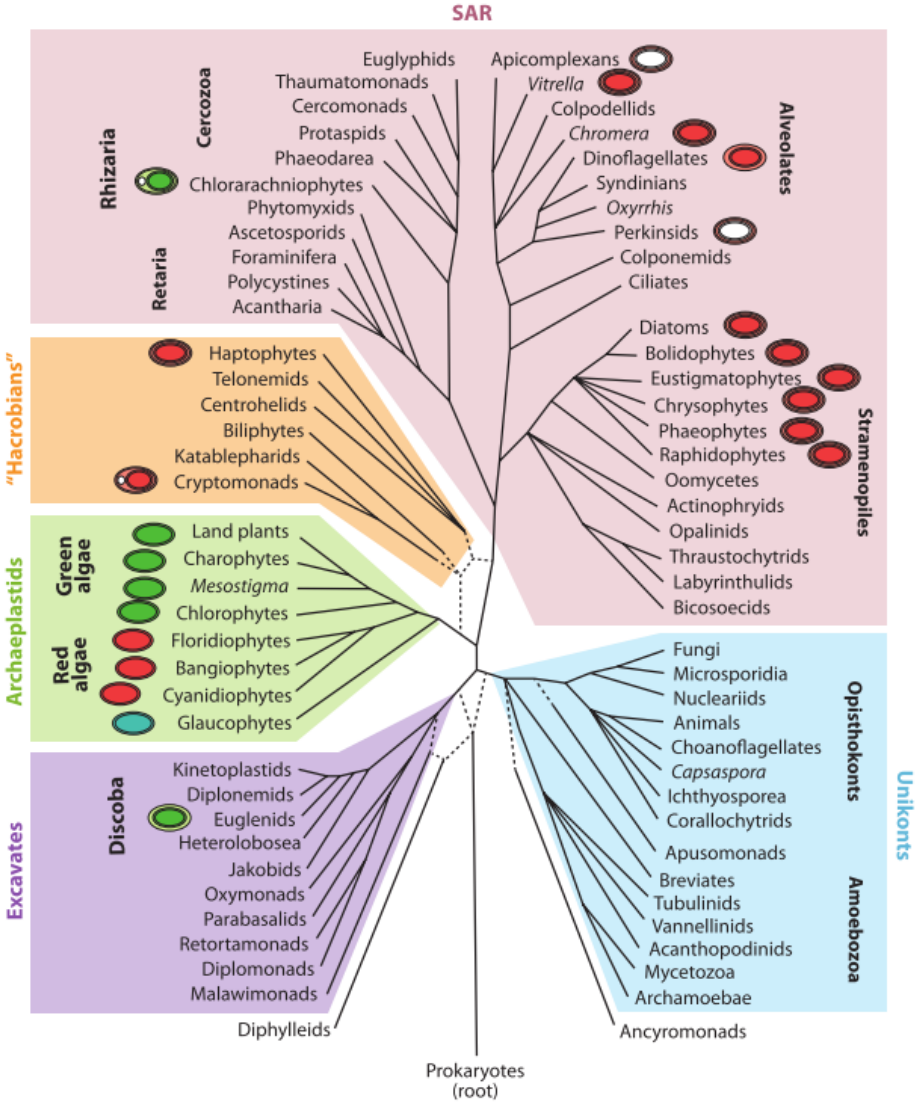


Figure 7. Distribution des lignées photosynthétiques chez les eucaryotes (source : [74]**)**

1. Endosymbiose primaire

Le chloroplaste a été acquis une et une seule fois il y a approximativement un milliard d'année chez l'ancêtre commun aux archaeplastida, un groupe comprenant les glaucophytes, les algues rouges, les algues vertes ainsi que les plantes terrestres. La monophylie des archaeplastida et de leurs chloroplastes [75] ainsi que celle du système de ré-adressage vers l'organite des protéines synthétisées dans le cytoplasme le confirme [76], et tout laisse à penser que l'hôte de la cyanobactérie était une cellule eucaryote hétérotrophe pourvue d'une mitochondrie [44]. L'intégration de l'endosymbionte photosynthétique a suivi les mêmes étapes que pour la mitochondrie : son génome a été fortement réduit, une large partie de ses gènes a été perdue alors qu'une autre a été transférée vers le génome nucléaire

par transferts de gènes endosymbiotiques, exacerbant encore un peu plus le mosaïcisme génomique des eucaryotes [73]. Plusieurs études ont tenté d'estimer l'impact génomique de la cyanobactérie sur le génome de l'hôte. En 2002, une étude de Martin a estimé que 18% des gènes nucléaires d'*Arabidopsis thaliana* sont d'origine cyanobactérienne [77]. D'autres études faites chez des glaucophytes ou des algues rouges ont produits des estimations moindres, de l'ordre de 10% [78,79]. Sans tenir compte de ces différences, l'endosymbionte a significativement remodelé le génome de l'hôte en apportant de nombreux gènes dont beaucoup sont impliqués dans la photosynthèse [80].

2. Endosymbioses secondaires

Une singularité de l'acquisition de la photosynthèse chez les eucaryotes est qu'à la différence de l'évolution de la mitochondrie, l'évolution de la photosynthèse implique de nombreux événements d'endosymbioses secondaires ayant permis la diffusion horizontale de la photosynthèse entre lignées eucaryotes. A la différence de l'endosymbiose primaire où une cyanobactérie est entrée en endosymbiose avec un eucaryote non photosynthétique, l'endosymbiose secondaire implique la phagocytose d'un eucaryote photosynthétique par un eucaryote non photosynthétique (Figure 8) [74]. Des endosymbioses tertiaires, c'est à dire la phagocytose d'un eucaryote secondaire par un eucaryote non photosynthétique, ont même été observées chez certaines lignées de dinoflagellés.

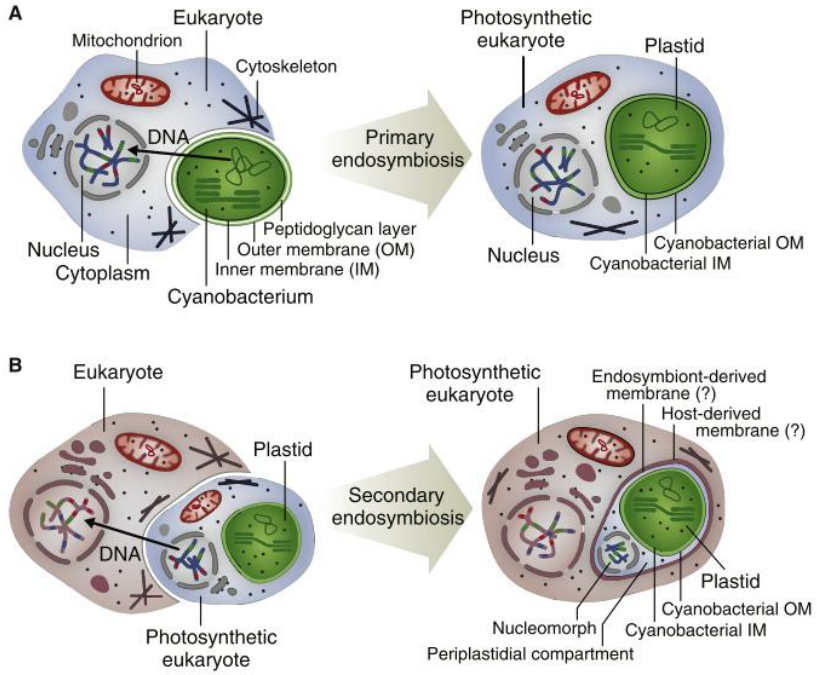


Figure 8. Endosymbiose primaire et secondaire (source : [44]**)**

Certains organismes conservent la trace de ces évènements comme le cryptophyte *Guillardia theta* et le chlorarachniophyte *Bigelowiella natans,* qui possèdent encore la trace du noyau, appelé nucléomorphe, de l'algue rouge et verte qu'ils ont phagocytée [81] (Figure 9).

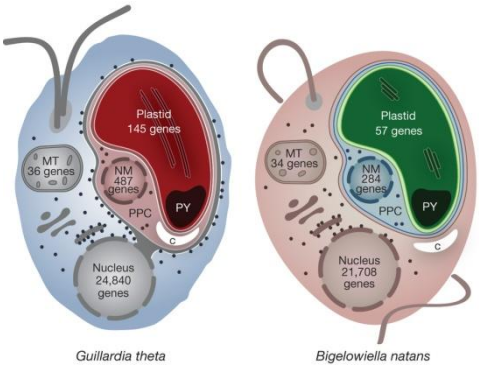


Figure 9. Biologie cellulaire de *Guillardia theta* et de *Bigelowiella natans* (source : [81]**)**

Au moins trois événements d'endosymbioses secondaires ont été recensés, un aboutissant à la lignée des euglénoïdes, un chez les chlorarachniophytes et au moins un impliquant l'acquisition d'un plaste d'algue rouge chez les Cryptophytes, les Alvéolés, les Straménopiles et les Haptophytes (groupe CASH) (Figure 10). Alors qu'il est maintenant bien accepté que les euglénoïdes et les chlorarachniophytes, qui possèdent un plaste provenant d'algue verte, ont deux origines indépendantes (les reconstruction phylogénétiques des plastes le confirment [82] ainsi que l'étude de leurs systèmes de ré-adressage [83]), le nombre et la nature des endosymbioses impliquant des algues rouges et le groupe CASH restent intensément débattus.

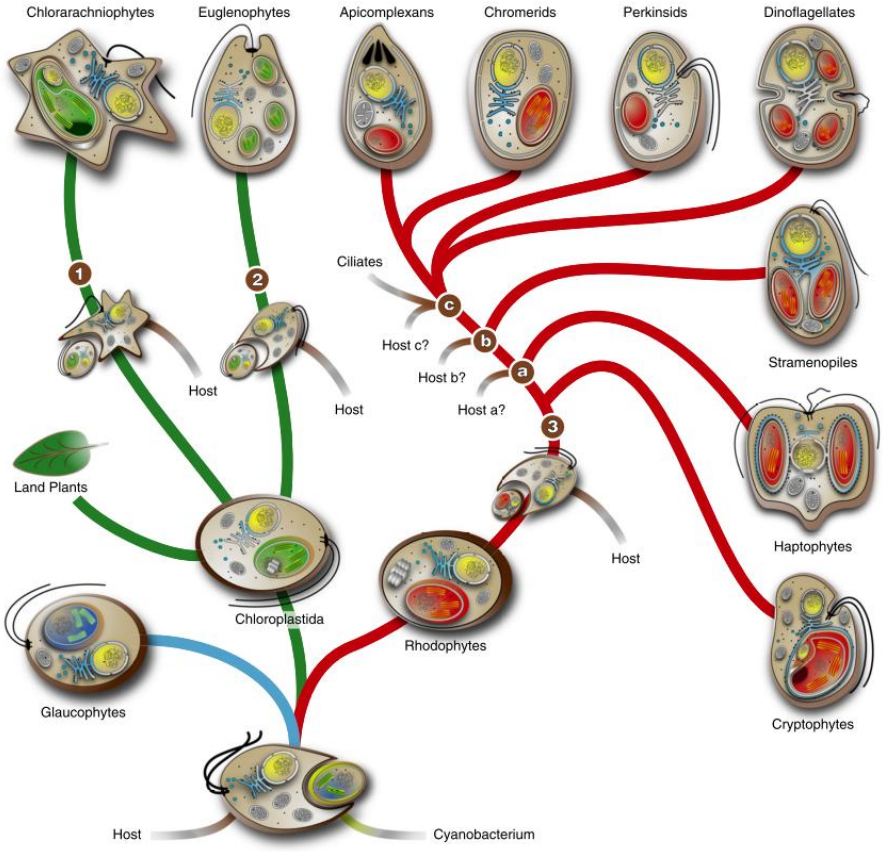


Figure 10. Histoire des endosymbioses (source : [84]**)**

L'histoire des lignées du groupe CASH est beaucoup plus complexe, notamment parce que beaucoup plus de lignées possèdent un plaste rouge acquis secondairement et que les phylogénies des espèces et des plastes sont incongruentes. Alors que l'analyse phylogénétique des génomes chloroplastiques soutient la monophylie du groupe CASH [85,86], l'analyse phylogénétique des espèces ne soutient pas la monophylie du groupe CASH [87]. La monophylie des alvéolés et des straménopiles est bien établie; en revanche les cryptophytes et des haptophytes, que l'on pensat proches, semblent être polyphylétiques, les cryptophytes branchant finalement avec les archaeplastida et les haptophytes branchant avec le groupe SAR (Straménopile, Alvéolés et Rhizaria) [87,88].

Un moyen complémentaire pour reconstruire l'histoire des événements d'endosymbiose est de s'intéresser au système d'importation des protéines cytosoliques vers le chloroplaste. Du fait des EGT, la majeure partie des protéines travaillant dans le chloroplaste se trouvent codées dans le génome nucléaire et sont donc synthétisées dans le cytoplasme avant d'être importées dans le chloroplaste à l'aide du complexe protéique TIC et TOC (pour Translocon of the Inner and Outer enveloppe of Chloroplast) [89,90]. Le système TIC et TOC est un complexe membranaire spécifique des eucaryotes photosynthétiques contenant de

nombreuses protéines dont beaucoup ne sont présentes que chez les eucaryotes. Il parait donc très peu probable que ce complexe protéique ait évolué indépendamment plusieurs fois. Par conséquent, l'étudier est un bon moyen d'élucider l'histoire des événements d'endosymbioses [47,91] et il a d'ailleurs été utilisé pour confirmer que l'ensemble des plastes ont une seule et même origine [76]. Un système analogue appelé TIM et TOM (pour Translocon of the Inner and Outer of Mitochondrial membrane) existe pour la mitochondrie et a aussi été utilisé pour confirmer l'origine unique des mitochondries [47,92].

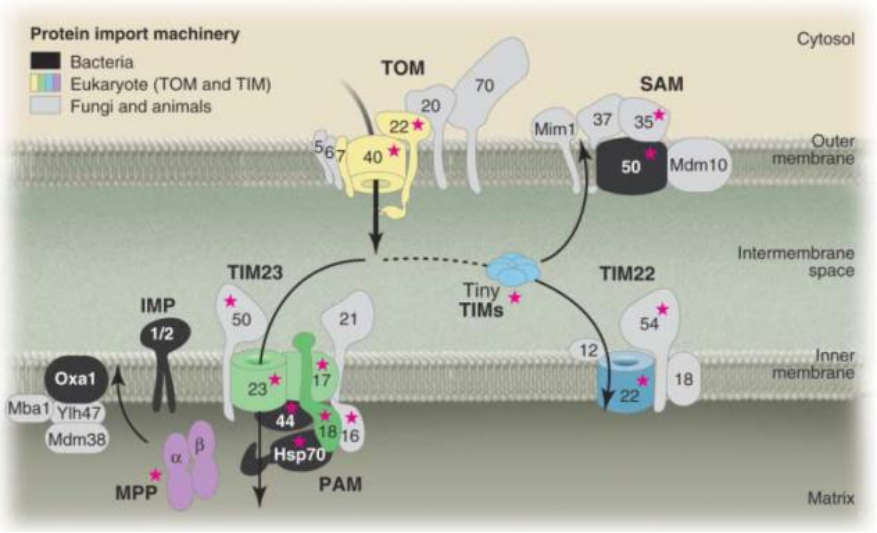


Figure 11. Le système TIM/TOM (source : [92]**)**

A la différence des plastes primaires qui contiennent deux membranes, les plastes du groupe CASH possèdent, à l'exception des dinoflagellés, deux membranes externes supplémentaires. La présence de ces deux nouvelles membranes pose problème pour l'importation des protéines car cela nécessite que les protéines synthétisées dans le cytoplasme passent deux nouvelles membranes avant d'atteindre le système TIC et TOC. La membrane la plus externe correspond au prolongement du reticulum endoplasmique de l'hôte [81] et son passage se fait grâce au complexe Sec61 de l'hôte. Le cas de la seconde membrane la plus externe est plus compliqué et a nécessité l'évolution d'une nouvelle machinerie pour permettre l'acheminement des protéines vers le système TIC et TOC. Les travaux de l'équipe de Uwe-G. Maier ont permis d'identifier le système d'importation à travers la seconde membrane externe [93]. Ce complexe protéique est appelé SELMA (pour Symbiont-specific ERAD-like Machinery) et dérive du système ERAD (pour Endoplasmic Reticulum Associated Degradation) qui permet d'exporter les protéines du réticulum endoplasmique vers le cytosol pour leur dégradation chez les eucaryotes. A l'exception des dinoflagellés qui possèdent trois membranes et ont sans doute perdu celle contenant le système SELMA,

l’ensemble des lignées du groupe CASH possèdent le système SELMA et il semble avoir lui aussi une seule origine comme TIC/TOC et TIM/TOM [91]. Ces résultats, en plus de la monophylie des plastes d’algues rouges, semblent indiquer que les lignées du groupe CASH ont hérité du même plaste. Cependant ils n’indiquent pas comment ces plastes ont été transmis vers les différentes lignées polyphylétiques. Une étude a proposé des endosymbioses en série (tertiaire et quaternaire) pour expliquer la distribution de ces lignées [94] mais cette hypothèse pose de nombreux problèmes notamment concernant le nombre de membranes observées [91].

D. Un gain métabolique pour expliquer les transitions égalitaires ? (Article 1)

La nature chimérique des eucaryotes nous montre que de nouvelles lignées peuvent évoluer *via* des processus introgressifs [3]. L'évolution n'est pas qu'un processus graduel au sens darwinien du terme, des événements non graduels et non verticaux ont joué des rôles essentiels dans certaines transitions évolutives majeures [40,41]. La fusion de deux organismes possédant des origines phylogénétiques distinctes peut aboutir à la création d'un nouvel organisme aux propriétés émergentes. Ces transitions sont dites égalitaires [40,41]. L'émergence des eucaryotes et celle des eucaryotes photosynthétiques sont clairement des transitions évolutives majeures impliquant des processus introgressifs. Expliquer ce qui conduit à ces transitions égalitaires est important mais extrêmement compliqué du fait de l'ancienneté et de la rareté de ces événements [95]. Récemment, deux études ont suggéré que tous les grands clades d'Archaea ont acquis massivement des gènes bactériens et, bien que cela soit débattu [96,97], ces acquisitions semblent directement liées à l'origine de ces clades [12,98]. Si l'on ajoute à cette découverte d'autres transitions égalitaires plus récentes et donc ayant des impacts moins importants en évolution [24,29,99], une règle émerge : l'acquisition de fonctions métaboliques semble être une cause importante de ces transitions égalitaires.

Spotlight

CellPress

Metabolic bacterial genes and the construction of high-level composite lineages of life

Raphaël Méheust, Philippe Lopez, and Eric Bapteste

UMR7138 Evolution Paris-Seine; Institut de Biologie Paris-Seine; Université Pierre et Marie Curie; 9 quai saint Bernard, 75005, Paris, France

Understanding how major organismal lineages originated is fundamental for understanding processes by which life evolved. Major evolutionary transitions, like eukaryogenesis, merging genetic material from distantly related organisms, are rare events, hence difficult ones to explain causally. If most archaeal lineages emerged after massive acquisitions of bacterial genes, a rule however arises: metabolic bacterial genes contributed to all major evolutionary transitions.

Making sense of the origins of major lineages of life, and therefore of the ways by which novel physiologies, ecological systems, and classes of organisms evolved is possibly as important as understanding the origin of species [1]. It could provide fundamental insights about the biology and the intimate make-up of many organisms investigated by microbiologists, ecologists, developmental biologists, geneticists and evolutionary biologists. Recent findings [2–4] proposed that unsuspected major evolutionary transitions occurred amongst prokaryotes, because most archaeal lineages emerged as the result of massive acquisitions of bacterial genes. Therefore, major archaeal lineages would be in part composed of bacterial genes. While the mere discovery of composite archaeal lineages is already thought-provoking, it takes an even greater significance when considered in a broader biological context (Figure 1). This latter comparison unravels a remarkable trend: all major evolutionary transitions leading to novel composite high-level lineages might have benefited from the merging of genetic material from bacteria with genetic material from other sources.

The contribution of bacterial genes to eukaryotic evolution is well-acknowledged [4–6]. Eukaryotes appear as genetic chimera, largely composed of metabolic genes from bacterial origin, while another part of their genomes likely originates from archaea. Eukaryogenesis would have indeed involved (at least) these two kinds of partners: an ancestral bacterial lineage and an ancestral archaeal lineage [4–6]. This dual origin of eukaryotes is without doubt one of the main evolutionary events that occurred on the planet. It is striking because it shows that lineages can evolve by introgressive events [7], and not just via divergence from a last common ancestor. As eukaryotes emerged from the merging of these two distantly related components, novel creatures took a central stage in the evolutionary history benefiting from novel biological properties, such as, typically, a new mode of generation of genetic variability, i.e. the general fuel for evolution, by meiosis. Such an event, which changed the course of life on Earth, has been described as an egalitarian evolutionary transition [7], because it involves the association and stabilization of elemental components with different phylogenetic origins and their transformation into a novel composite life form. Because mitochondria are considered remnants of this ancient evolutionary transition, bacterial genes are generally thought to have endowed eukaryotes with their metabolic capabilities [4–6].

A major evolutionary transition such as eukaryogenesis, giving birth to a high-level lineage, by the merging of genetic material from distantly related organisms, is usually assumed to be rare (no more than a few events per billion years). Being rare, however, does not entail that such events do not obey rules, but simply that causal rules are difficult to discover. In a recent series of original works, Nelson-Sathi *et al*. [2,3] further showed that, possibly, most major archaeal lineages likewise emerged from the merging of metabolic genes from bacteria with the genetic material of methanogenic archaea. Again, metabolic genes appeared as key elements for the birth of these novel lineages via introgressive processes, producing novel successful lines of composite beings on Earth. These findings in the archaeal domain suggest that the evolution of eukaryotes was not just one random chance event, but the outcome of a recurrent process in which metabolic genes from bacterial lineages provide genetic bases for the make-up of novel life forms. In other words, in many occasions bacterial metabolic genes were subjected to introgressive processes, and only in a limited number of occasions, these introgressions resulted in the emergence of lineages of novel beings better fitting with the adaptive demands. Therefore, the contribution of metabolic genes from bacterial lineages would be one of the rules of egalitarian evolutionary transitions. By contrast, so far, genes from archaeal [2] and from eukaryotic lineages do not seem to have contributed to the emergence of novel bacterial lineages. A major lesson from works on the origins of composite high-level phyla might be that metabolic bacterial genes are amongst the greatest and most creative evolutionary plugins of life. Acquisition of such metabolic genes likely opened or defined new niches for composite lineages, in which environmental adaptation and further habitat preference took place.

Corresponding author: Bapteste, E. (eric.bapteste@snv.jussieu.fr).
Keywords: evolutionary transition; prokaryotic evolution; tree of life; web of life; lateral gene transfer; eukaryogenesis.

0169-5347/
 http://dx.doi.org/10.1016/j.tree.2015.01.001

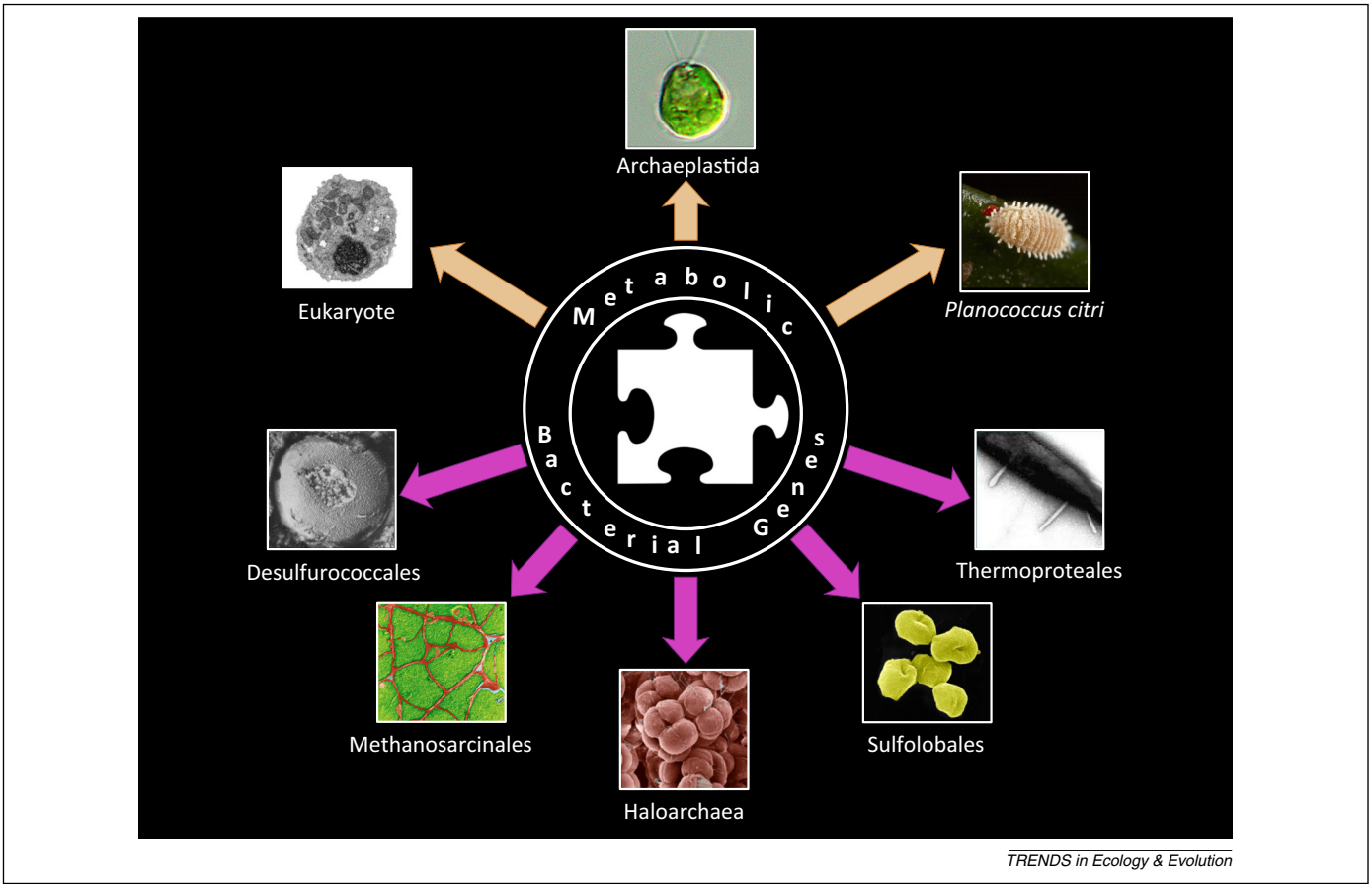


Figure 1. Introgression of metabolic bacterial genes: a recurrent evolutionary theme at the origin of novel composite lineages. First reports of bacterial genes contributions to the evolution of lineages were documented in eukaryotes (orange arrows), with the discoveries of eukaryogenesis, the primary chloroplastic endosymbiosis at the origins of Archaeaplastida, or of more recent endosymbioses endowing several eukaryotic lineages with additional metabolic capabilities, exemplified here by the tripartite nested mealybug symbiosis. Nelson-Sathi *et al.* [1] profoundly expanded this view, as they propose that numerous major archaeal lineages (pink arrows) also originated from the massive acquisition of bacterial genes.

This unparalleled role of metabolic bacterial genes in shaping lineages has also been observed in several endosymbiotic events, which turned non-photosynthetic eukaryotes into photosynthetic ones [8], as well as in other fascinating introgressions [9]. Such a role is directly consistent with the frequent exchanges of metabolic genes between bacterial lineages themselves [10]. Therefore, not all genes on Earth have the same evolutionary fate and impact. Within the cellular world, metabolic bacterial genes, characterized by their amazing evolvability and their repeated crucial contribution to the make-up of lineages across all life, seem to be, by far, a most valuable source of adaptations. Why bacterial genes, rather than archaeal or eukaryotic genes, constitute such a widespread powerful evolutionary material deserves further investigation. One can only speculate. Careful analyses of genes flux, for example analyzing the turn-over of genes belonging to distinct functional categories in genomes, might show that metabolic bacterial genes persist for longer time periods than most other genes in genomes. Such a higher persistence, if observed, might explain why contributions of metabolic bacterial genes to the long-time evolution of composite lineages have been repeatedly detected. Importantly, works by Nelson-Sathi *et al.* encourages strengthening the research program on evolutionary transitions, by specifically tracking the motion and the transformative role of metabolic genes from bacterial origin in the web of life. It could provide novel and broader angles to address the issue of the origins of lineages, and the diversity of life on Earth.

Acknowledgments

EB is funded by FP7/2007-2013 Grant Agreement # 615274, and thanks members of the ANR-13-BSH3-0007 ('The space of explanations in evolutionary biology') project for stimulating discussions.

References

1 Szathmáry, E. and Smith, J.M. (1995) The major evolutionary transitions. *Nature* 374, 227–232
2 Nelson-Sathi, S. *et al.* (2015) Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* 517, 77–80
3 Nelson-Sathi, S. *et al.* (2012) Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proc. Natl. Acad. Sci. U.S.A.* 109, 20537–20542
4 Williams, T.A. *et al.* (2013) An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* 504, 231–236
5 McInerney, J.O. *et al.* (2014) The hybrid nature of the Eukaryota and a consilient view of life on Earth. *Nat. Rev. Microbiol.* 12, 449–455

6 Rivera, M.C. and Lake, J.A. (2004) The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* 431, 152–155
7 Bapteste, E. *et al.* (2012) Evolutionary analyses of non-genealogical bonds produced by introgressive descent. *Proc. Natl. Acad. Sci. U.S.A.* 109, 18266–18272
8 Gould, S.B. *et al.* (2008) Plastid evolution. *Annu. Rev. Plant Biol.* 59, 491–517
9 Husnik, F. *et al.* (2013) Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell* 153, 1567–1578
10 Chai, J. *et al.* (2014) Functional phylogenomics analysis of bacteria and archaea using consistent genome annotation with UniFam. *BMC Evol. Biol.* 14, 207

III. Introgression au niveau génomique et innovations (Articles II et III)

Les phénomènes introgressifs ne transforment pas seulement les objets biologiques aux niveaux des cellules et des organismes, ils réorganisent aussi les génomes et les gènes. Les gènes qui composent les génomes ne sont pas fixés, les génomes sont dynamiques et leurs contenus se modifient au cours du temps et des générations. Des gènes peuvent être perdus [100–102], le génome peut aussi acquérir de nouveaux gènes par de nombreux moyens [103].

L'article suivant [5] fait un état de l'art des méthodes de réseaux utilisées pour étudier l'évolution réticulée au niveau génomique. L'article propose une formalisation des réseaux en évolution en quatre types :

- les réseaux de similarités de séquences
- les réseaux de génomes
- les réseaux multiplexés
- les réseaux bipartis

Le second article [104] s'est fait en collaboration avec François-Joseph Lapointe de l'université de Montréal. Il correspond au développement d'une méthode étudiant le comportement d'un réseau après ajout de nouveaux nœuds et de nouvelles arêtes.

Trends in Microbiology

CellPress

Review

Network-Thinking: Graphs to Analyze Microbial Complexity and Evolution

Eduardo Corel,[1,*] Philippe Lopez,[1] Raphaël Méheust,[1] and Eric Bapteste[1]

The tree model and tree-based methods have played a major, fruitful role in evolutionary studies. However, with the increasing realization of the quantitative and qualitative importance of reticulate evolutionary processes, affecting all levels of biological organization, complementary network-based models and methods are now flourishing, inviting evolutionary biology to experience a network-thinking era. We show how relatively recent comers in this field of study, that is, sequence-similarity networks, genome networks, and gene families–genomes bipartite graphs, already allow for a significantly enhanced usage of molecular datasets in comparative studies. Analyses of these networks provide tools for tackling a multitude of complex phenomena, including the evolution of gene transfer, composite genes and genomes, evolutionary transitions, and holobionts.

Trends

Introgressive processes shape the microbial world at all levels of organisation.

This reticulated evolution is increasingly studied by sequence-similarity networks.

They provide an inclusive accurate multilevel framework to study the web of life.

Networks enhance analyses of microbial genes, genomes, communities, and of symbiosis.

New Methods for Studying the Web of Life

The tree model has been largely and rightly used in evolutionary analyses since Darwin's seminal work [1]. The genealogical relationships between evolving objects are indeed critical to explain life's diversity, not only from a processual perspective (where common ancestry explains some similarities), but also as a powerful pattern to classify all related evolved forms [2]. However, the tree structure, especially when assumed to be universal, strongly constrains our description of the evolution of life [3–5]. By definition, a tree can only describe divergence from a last common ancestor (often with dichotomies, or with polytomies describing fast radiations). In vertical descent, the genetic material of a particular evolutionary unit is propagated by replication inside its own lineage. When such lineages split, and become genetically isolated from one another, this produces a tree. By contrast, in introgressive descent, the genetic material of a particular evolutionary unit propagates into different host structures and is replicated within these host structures [4]. However, a tree with a single ancestor for each object cannot represent such a merging of distinct lineages into a novel common host structure. Typically, organisms produced by sexual reproduction in eukaryotes originate from two parents which merged their genetic material. Genealogical trees with a single ancestor do not describe relationships within eukaryotic sexual populations. Indeed, this genuine genealogical relationship cannot be depicted with a traditional tree representation since this pattern would impose that one considers an offspring either more closely related to only one of its parents, or to be the progenitor of its own ascendants [4].

The distinction between vertical and introgressive descent is not a minor one; **introgression** (see Glossary) affects all levels of biological organization: from molecules, when sequences legitimately or illegitimately recombine, to genomes, when sequences enter genomes by lateral

[1]Equipe AIRE, UMR 7138, Laboratoire Evolution Paris-Seine, Université Pierre et Marie Curie, 7 quai St Bernard 75005 Paris, France

*Correspondence: eduardo.corel@upmc.fr (E. Corel).

Box 1. Mosaicism of Life

Introgression, the merging of entities from different lineages, affects multiple levels of biological organization. For example, Figure IA describes the introgression of genes from distinct gene families, which results in composite genes, such as multidomain genes [46]. These sequences can come from within a given genome, or when they come from different genomes, such as the genomes of an endosymbiont and of its host, the resulting composite gene, composed partly of material from an endosymbiont, is therefore a symbiogenetic gene. Figure IB describes the introgression of a gene into a host genome, occurring for instance during a lateral or an endosymbiotic gene transfer, which results in a composite genome [63], or when sequences transfer across mobile genetic elements, producing mosaic mobile elements [37]. Of note, more than one gene can be so acquired by a genome [59,60]. Figure IC describes the introgression of a genome into another genome, occurring for instance when the genome of *Wolbachia pipiensis* becomes inserted into the genome of a *Drosophila ananassae*, or when the genome of a virophage such as Sputnik becomes inserted into the genome of a giant virus such as *Mimivirus*, which results in a composite genome [87–89]. Figure ID describes the introgression of a mobile element, such as a plasmid, within a host cell (or organism), occurring for instance when a symbiotic plasmid carrying hypermutagenesis determinants (e.g., the *imuABC* cassettes) invades soil bacteria, enhancing the *ex planta* phenotypic diversification of these novel composite cells [90,91]. Figure IE describes the introgression of a genome into a host cell, occurring for instance during events of kleptoplasty [92,93] or as a result of an extreme reductive evolution after secondary or tertiary plastid acquisition, which results in a (transient or persistent) composite organism [7]. Figure IF describes introgression of cells or organisms, occurring for instance during the evolution and growth of multispecies biofilms [94], endosymbiosis [8,95], during the development and speciation of animals [82,83]. Typically, sequence-similarity networks can be used to investigate for A; genome networks for B, C, and E; multiplex genome networks for B, C, and E; and bipartite networks for B, D, E, and F.

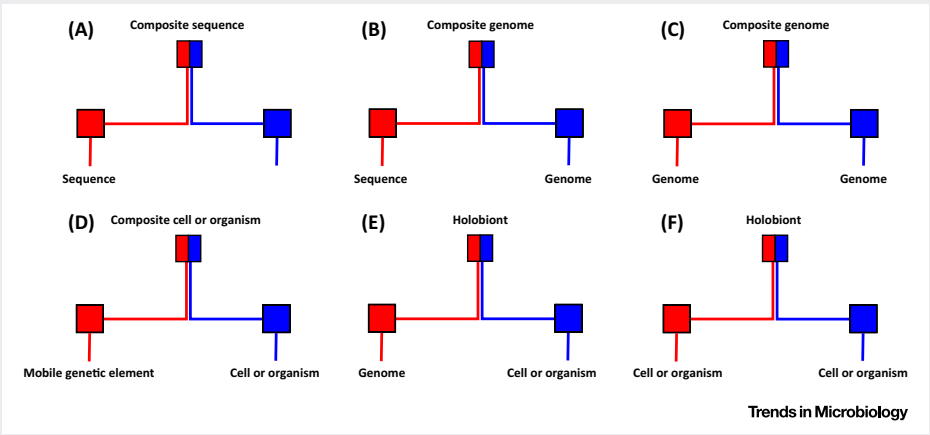


Figure I. Several Illustrations of Mosaicism through Merging Events. (A) Composite genes result from the fusion of different gene domains. (B) Composite genomes can result from the introgression of a gene into a genome, or (C) from the introgression of a genome into a genome. (D) Composite organisms can arise from the introgression of a mobile genetic element. Holobionts result from the introgression of a genome (E) or of another cell (F) into a cell.

gene transfer, and to holobionts, when organisms form a collective system (such as the tight association observed between host and endosymbionts) [6–8] (Box 1). Introgressive descent does not always imply lateral gene transfer: for example, independently replicated gene families, each having their own tree, can merge, and this results in a novel composite gene family. Since even introgressive descent is descent, it encompasses a vertical dimension. The tree representation emphasizes how entities evolve *ex unibus plurum*, whereas the network representation emphasizes how entities evolve *ex pluribus unum*. Of course, evolution progresses in both dimensions. Thus, the tree of life and the network of life are not mutually exclusive models. When lineages that evolved in a tree-like fashion merge, this creates reticulation between branches of trees; likewise, after a reticulation event, phylogenetically composite entities can undergo a tree-like evolution: a tree starts growing on the ground of an initial reticulation. Consequently, future synthetic representations could aim at displaying simultaneously both vertical and lateral parts of biological evolution.

Glossary

Articulation point (or cut-vertex): node in a graph whose removal increases the number of connected components of the resulting graph.
Betweenness: centrality measure for a node in a graph, namely, the proportion of shortest paths between all pairs of nodes that pass through this specific node. Nodes having a betweenness close to 1 are said to be more central, and those close to 0, more peripheral.
Bipartite graph: a graph with two types of node (top nodes and bottom nodes) such that an edge only connects nodes of one type with nodes of the other type.
Club of genomes: a coalition of entities replicating in separate events and exploiting some common genetic material that does not necessarily trace back to a single last common ancestor.
Community: in graph theory, groups of nodes that are more connected between themselves than with the rest of the graph. This technical meaning should not be confused with its use in expressions such as 'microbial communities'.
Connected component: set of nodes in a graph for which there is always an interconnecting path.
Degree: number of incident edges to a given node.
Introgression: descent process through which the genetic material of a particular evolutionary unit propagates into different host structures and is replicated within these host structures.
Multiplex graph: a graph having possibly several edges of different types between two nodes.
Neighbors: nodes that are directly connected by an edge.
Public genetic goods: the common genetic material shared by a club of phylogenetically distant genomes.
Quotient graph: simplified graph whose nodes represent disjoint subsets of nodes of the original graph; an edge in this new graph connects two such new nodes whenever an edge in the original graph connects at least one element of a new node with at least one from the other.
Support: the common set of neighbors of a twin class.
Twins: nodes in a graph that have exactly the same set of neighbors.

Importantly, not all evolving objects entertain genealogical relationships: for instance, viruses and cells, both critical players of biological evolution, are not assumed to be related in this way [9–14], nor are plasmid and plasmid-like transferable objects, as integrative-conjugative elements (ICEs). Cells, viruses, plasmids, and ICEs lack recognized genealogical relationships, either because they genuinely evolved from separate roots, or because their putative common ancestor(s) cannot be inferred from the data, for example, if other descendants of such ancestors became extinct, or have not been sequenced to date. This apparent genealogical disconnection does not exclude vertical evolution within lineages of mobile elements. There is, for example, evidence for both vertical and introgressive descent in plasmids and ICEs of firmicutes [15]. But it means that one genealogical tree cannot represent all the evolutionary history [6]. Therefore, a traditional approach to analyzing evolution incurs the risk of missing explananda (many phenomena that are not described by a genealogical tree) and missing explanans (many evolutionary processes responsible for life's diversity). Trees and networks are representations that allow for scientific analysis. Consistently, tree-thinking has already largely been exploited, and it is now timely and heuristic to turn to network-thinking to illuminate additional and complex aspects of the biology. In this review, we argue that sequence-similarity networks, already used to investigate the evolution of protein coding genes, can also be used to analyze many mosaicisms of life, such as bacterial genome evolution, prokaryotes' and protists' organismal evolution, and the evolution of holobionts and communities in which microbes play a role, in particular as symbionts (Box 1). We explain introgression results in at least three major phenomena: (i) microbial social life, understood here as genetic transfers between different genomes, (ii) chimerism (occasionally implying major evolutionary transitions), and (iii) holobionts. All three examples resist classic tree-based analyses and challenge our evolutionary knowledge. A tree model alone does not describe these introgressive processes, that is, the fact that they involve multiple lineages, and their outcomes, that is, the fact that they produce collective, composite, entities. We describe how and why these phenomena can be studied using three classes of networks [sequence-similarity networks, genome networks, and **bipartite graphs** (Figure 1, Key Figure)], enlarging the analytical toolkit of evolutionary microbiologists. On the one hand, the display of large networks will constitute a challenge for the future development of network-thinking. On the other hand, in terms of interpretation, even very large and dense networks can be effectively simplified, for example, using twin analyses. Thus, we expect a network-thinking era to soon be at the forefront in microbiology.

Investigating Microbial Social Life with Genome Networks

Gene transfer between prokaryotic organisms and mobile genetic elements (i.e., viruses and plasmids) has largely shaped cellular genome content, as illustrated by the observation of prokaryotic pangenomes [15,16], for which the collection of gene families used by the members of a given species is larger than the number of gene families present in any individual genome from that species. The flow of genes between genomes, often mediated by mobile genetic elements, explains this observation, but complicates classic inferences about the past (such as genome reconstruction attempts) [4,5,17–20]. For a given lineage, the contents of ancestral genomes may be largely different from the union of extant genomes because prokaryotic genomes act as 'read–write' storage organelles rather than 'read–only' memories [21], and genomes can lose genes. Thus, describing evolution requires not only the tracking of mutations that accumulate within gene families, or loss of gene families [22], but also genes that are gained by introgression [23]. The latter encourages exploring horizontal gene transfer within prokaryotic communities. This brings forward difficult questions [20,24–30] since there are many routes through which genes pass from one microbial host to the other, that is, multiple channels [31] for gene transmission. For example, is gene transmission random in terms of cellular, viral, or plasmidic targets (however producing asymmetrical results due to some further host selection acting on the incoming genetic material)? Is it random in terms of what gene families are transmitted? Can we find groups of cotransmitted genes?

Key Figure

Different Graph Representations of the Same Gene Sharing among Genomes

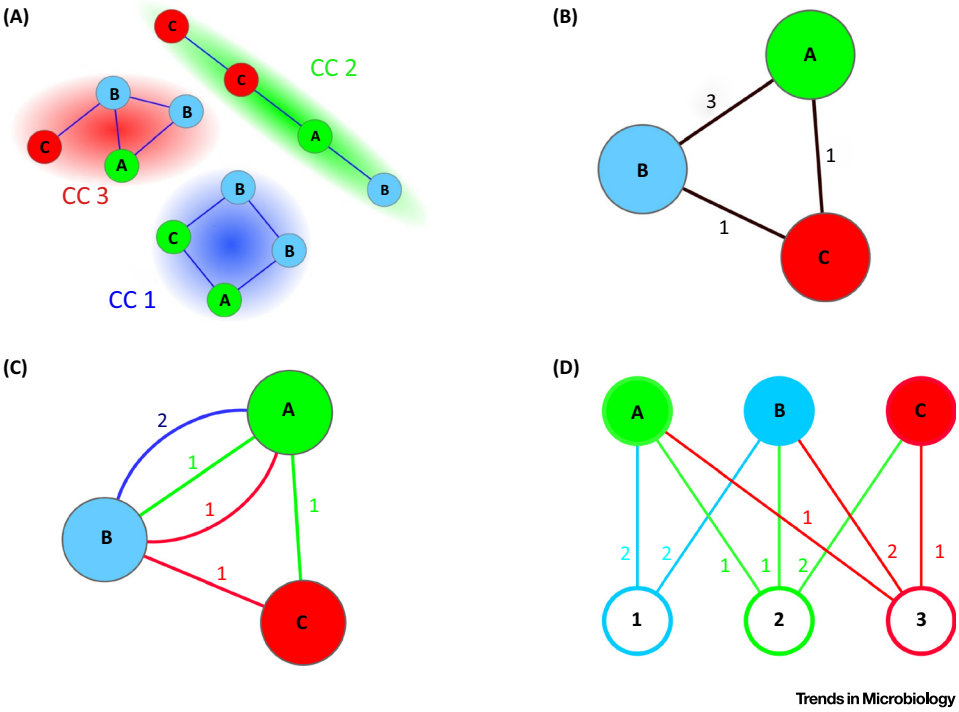


Figure 1. (A) Sequence-similarity network (SSN): each node (circle) represents a protein-coding gene sequence; the color and the label of the node represent the genome where the gene is found. Two nodes are connected by an edge (a line linking two nodes) if the pair of sequences fulfills given similarity criteria such as a minimum percentage identity and coverage (i.e., the ratio between the length of the matching parts and the total length of any two sequences). Sequence-similarity networks are analyzed as a partition into connected components (CCs, highlighted as color halos). This partition defines groups of putative gene families, when reciprocal sequence coverage and identity percentage are high [68]: for instance, we can interpret CC1 as a gene family for which two copies are present both in genomes A and B. (B) Genome networks (GNs) can be obtained from SSNs: nodes are genomes (described by color and label); edges connect genomes that share at least one gene family; GNs can be weighted: weights count the number of gene families shared by the two genomes. In the example, A and B share three gene families, but the graph does not specify which ones. (C) Multiplexed networks (MNs) can be, in turn, obtained from GNs by labelling edges in order to identify what gene families are shared: nodes represent genomes; multi-edges represent distinct shared gene families (same color code as the CCs in the SSN); weights count the number of shared genes in each family: the blue edge between A and B corresponds to CC1 in (A) and has therefore weight 2. (D) Bipartite graphs can also be obtained from SSNs; top nodes are genomes; bottom nodes are gene families; edges connect a genome to a gene family if that genome contains at least one representative of the corresponding gene family; weights count the number of genes of that family present in that genome: in the example, node 1 corresponds to CC1 in (A), and has therefore edges incident to genomes A and B, each of weight 2.

Shared gene networks were introduced precisely to tackle these issues (Figure 1) [17,19,32]. These networks represent which genomes share genetic material, without prejudice regarding the processes involved (vertical descent, but also introgressions [19,33,34]). In genome networks, all entities are not necessarily genealogically related, allowing for simultaneous analysis of mobile genetic elements and cellular evolution. In that respect, the social microbial network is more inclusive than the tree of life, which is restricted to one type of relationship between one

fraction of the biological diversity [6]. Two genomes with a direct connection in such a graph are similar in the sense that they share at least one gene family, whereas two genomes connected only by an indirect path are not similar in terms of gene content. These genome networks display some structure. First of all, plasmids are more central (higher **betweenness** [35] for a given **degree**) and viruses more peripheral, testifying that plasmids are general couriers for gene transmission amongst microbes [19]. Second, genome networks have several **connected components**, that is, several sets of genomes for which there is always an interconnecting path. Each of these connected components groups genomes with exclusive, non-overlapping sets of gene families, and thus corresponds to pools of genes uniquely associated with these genomes [19]. The existence of different connected components suggests the existence of restrictions to introgression.

Within a connected component, a genome network only shows that genomes share genes, but not what the shared genes are. Typically, a triangle of three connected genomes (A, B, C) may result from the sharing of different genes for each pair (AB, BC, AC) within this triangle [4] (see Figure 1B,C). Thus, genomes may form tightly clustered **communities** [20] in these graphs while sharing different genes. Genome networks provide general information about barriers to transmission and about genetic partnerships, suggesting clubs of genomes enjoying **public genetic goods** [4,20]. These genome networks require, however, further specifications (for example, on their edges) to address detailed questions about gene transmission and its barriers. A more informative representation displays the identity of shared genes along each edge of a genome network, like in [36], which showed some gene sharing between bacteriophages (as early as 1999), or as in [37] that unraveled genetic transmission between mobile genetic elements of giant viruses (as recently as 2013). Such **multiplex graphs** are unquestionably attractive and rather natural representations of genetic sharing. However, their display becomes rapidly complex for large datasets, and from an analytical point of view, other graphs can offer practical advantages to analyze gene transmission beyond the genome network framework.

Introducing Bipartite Graphs in Evolutionary Studies

The information on the identity of shared edges (here, gene families) can be conserved in a less cluttered fashion by using bipartite 'gene families–genomes' graphs. In these graphs, the precise information regarding gene sharing is directly encoded as edges between these two kinds of nodes. Multiplex genome networks can be seen as unimodal projections [38] of such bipartite 'gene families–genomes' graphs (Figure 1D). Bipartite graphs include the same diversity of genomes as the genome networks described above, but they are more accurate. Importantly, simple specific bioinformatic treatments of these multilevel graphs allow one to rapidly identify which groups of genes are shared by which groups of genomes [39], and to display and compare different channels of gene transmission, that is, the routes across generations through which hereditary resources or information pass from parent to offspring [31].

As in genome networks, connected components produce an informative partition of the data. This partition can moreover be examined at different levels of similarity by tuning, for example, the sequence identity percentage. When the data consist of all the protein sequences from all the complete viral (3749), plasmidic (4350), and archaeal (152) genomes, together with a representative subsample of the eubacteria (230) from NCBI, we get the numbers shown in Table 1.

Assuming a rough molecular clock, these thresholds are useful for investigating events of different ages. Sequences with $\geq$90% identity have a relatively weak divergence with respect to sequences with 30% identity; indeed, these latter have likely diverged faster or for a longer period of time.

This representation of gene families–genomes bipartite graphs is explicitly multilevel. Interestingly, its analysis does not require any graph clustering algorithm (whose results tend to vary

Table 1. Statistics of the Prokaryote–Virus–Plasmid Gene Families–Genomes Bipartite Graphs[a]

Minimal identity percentage to connect sequences	30%	60%	90%
Number of connected components (CC)	156	375	488
Number of CC having only plasmids	25	73	155
Number of CC having only viruses	130	299	297
Size of the giant connected component (number of nodes)	6362	5143	2769

[a]For reciprocal 80% length cover, and different identity thresholds.
The data consist of all protein sequences from all complete plasmidic, viral, and archaeal genomes from NCBI (as of 11/2013), as well as one complete eubacterial complete genome for each family. The identity percentage describes the similarity, in terms of the conservation of primary sequences, between pairs of molecules. The higher this 'identity threshold' the more similar pairs of sequences must be to be directly connected in a sequence-similarity network. For high 'identity threshold', connected components consist of highly conserved sequences. In a first molecular clock-like approximation, higher 'identity thresholds' define groups of sequences that diverged more recently from one another than groups defined with lower 'identity threshold'.

considerably with their implementation). Genetic transmission among microbes can be investigated by simple topological notions of bipartite graphs that result in biologically relevant observations: **twins** and **articulation points** [40] that we detail below.

We apply here these notions only to gene family nodes. 'Twin' is a notion of graph theory; applied to gene families–genomes graphs, it singles out 'fellow travellers': gene families are twins when they are present in exactly the same set of genomes. In the language introduced in [34], the **support** of such a twin defines a **club of genomes.** Clubs of genomes, when composed of individuals pertaining to different species, could encourage further studies of 'kin-coevolution', for example, the fact that genetic divergence affecting multiple ecologically coexisting lineages, that exchanged genes at some point of their evolution, produces multilineage persistent clubs. The bipartite graph can be simplified by grouping together sets of gene families that are shared by exclusive groups of genomes, and by replacing each such group of gene families by a super-node. Nodes that remain untouched by this reduction process are considered as trivial twin classes (and result in trivial super-nodes). Technically, there is no difference between trivial and non-trivial twins, although, from the biological perspective, the latter correspond to groups of gene families that are more likely to be transmitted together. The resulting **quotient graph** is reduced, because every club of genomes is now defined by one super-node (individual gene family or group of gene families hosted in this club of genomes) while no information is lost (Figure 2). This property means that even very large graphs can be investigated. In the dataset presented in Table 1, we typically find clubs, such as the one composed of the firmicute *Enterococcus faecalis* and nine plasmids (present in lactococci or enterococci) that simultaneously and exclusively share the following gene families (at 90% identity): ribose 5-phosphate isomerase RpiB, galactose mutarotase and related enzymes, β-glucosidase/6-phospho-β-glucosidase/β-galactosidase, and phosphotransferase system cellobiose-specific component IIA. These shared mobilized gene families are involved in neighbor pathways of sugar metabolisms (specifically in glycolysis and in the pentose phosphate metabolic pathways), which likely explains their collective mobilization in plasmids.

Articulation points in a gene families–genomes bipartite graph correspond to gene families shared by many genomes with otherwise totally distinct gene contents (for a given similarity threshold). Although strictly topological, the notion of an articulation point is thus expected to help detect public genetic goods [34], that is, genetic material that is being shared by taxonomically distant genomes, which possibly benefit from the properties they confer, for some reason other than genealogy (i.e., genes coding for environmental adaptation or hitch-hiking with those). However, an articulation point can also detect selfish genes, such as the abundant transposases [41], which are spreading across multiple distantly related genomes (Box 2).

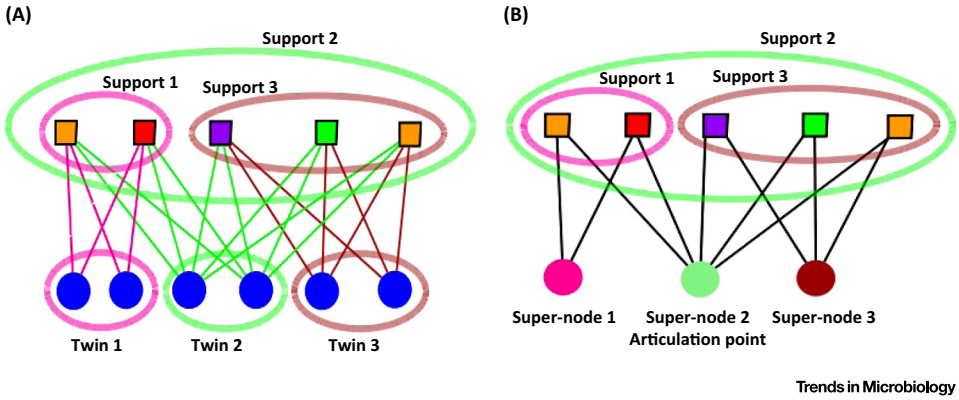


Figure 2. Twins and Articulation Points in a Bipartite Graph. (A) Top nodes in this bipartite graph are genomes and bottom nodes gene families. Nodes in each colored ellipse at the bottom form a twin class, since their sets of **neighbors** (supports encircled by similarly colored ellipses on the top level) are identical (as highlighted by the coloring of their incident edges). (B) Collapsing twin nodes into super-nodes yields a reduced graph, without further bottom twin nodes. The supported groups of host genomes are unchanged, and are now defined as the neighbors of a single super-node. Due to the graph reduction, the green super-node is now an articulation point, since its removal disconnects the nodes in the pink and brown supports.

Investigating Composite Genes, Organisms, and Evolutionary Transitions with Sequence-Similarity Networks

Introgression can also be investigated below the gene level and above the organismal level. For instance, composite genes, such as the genes produced by evolutionary tinkering [42], famous for encoding multidomain proteins, are well documented in cellular genomes [43–45], and have been reported in viruses and plasmids [46,47]. Such genes are composed of genetic fragments (e.g., components, which can be domains or full genes) that are otherwise found in separate gene families [48]. The fusion of a receptor-binding protein with a tail fiber protein in the lactococcal bacteriophage bIBB29, producing a composite gene involved in host specificity, offers a good example of this sort of molecular mosaic [49]. While many substitution models have been developed to account for gradual evolution by point mutation in phylogenetic inferences, models describing the rules and rates of emergence (or fission) of composite genes are still rare [50–52], especially for unicellular organisms and mobile genetic elements [46,47]. However, many gene families are not just evolving gradually within the boundaries of a single gene family [53]. The accretion of two protein domains into a novel host structure constitutes a case of saltatory molecular evolution by introgression. The rules of evolution and fragment combination largely remain to be discovered [54,55]. Sequence-similarity networks could contribute to this task. Indeed, these graphs can: (i) provide a systematic description of both composite and component genes in genomes (and metagenomes); (ii) be used to polarize fusion and fission events (by comparing the taxonomical distribution of genes hosts in associated component and composite gene families); (iii) be directly used to compare the relative conservation of overlapping component and composite sequences, for example, to determine whether domains found in different combination have different rates of evolution. The detection of composite genes using sequence-similarity networks can further contribute to understanding the rules of evolution of other biological networks, such as protein–protein interaction networks [56]. For instance, when, as a result of exon- or domain-shuffling, composite genes produce novel combinations of domains of interaction, composite genes can introduce novel nodes and edges in protein–protein interactions. Likewise, composite genes can impact the robustness of protein–protein interaction networks, when genes coding for separate proteins involved in a functional interaction become fused, ‘crystallizing’ an edge of the protein–protein interaction network.

Box 2. Articulation Points Reveal Potential Public Genetic Goods

In a prokaryote–virus–plasmid dataset, we typically find clubs of genomes, such as the one (represented in Figure I) composed of two mesophilic sulphur-reducing acetate-metabolizing Proteobacteria (*Geobacter sulfurreducens* and *Desulfobacca acetoxidans*) and two thermophilic hydrogenotrophic methanogen Euryarchaeota (*Methanocella conradii* and *Methanocella paludicola*). These taxa are linked by an articulation point, which indicates the sharing of a conserved gene family (at >90% identity), functionally annotated as a tRNA (1-methyladenosine) methyltransferase. This kind of association between sulphate-reducer and methanogens is well-documented in the literature [96,97]. The sharing of genes between different prokaryotes suggested by this network analysis makes sense, since these prokaryotes are found in common anoxic environments, such as rice paddy soils [98]. Also, *G. sulfurreducens* and *M. paludicola* contain a laterally-transferred two-gene cluster, *hgcAB*, related to the ability to methylate mercury [99]. Thus, a graph analysis produces a novel testable hypothesis, namely, to see if the shared tRNA methyltransferase is involved in the adaptation to the environment of these taxa, or if it hitch-hiked with other genes transferred between these taxa, such as the *hcgAB* cluster.

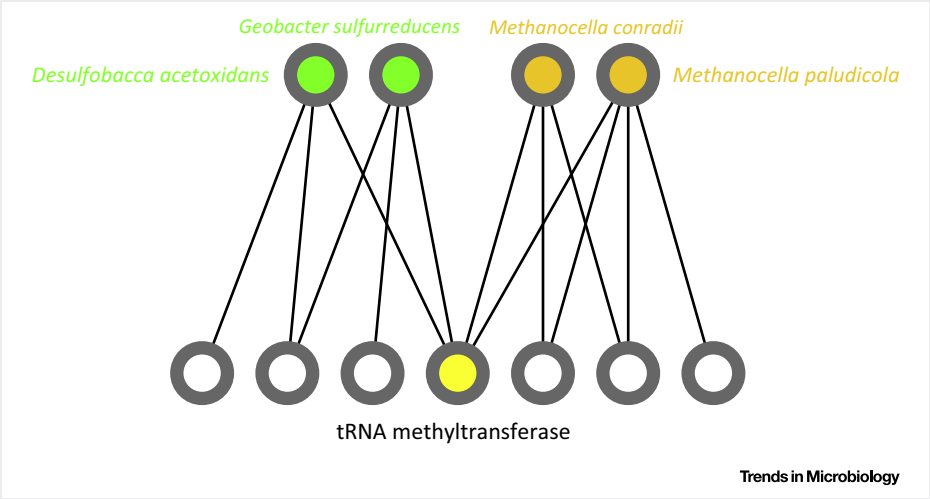


Figure I. Excerpt of a Typical Reduced Gene Familes–Genomes Bipartite Graph around an Articulation Point. The top nodes compose the club defined by the sharing of a conserved tRNA methyltransferase (bottom node in yellow). For simplicity, only the direct neighbors of the members of the club have been included in the picture of the graph. The removal of the articulation point (in yellow) isolates the two taxonomically homogeneous groups from each other.

This issue takes on particularly fundamental importance in organisms hosting genes from multiple origins. These introgressed genes have distinct evolutionary past histories and, hence, possibly different future evolvabilities. For example, eukaryotic genomes [57,58] as well as major archaeal lineages are composed of genes from both bacterial and archaeal origins [59,60]. Some studies have focused on the evolution of complete genes of distinct origins in these mosaic taxa [i.e., contrasting the essentiality or centrality of genes from bacterial and archaeal origins in regulatory or metabolic eukaryotic networks [61], or simply performing classic phylogenetic analyses of these genes to identify endosymbiotic gene transfer (or EGT) [8]]. A common fate for proteins derived from such transferred organellar genes is to be targeted back to the compartment of origin to perform their original function, but not only [62]. Regarding these proteins and genes, the study of composite organisms has opened the door to an exciting evolutionary question that, we argue, networks can now better address: what happens after distinct genetic material becomes integrated into a new host? Genes from distinct origins could have different propensities to be lost or to diverge during subsequent evolution of their novel composite host lineage [63]. Likewise, at the infragenic level, the evolutionary impact of introgression deserves consideration. Do composite organisms host novel symbiogenetic composite genes with components from different phylogenetic origins that could only be born in such genetic melting-pots as a result of the original mixing of gene fragments? A positive answer, that is, the detection of such novel composite genes in composite organisms, could

revolutionize our understanding of the origins of biological traits. A negative answer, that is, the lack of novel composite genetic material from different organismal sources despite their new physical proximity, would indicate strong selective pressures preventing the birth of novel gene families in spite of changes in their genomic context. Thus, it would be worth testing if introgression at one level of biological organization (i.e., between cells) can favor introgression at another level (i.e., between genes). For example, organisms with composite genomes, or holobionts, might be composed of more composite symbiogenetic genes than organisms devoid of endosymbionts, or less subjected to gene transfer.

Sequence-similarity networks are ideal tools for investigating these issues (Figure 3). These very inclusive graphs [47,53] allow for comparative analyses of massive datasets without the need for multiple sequence alignments [4,64–66]. Similarity is typically detected in a BLAST all-versus-all analysis to produce a table of pairwise hits [67]. Sequence-similarity networks are displayed and analyzed as a set of connected components (Figure 1A) [68]. When the coverage between sequences is high, this partition of the nodes defines groups of putative homologous sequences or gene families. Thus, sequence-similarity networks have been used with relatively stringent criteria (i.e., hits between two sequences must show >30% identity, cover ≥80% of both sequences length, and have a maximal E-value of 10^{-5} in BLAST comparative analyses) coupled with clustering methods to identify clusters of nodes corresponding to homologous gene families [69–71]. In the past 20 years, sequence-similarity networks have indeed mainly been used to investigate the evolution of protein-coding genes [4,64–66,71–75], and to perform functional annotation. For instance, the COG categories correspond to groups of similar sequences (with remarkable topological properties in sequence-similarity networks) that have likely evolved from a single ancestral gene. In comparative analyses, COG are often used as proxy for functional annotations because their remarkable conservation suggests that sequences from the same COG may have preserved some common functions [71]. This standard approach, however, would not readily detect composite genes [76]. Using less stringent thresholds for mutual sequence coverage (Figure 3A) or identity percentage, sequence-similarity networks can be used to detect superfamilies [66,77–79], divergent homologues, or composite genes [when, for example, the length coverage condition is relaxed to take into account (partial) similarity (Figure 3C), such as domain sharing, between sequences] [46].

These kinds of analyses with flexible definitions confirm that not all eukaryotic gene families have homologs in prokaryotes. When they do, sequence-similarity network analyses indicate that eukaryotic gene families homologous to those of bacteria (for which sequences of eukaryotes exclusively cluster with sequences from bacteria [63]) and eukaryotic gene families homologous to those of archaea (for which sequences of eukaryotes exclusively cluster with sequences from archaea) have different rates of evolution. For example, eukaryotic gene families with bacterial origins are more easily expanded or lost when eukaryotic genomes expand or shrink, while the number of eukaryotic gene families with archaeal origins is much more stable [22,63].

Moreover, sequence-similarity networks demonstrated their efficiency to unravel distant homologues in eukaryotic genomes, that is, gene families for which some present-day eukaryotes possess a version that originated from a bacterial progenitor, while other present-day eukaryotes possess an homologous version that originated from an archaeal progenitor, or when the same eukaryotes possess both diverged versions in its nuclear genome, one from a bacterial origin, the other from an archaeal origin [63] (Figure 3B). The latter presence of such distant homologues characterizes the occurrence of EGT [7,59], an introgressive process where a gene from an organelle (such as mitochondria or plastids) has been imported into the eukaryotic nuclear DNA, where an homologous nuclear copy from archaebacterial origin was already present (Figure 3D). These networks are promising to look for possibly still-hidden EGT, and past endosymbioses when they are applied to new genomic data.

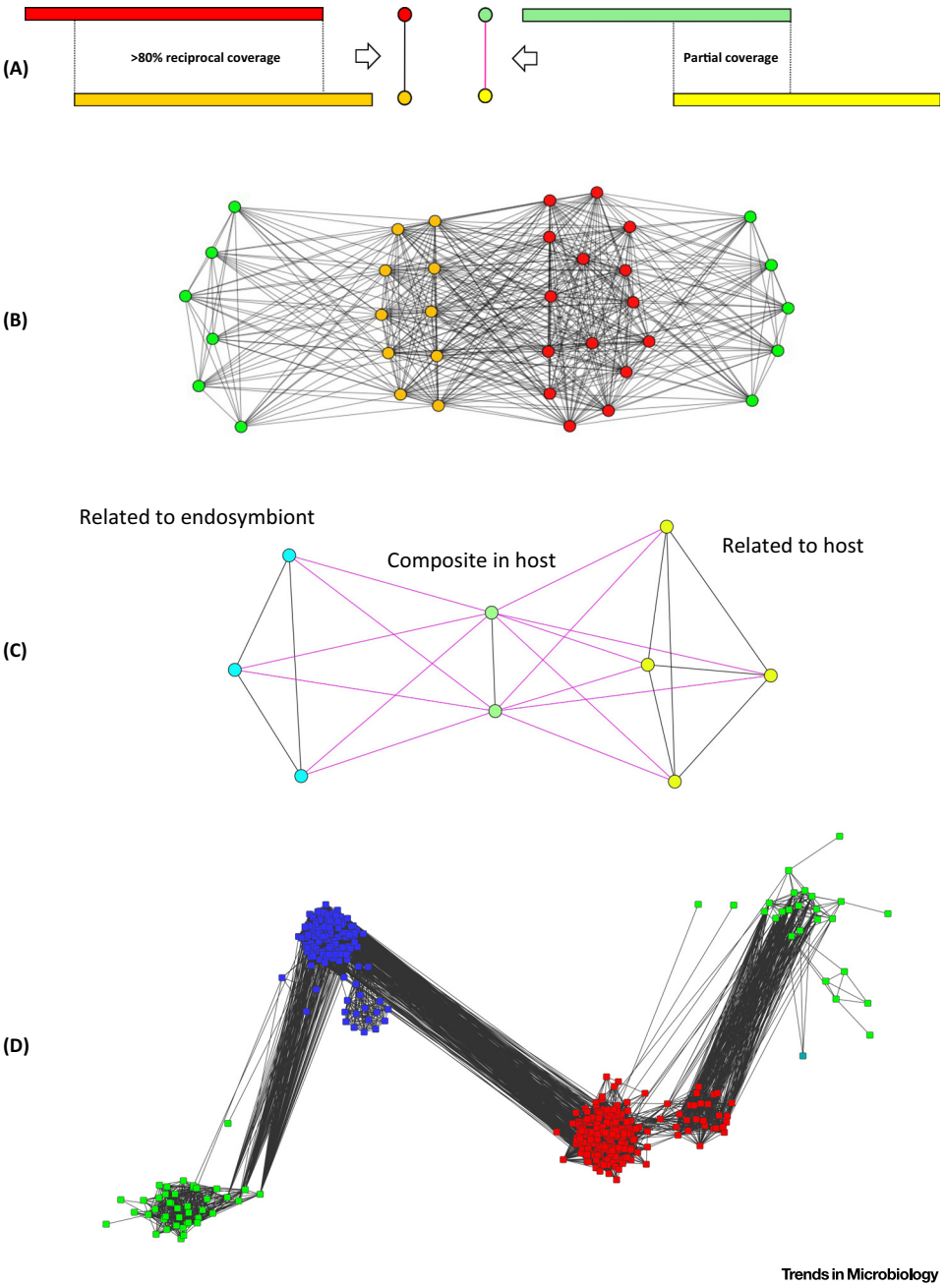


Figure 3. Typical Patterns for Candidate Endosymbiotic Gene Transfer (EGT) and Composite Genes in Sequence-Similarity Networks. (A) Sequence-similarity networks can be used for the detection of distant homologues in eukaryotic genomes. Complete (left) and partial (right) sequence similarity, and how they are translated as different types of edges in the sequence-similarity network (SSN). In black, the percentage of reciprocal cover is high; the sequences are homologous over their entire length. In purple, the cover percentage is low; the sequences are only partly similar, that is, they share a homologous domain. (B) Shortest-path analysis in a sequence-similarity graph can be used for detecting possible endosymbiotic gene transfer (EGT). Indeed, EGT results in a characteristic network pattern: an indirect short path along which all edges indicate homology, connecting two nodes corresponding to diverged sequences present in a given host organism. Green nodes represent eukaryotic sequences; red, bacterial sequences; and yellow, archaeal sequences. Black edges denote complete sequence similarity (>80% length). All shortest paths between eukaryotic sequences that pass through the bacterial and archaeal components are likely candidates for EGT, because this indicates that a first type of eukaryotic sequence has affinities to bacterial sequences while a second type has affinities to archaeal ones. (C) Sequence-similarity networks with edges for complete and partial coverage are also useful for the detection of composite genes. The

Sequence-similarity networks are also most useful for identifying composite genes (Figure 3C), and their use for detecting genes composed of parts from different origins will likely soon aid reticulate evolution analyses [46,47,53]. Indeed, the level of molecular intricacy between hosts and symbionts may well exceed whole gene introgression in the genome of composite organisms. Preliminary results show that photosynthetic eukaryotes contain some novel nuclear composite genes, featuring unique couplings of domains from plastid origin, without any counterpart in the prokaryotic world. For example, photosynthetic dinophytes contain a composite gene coding for a protein consisting of two domains: one SufE domain of cyanobacterial origin (i.e., probably originating from the chloroplast genome) and a tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase of proteobacterial origin. Interestingly, SufE displays desulfurylase activity [80], while the tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase posesses a thiol group (R-S-H) containing a sulfur atom. It is possible that the sulfur atom required for the thiol group is provided by the activity of SufE to the new physical coupling of these domains in a symbiogenetic gene. Such findings encourage experimental studies to establish whether and which biological properties emerged from the physical coupling of domains in a novel eukaryotic gene with endosymbiotic origin.

Understanding the entanglement of molecular building blocks, below and above the gene level, is probably the next step required to analyze molecular processes going on during evolutionary transitions mediated by the merging of lineages [4,57–60].

Concluding Remarks: Networks Enhance Our Comprehension of Life's Complexity

The complexity and diversity of phenomena acknowledged and investigated by evolutionary biologists is striking, and growing: it now goes well beyond the identification of lineage divergence from a single common ancestor, enhancing what is considered as the Darwinian paradigm. When pushed to its limits, introgression might result in the integration of laterally acquired features into a sustainable structure, controllable by regulatory systems, which may themselves be the result of introgression. A technical and theoretical transition has accompanied this broadening of scope within the evolutionary paradigm. Namely, network models and methods, never truly absent in biological studies [81], have been developed and implemented. Hence, they now offer powerful complementary approaches to evolutionary studies, which will enhance the exploitation of molecular datasets in multiple directions. The routes and genetic goods of microbial social life, the origins and combination rules of composite genes, and the genetic transformation coupled with major evolutionary transitions, can readily be investigated using powerful, inclusive, comparative network-based tools. The diversity of such tools is itself constantly increasing: the multi-thresholded sequence-similarity networks, (multiplex) genome networks, and the bipartite graphs presented here, allow one to perform multi-agent and multilevel comparative analyses, and may become as familiar to evolutionary biologists as phylogenetic trees in the near future. Importantly, these network tools have not yet been used

figure shows a pattern associated with the detection of composite genes. Black edges denote complete (>80% cover) and purple edge denote partial (<80% cover) sequence similarity. The green family is a candidate symbiogenetic composite gene, derived from endosymbiotic lateral gene transfer, since it displays one part with similarity to host-related sequences (yellow) and another part with similarity to endosymbiont-related (blue) genes. (D) A concrete example of a possible EGT: archaeal sequences are represented in blue, eubacterial in red, and eukaryotic genes in green (there is also a single plasmidic sequence in blue-green on the right). Eukaryotic sequences clearly form two groups, one closer to archaea, one more related to eubacteria. All the sequences have a generic annotation as RNA-pseudouridine synthase, but while the eubacterial (and related eukaryotic) sequences are exclusively tRNA synthases (thus putatively of mitochondrial origin), on the archaeal side (thus possibly of host origin) we find tRNA- as well as rRNA-pseudouridine synthases. It indeed turns out that this family contains two pseudouridine synthase genes that are both present in *Saccharomyces cerevisiae*, having a similar function but acting on a different substrate: one on the archaeal side, coding for *Cbf5p* that acts on large and small rRNA [100,101], and the other on the eubacterial side, coding for *Pus4*, that acts on mitochondrial and cytoplasmic tRNA-uridine [102].

Outstanding Questions

What are the rules of domain and gene shuffling in microbes? Sequence-similarity networks provide fast and effective means for systematic analyses of the evolution of composite genes, by simultaneously detecting families of components contributing to composite gene families. The phylogenetic origins and the functional categories of these components will show whether microbes are using transferred genes to create new composite genes in their genomes. For example, do the notoriously mosaic haloarchaeal genomes harbor composite genes of bacterial origin? Does the proportion of composite genes in microbes change with the environment? Can one introduce models of nucleotide substitution into sequence-similarity networks in order to make them more realistic with regard to sequence evolution?

Is every gene everywhere? Gene-similarity networks applied to large-scale metagenomic data and gene-sharing networks featuring environments instead of genomes as their nodes will provide inclusive novel ways to address this important question. These graphs will show whether similar sequences are found in geographically or ecologically similar environments, and serve to detect ubiquitous and endemic genes sets.

What phenotypes in holobionts have multiple origins, that is, did not evolve within a single phylum but emerged from a biological collective? Bipartite graphs with microbial taxa or microbial gene families as bottom nodes and with animal or human hosts as top nodes will immediately allow for the identification of phylogenetically heterogeneous groups of microbes, or groups of gene families in microbes, always associated with a particular host-level phenotype.

How do processes of molecular evolution occurring at the level of the microbiota affect eukaryotic hosts? The microbial gene families–eukaryotic host bipartite graphs described above can be refined to take into account information about the molecular evolution of the gene families (e.g., their rate of evolution, or whether to what extent and by what mobile elements each gene family was eventually transferred). This adds an explicit evolutionary dimension to the bottom-level nodes, allowing one to evaluate, for example,

at their full potential (see Outstanding Questions). In particular, they could also be used to analyze the evolution of communities of synthetic microorganisms, biofilms, and holobionts. These latter collective systems encompass a challenging complexity. For example, holobionts rely on a multiplicity of interacting transmission systems and channels for their development and evolution that differ in the microbes and in their hosts. This heterogeneity complicates the understanding of the causes of holobionts' collective phenotypes by traditional methods, even in the metazoan world [82]. Applying a network analytical framework to holobiont studies may be an innovative way to decipher what traits, long held as characteristic of a single animal (i.e., species incompatibility, self-immunity, or possibly behavior [83,84]), or of an individual organism/biofilm (i.e., health conditions [85,86] or drug resistance), originate from complex interactions, at multiple biological levels, and how these involve microbes and their genes. More generally, network-thinking has lots to contribute to microbiology.

the impact of lateral gene transfer, operating at the microbial level, on the phenotypes of the eukaryotic host. For example, it becomes easy to test whether laterally transferred genes, mobilized by a broader range of mobile elements, are more largely distributed in human hosts than are resident gene families of the microbiome.

Can one extend the methods from bipartite to tripartite graphs, to account for more levels of biological organization? This defines, as a realistic objective, the implementation of genes–genomes–environments tripartite graphs, which can then be clustered to provide a global yet accurate representation of the structure of genetic diversity on Earth in a single comparative analysis.

Acknowledgments

E.C. and E.B. are funded by FP7/2007-2013 Grant Agreement #615274.

References

1. Darwin, C. (1859) *On the Origin of Species by Means of Natural Selection,* John Murray
2. O'Hara, R.J. (1997) Population thinking and tree thinking in systematics. *Zool. Scr.* 26, 323–329
3. Doolittle, W.F. and Bapteste, E. (2007) Pattern pluralism and the Tree of Life hypothesis. *Proc. Natl. Acad. Sci. U.S.A.* 104, 2043–2049
4. Bapteste, E. *et al.* (2012) Evolutionary analyses of non-genealogical bonds produced by introgressive descent. *Proc. Natl. Acad. Sci. U.S.A.* 109, 18266–18272
5. Doolittle, W.F. (1999) Phylogenetic classification and the universal tree. *Science* 284, 2124–2129
6. Bapteste, E. (2014) The origins of microbial adaptations: How introgressive descent, egalitarian evolutionary transitions and expanded kin selection shape the network of life. *Front. Microbiol.* 5, 1–4
7. Archibald, J.M. (2015) Genomic perspectives on the birth and spread of plastids: Fig 1. *Proc. Natl. Acad. Sci. U.S.A.* 112, 10147–10153
8. Lane, C.E. and Archibald, J.M. (2008) The eukaryotic tree of life: endosymbiosis takes its TOL. *Trends Ecol. Evol.* 23, 268–275
9. Claverie, J-M. and Ogata, H. (2009) Ten good reasons not to exclude giruses from the evolutionary picture. *Nat. Rev. Microbiol.* 7, 615
10. Koonin, E.V. *et al.* (2009) Compelling reasons why viruses are relevant for the origin of cells. *Nat. Rev. Microbiol.* 7, 615
11. Moreira, D. and López-García, P. (2009) Ten reasons to exclude viruses from the tree of life. *Nat. Rev. Microbiol.* 7, 306–311
12. Navas-Castillo, J. (2009) Six comments on the ten reasons for the demotion of viruses. *Nat. Rev. Microbiol.* 7, 615
13. Raoult, D. (2009) There is no such thing as a tree of life (and of course viruses are out!). *Nat. Rev. Microbiol.* 7, 615
14. Villarreal, L.P. and Witzany, G. (2010) Viruses are essential agents within the roots and stem of the tree of life. *J. Theor. Biol.* 262, 698–710
15. Lukjancenko, O. *et al.* (2010) Comparison of 61 sequenced *Escherichia coli* genomes. *Microb. Ecol.* 60, 708–720
16. Tettelin, H. *et al.* (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. U.S.A.* 102, 13950–13955
17. Dagan, T. *et al.* (2008) Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc. Natl. Acad. Sci. U.S.A.* 105, 10039–10044
18. Dagan, T. and Martin, W. (2007) Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc. Natl. Acad. Sci. U.S.A.* 104, 870–875
19. Halary, S. *et al.* (2010) Network analyses structure genetic diversity in independent genetic worlds. *Proc. Natl. Acad. Sci. U.S.A.* 107, 127–132
20. Skippington, E. and Ragan, M.A. (2011) Lateral genetic transfer and the construction of genetic exchange communities. *FEMS Microbiol. Rev.* 35, 707–735
21. Shapiro, J.A. (2007) Bacteria are small but not stupid: cognition, natural genetic engineering and socio-bacteriology. *Stud. Hist. Philos. Biol. Biomed. Sci.* 38, 807–819
22. Ku, C. *et al.* (2015) Endosymbiotic origin and differential loss of eukaryotic genes. *Nature* 524, 427–432
23. Lobkovsky, A.E. *et al.* (2014) Estimation of prokaryotic supergenome size and composition from gene frequency distributions. *BMC Genomics* 15, S14
24. Kloesges, T. *et al.* (2011) Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. *Mol. Biol. Evol.* 28, 1057–1074
25. Popa, O. and Dagan, T. (2011) Trends and barriers to lateral gene transfer in prokaryotes. *Curr. Opin. Microbiol.* 14, 615–623
26. Popa, O. *et al.* (2011) Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res.* 21, 599–609
27. Jain, R. *et al.* (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl. Acad. Sci. U.S.A.* 96, 3801–3806
28. Park, C. and Zhang, J. (2012) High expression hampers horizontal gene transfer. *Genome Biol. Evol.* 4, 523–532
29. Sorek, R. *et al.* (2007) Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* 318, 1449–1452
30. Cohen, O. *et al.* (2011) The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. *Mol. Biol. Evol.* 28, 1481–1489
31. Lamn, E. (2014) Inheritance systems. In *The Stanford Encyclopedia of Philosophy* (Winter 2014) (Zalta, E.N., ed.), http://plato.stanford.edu/archives/win2014/entries/inheritance-systems
32. Lima-Mendez, G. *et al.* (2008) Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol. Biol. Evol.* 25, 762–777
33. Halary, S. *et al.* (2013) EGN: a wizard for construction of gene and genome similarity networks. *BMC Evol. Biol.* 13, 146
34. McInerney, J.O. *et al.* (2011) The public goods hypothesis for the evolution of life on Earth. *Biol. Direct* 6, 41
35. Brandes, U. (2008) On variants of shortest-path betweenness centrality and their generic computation. *Soc. Netw.* 30, 136–145
36. Hendrix, R.W. *et al.* (1999) Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. *Proc. Natl. Acad. Sci. U.S.A.* 96, 2192–2197
37. Yutin, N. *et al.* (2013) Virophages, polintons, and transpovirons: a complex evolutionary network of diverse selfish genetic elements with different reproduction strategies. *Virol. J.* 10, 158

38. Ahn, Y-Y. *et al.* (2011) Flavor network and the principles of food pairing. *Sci. Rep.* 1, 196
39. Rivera, C.G. *et al.* (2010) NeMo: network module identification in cytoscape. *BMC Bioinformatics* 11 (Suppl. 1), S61
40. Diestel, R. (2006) *Graph Theory,* Springer Science & Business Media
41. Aziz, R.K. *et al.* (2010) Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res.* 38, 4207–4217
42. Derouiche, A. *et al.* (2015) Evolution and tinkering: what do a protein kinase, a transcriptional regulator and chromosome segregation/cell division proteins have in common? *Curr. Genet.* Published online August 19, 2015. http://dx.doi.org/10.1007/s00294-015-0513-y
43. Kawashima, T. *et al.* (2009) Domain shuffling and the evolution of vertebrates. *Genome Res.* 19, 1393–1403
44. Chothia, C. (2003) Evolution of the protein repertoire. *Science* 300, 1701–1703
45. de Souza, S.J. (2012) Domain shuffling and the increasing complexity of biological networks. *Bioessays* 34, 655–657
46. Jachiet, P.A. *et al.* (2013) MosaicFinder: Identification of fused gene families in sequence similarity networks. *Bioinformatics* 29, 837–844
47. Jachiet, P. *et al.* (2014) Extensive gene remodeling in the viral world: new evidence for non-gradual evolution in the mobilome network. *Genome Biol. Evol.* 6, 2195–2205
48. Cheng, S. *et al.* (2014) Sequence similarity network reveals the imprints of major diversification events in the evolution of microbial life. *Front. Ecol. Evol.* 2, 1–13
49. Hejnowicz, M.S. *et al.* (2009) Analysis of the complete genome sequence of the lactococcal bacteriophage bIBB29. *Int. J. Food Microbiol.* 131, 52–61
50. Pasek, S. *et al.* (2006) Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics* 22, 1418–1423
51. Kummerfeld, S.K. and Teichmann, S.A. (2005) Relative rates of gene fusion and fission in multi-domain proteins. *Trends Genet.* 21, 25–30
52. Snel, B. *et al.* (2000) Genome evolution. *Trends Genet.* 16, 9–11
53. Haggerty, L.S. *et al.* (2014) A pluralistic account of homology: adapting the models to the data. *Mol. Biol. Evol.* 31, 501–516
54. Patthy, L. (2003) Modular assembly of genes and the evolution of new functions. *Genetica* 118, 217–231
55. Nakamura, Y. *et al.* (2007) Rate and polarity of gene fusion and fission in *Oryza sativa* and *Arabidopsis thaliana*. *Mol. Biol. Evol.* 24, 110–121
56. Dohrmann, J. *et al.* (2015) Global multiple protein–protein interaction network alignment by combining pairwise network alignments. *BMC Bioinformatics* 16, S11
57. McInerney, J.O. *et al.* (2014) The hybrid nature of the Eukaryota and a consilient view of life on Earth. *Nat. Rev. Microbiol.* 12, 449–455
58. Williams, T.A. *et al.* (2013) An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* 504, 231–236
59. Nelson-Sathi, S. *et al.* (2012) Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proc. Natl. Acad. Sci. U.S.A.* 109, 20537–20542
60. Nelson-Sathi, S. *et al.* (2015) Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* 517, 77–80
61. Alvarez-Ponce, D. and McInerney, J.O. (2011) The human genome retains relics of its prokaryotic ancestry: human genes of archaebacterial and eubacterial origin exhibit remarkable differences. *Genome Biol. Evol.* 3, 782–790
62. Deane, J.A. *et al.* (2000) Evidence for nucleomorph to host nucleus gene transfer: light-harvesting complex proteins from cryptomonads and chlorarachniophytes. *Protist* 151, 239–252
63. Alvarez-Ponce, D. *et al.* (2013) Gene similarity networks provide tools for understanding eukaryote origins and evolution. *Proc. Natl. Acad. Sci. U.S.A.* 110, E1594–E1603
64. Yona, G. *et al.* (2000) ProtoMap: automatic classification of protein sequences and hierarchy of protein families. *Nucleic Acids Res.* 28, 49–55
65. Frickey, T. and Lupas, A. (2004) CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* 20, 3702–3704
66. Atkinson, H.J. *et al.* (2009) Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS ONE* 4, e4345
67. Forster, D. *et al.* (2014) Testing ecological theories with sequence similarity networks: marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms. *ISME J.* 13, 1–16
68. Bittner, L. *et al.* (2010) Some considerations for analyzing biodiversity using integrative metagenomics and gene networks. *Biol. Direct* 5, 47
69. Altenhoff, A.M. *et al.* (2015) The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Res.* 43, D240–D249
70. Sayers, E.W. *et al.* (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 39, D38–D51
71. Tatusov, R.L. *et al.* (1997) A genomic perspective on protein families. *Science* 278, 631–637
72. Bapteste, E. *et al.* (2013) Networks: expanding evolutionary thinking. *Trends Genet.* 29, 439–441
73. Enright, A.J. and Ouzounis, C.A. (2000) GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics* 16, 451–457
74. Li, L. *et al.* (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189
75. Enright, A.J. *et al.* (2003) Protein families and TRIBES in genome sequence space. *Nucleic Acids Res.* 31, 4632–4638
76. Song, N. *et al.* (2008) Sequence similarity network reveals common ancestry of multidomain proteins. *PLoS Comput. Biol.* 4, e1000063
77. Sasson, O. *et al.* (2003) ProtoNet: hierarchical classification of the protein space. *Nucleic Acids Res.* 31, 348–352
78. Matsui, M. *et al.* (2013) Comprehensive computational analysis of bacterial crp/fnr superfamily and its target motifs reveals stepwise evolution of transcriptional networks. *Genome Biol. Evol.* 5, 267–282
79. Rappoport, N. *et al.* (2013) ProtoNet: charting the expanding universe of protein sequences. *Nat. Biotechnol.* 31, 290–292
80. Ollagnier-de-Choudens, S. *et al.* (2003) Mechanistic studies of the SufS-SufE cysteine desulfurase: evidence for sulfur transfer from SufS to SufE. *FEBS Lett.* 555, 263–267
81. Ragan, M.A. (2009) Trees and networks before and after Darwin. *Biol. Direct* 4, 43
82. Selosse, M-A. *et al.* (2014) Microbial priming of plant and animal immunity: symbionts as developmental signals. *Trends Microbiol.* 22, 607–613
83. Brucker, R.M. and Bordenstein, S.R. (2012) Speciation by symbiosis. *Trends Ecol. Evol.* 27, 443–451
84. Brucker, R.M. and Bordenstein, S.R. (2013) The hologenomic basis of speciation: gut bacteria cause hybrid lethality in the genus *Nasonia*. *Science* 341, 667–669
85. Hur, K.Y. and Lee, M-S. (2015) Gut microbiota and metabolic disorders. *Diabetes Metab. J.* 39, 198–203
86. Gilbert, S.F. *et al.* (2012) A symbiotic view of life: we have never been individuals. *Q. Rev. Biol.* 87, 325–341
87. Dunning Hotopp, J.C. *et al.* (2007) Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* 317, 1753–1756
88. La Scola, B. *et al.* (2008) The virophage as a unique parasite of the giant mimivirus. *Nature* 455, 100–104
89. Boyer, M. *et al.* (2011) Mimivirus shows dramatic genome reduction after intraamoebal culture. *Proc. Natl. Acad. Sci. U.S.A.* 108, 10296–10301
90. Lanza, V.F. *et al.* (2015) The plasmidome of Firmicutes: impact on the emergence and the spread of resistance to antimicrobials. *Microbiol. Spectr.* 3, PLAS-0039-2014
91. Remigi, P. *et al.* (2014) Transient hypermutagenesis accelerates the evolution of legume endosymbionts following horizontal gene transfer. *PLoS Biol.* 12, e1001942

92. Serôdio, J. *et al.* (2014) Photophysiology of kleptoplasts: photosynthetic use of light by chloroplasts living in animal cells. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.* 369, 20130242

93. Rauch, C. *et al.* (2015) Why it is time to look beyond algal genes in photosynthetic slugs. *Genome Biol. Evol.* 7, 2602–2607

94. Ereshefsky, M. and Pedroso, M. (2015) Rethinking evolutionary individuality. *Proc. Natl. Acad. Sci. U.S.A.* 112, 10126–10132

95. Martin, W.F. *et al.* (2015) Endosymbiotic theories for eukaryote origin. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.* 370, 20140330

96. Orphan, V.J. *et al.* (2001) Comparative analysis of methane-oxidizing archaea and sulfate-reducing bacteria in anoxic marine sediments. *Appl. Environ. Microbiol.* 67, 1922–1934

97. Ozuolmez, D. *et al.* (2015) Methanogenic archaea and sulfate reducing bacteria co-cultured on acetate: teamwork or coexistence? *Front. Microbiol.* 6, 492

98. Sun, M. *et al.* (2015) Microbial community analysis in rice paddy soils irrigated by acid mine drainage contaminated water. *Appl. Microbiol. Biotechnol.* 99, 2911–2922

99. Liu, Y.R. *et al.* (2014) Patterns of bacterial diversity along a long-term mercury-contaminated gradient in the paddy soils. *Microb. Ecol.* 68, 575–583

100. Lafontaine, D.L. *et al.* (1998) The box H + ACA snoRNAs carry Cbf5p, the putative rRNA pseudouridine synthase. *Genes Dev.* 12, 527–537

101. Zebarjadian, Y. *et al.* (1999) Point mutations in yeast CBF5 can abolish *in vivo* pseudouridylation of rRNA. *Mol. Cell. Biol.* 19, 7461–7472

102. Becker, H.F. *et al.* (1997) The yeast gene YNL292w encodes a pseudouridine synthase (Pus4) catalyzing the formation of psi55 in both mitochondrial and cytoplasmic tRNAs. *Nucleic Acids Res.* 25, 4493–4499

PLOS ONE

RESEARCH ARTICLE

BRIDES: A New Fast Algorithm and Software for Characterizing Evolving Similarity Networks Using Breakthroughs, Roadblocks, Impasses, Detours, Equals and Shortcuts

Etienne Lord[1,2], Margaux Le Cam[2], Éric Bapteste[3,4], Raphaël Méheust[3], Vladimir Makarenkov[1], François-Joseph Lapointe[2]*

1 Département d'informatique, Université du Québec à Montréal, Montréal, Québec, Canada, **2** Département de sciences biologiques, Université de Montréal, Montréal, Québec, Canada, **3** Sorbonne Universités, UPMC Université Paris 06, Institut de Biologie Paris-Seine, Paris, France, **4** CNRS, UMR7138, Institut de Biologie Paris-Seine, Paris, France

* francois-joseph.lapointe@umontreal.ca

Citation: Lord E, Le Cam M, Bapteste É, Méheust R, Makarenkov V, Lapointe F-J (2016) BRIDES: A New Fast Algorithm and Software for Characterizing Evolving Similarity Networks Using Breakthroughs, Roadblocks, Impasses, Detours, Equals and Shortcuts. PLoS ONE 11(8): e0161474. doi:10.1371/journal.pone.0161474

Editor: Xia Li, College of Bioinformatics Science and Technology, CHINA

Received: April 27, 2016

Accepted: August 6, 2016

Published: August 31, 2016

Data Availability Statement: The R and C++ source codes are available from the Github repository (https://github.com/etiennelord/BRIDES/) with a GPL version 3 license. The sample networks (Figs 1 and 2) are located in either the R or C++ source directories. The source code for the simulations (Fig 3) is located in the Simulation directory of the github repository. The genome similarity networks (Fig 4, Table 2) are available in the GenomeNetwork directory.

Abstract

Various types of genome and gene similarity networks along with their characteristics have been increasingly used for retracing different kinds of evolutionary and ecological relationships. Here, we present a new polynomial time algorithm and the corresponding software (BRIDES) to provide characterization of different types of paths existing in evolving (or augmented) similarity networks under the constraint that such paths contain at least one node that was not present in the original network. These different paths are denoted as **B**reakthroughs, **R**oadblocks, **I**mpasses, **D**etours, **E**qual paths, and **S**hortcuts. The analysis of their distribution can allow discriminating among different evolutionary hypotheses concerning genomes or genes at hand. Our approach is based on an original application of the popular shortest path Dijkstra's and Yen's algorithms. The C++ and R versions of the BRIDES program are freely available at: https://github.com/etiennelord/BRIDES.

Introduction

Network structures provide useful representations of interactions between the elements of complex systems [1, 2]. They can, for example, represent relationships between microbial communities in different environments [3] or between proteins in different bacteria and eukaryotes [4]. The abundance of the network elements (i.e. represented by nodes) as well as their interactions (i.e. represented by edges) often vary over time. Comparing evolving networks containing sets of attributes (or annotations) at their nodes is currently becoming central to different fields of biology, including ecology, evolution, cell biology and medicine [1, 5–7]. For example, genome similarity networks, where each node represents a genome and the edge weights correspond to the number of shared gene families between genomes, have been used to identify horizontal gene transfer events [8] and other reticulate phylogenetic relationships [3]. Genome similarity networks are typically constructed at different stringency thresholds (e.g. 50, 60, 70,

Funding: EL is supported by a Natural Sciences and Engineering Research Council (NSERC) scholarship. VM and FJL respectively hold NSERC discovery grants OGP0155251 and OGP249644. EB is funded by the European Research Council under the European Community's Seventh Framework Program FP7 (Grant Agreement n°615274).

Competing Interests: The authors have declared that no competing interests exist.

90, 99%) [6], or by using constantly increasing datasets, thus producing a range of inclusive networks. Such a strategy allows the detection of ancient evolutionary connections [1] or the verification of ecological distribution of taxa [9]. Furthermore, network analysis can be used with heterogeneous types of biological data, e.g. for linking protein structures to their functions [10,11].

Previous works in this field have focused on the use of conventional graph-theoretic measures describing the evolution of networks, such as the numbers of nearest neighbors or certain network motifs [12,13], or the distribution of shortest paths [9]. In this study, we present a number of novel features, which characterize evolving networks. All of them are based on the presence of additional nodes and edges in the augmented network. These *added nodes* and *edges* can be used to connect the original network nodes through different types of simple paths (i.e. loopless paths or paths that do not visit the same nodes twice) [14,15]. Precisely, we will describe a new polynomial time algorithm for estimating the number of **B**reakthroughs, **R**oadblocks, **I**mpasses, **D**etours, **E**qual paths and **S**hortcuts (BRIDES) in evolving networks. Moreover, we have developed C++ and R functions implementing the new algorithm. We will also present the results of our simulation study comparing the performances of four different versions of our algorithm as well as the application of the most successful version of BRIDES to real genome similarity networks. It is worth noting that, contrary to previous work, we were neither interested in counting the number of colored motifs in networks [16], nor in determining their types [12, 13].

Materials and Methods

Description of the BRIDES algorithm

This section describes the problem we address here from a mathematical point of view and presents the most important computational details of our algorithm. The main questions that we try to answer in this paper are the following:

1. Is there a simple path between two given nodes i and j (i.e. original network nodes in our study) that contains at least one node from a specific set of nodes (i.e. set of added nodes in our study)?
2. If such a simple path exists, is it the shortest path between i and j?

Since the number of simple paths between two given nodes in a graph can be exponential in the size of the graph, visiting and counting all of them is a problem belonging to #P [14, 17]. On the other hand, the number of simple shortest paths between two nodes of an undirected graph can also be exponential in the number of the graph nodes. Furthermore, the problem of finding a simple shortest path including a set of *must-include* nodes is NP-hard [18, 19]. Therefore, effective heuristic algorithms should be applied to answer questions (a) and (b), especially when large genetic or genomic similarity networks are considered. Vardhan and colleagues [19] proposed a fast heuristic algorithm to compute a simple path that contains a given ordered set of must-include nodes. However, this problem is slightly different from our problem, since our main objective is to find a shortest simple path including *at least one node* from a set of specified nodes. Li and colleagues [20] tried to address the latter problem, presenting a fast heuristic approach based on the principle of optimality of dynamic programming. However, their elegant algorithm can be applied only to graphs with a specific–series/parallel–topology [20].

The BRIDES algorithm takes as input two networks: (1) network X with an original set of nodes N_X and edges E_X (Fig 1A) and (2) network Y with an augmented set of nodes N_Y and edges E_Y (Fig 1B). All nodes of network X should also be present in network Y (i.e. $N_X \subset N_Y$).

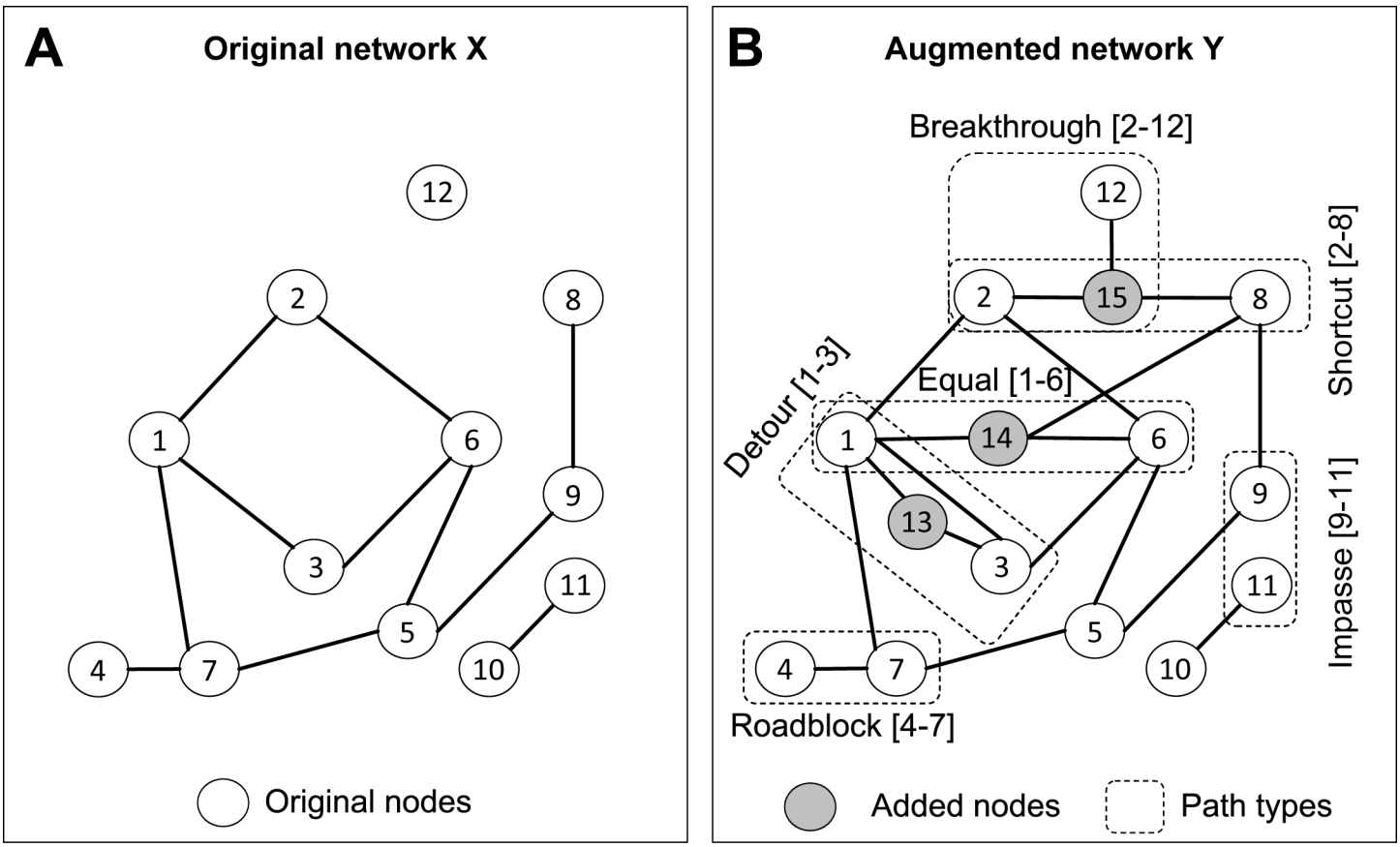


Fig 1. Examples of the BRIDES (Breakthrough, Roadblock, Impasse, Detour, Equal and Shortcut) paths in evolving networks. Panel (A) presents an original network *X* with 12 nodes. Panel (B) presents an augmented network *Y* with 15 nodes (including 12 original and 3 added nodes, which are colored in *grey*). Six different types of paths are shown in the augmented network *Y*.

doi:10.1371/journal.pone.0161474.g001

However, it is not required that $E_X \subset E_Y$. We first compute the shortest paths between the pairs of nodes in *X*, and then reassess their length in *Y*, by forcing these paths to include at least one added node (i.e. a node present in *Y*, but not in *X*). Our heuristic relies on a repeated application of Dijkstra's algorithm [21] to evaluate the impact of added nodes to the length of the shortest paths between the original nodes (see Algorithm 1).

Now we can define six distinct types of paths, related to the existence of added nodes in *Y*, which can be used to characterize complex relationships in evolving networks (Fig 1):

Breakthrough: a path that is impossible in network *X* but is possible in network *Y* (e.g. path between nodes 2 and 12, passing by added node 15 in Fig 1B);

Roadblock: a path that is possible in network *X* but is impossible in network *Y* (e.g. a simple path between nodes 4 and 7 that passes by an added node in *Y* is impossible, see Fig 1B);

Impasse: a path that is impossible in both networks, *X* and *Y* (e.g. there are no possible paths between nodes 9 and 11 in Fig 1A and 1B);

Detour: a path that is shorter in network *X* than in network *Y* (e.g. path between nodes 1 and 3 in Fig 1A and 1B);

Equal: a path that has the same length in networks *X* and *Y* (e.g. path between nodes 1 and 6, assuming that all edge lengths in *X* and *Y* are equal, Fig 1A and 1B);

Shortcut: a path that is longer in network *X* than in network *Y* (e.g. path between nodes 2 and 8 in Fig 1A and 1B).

Fig 2 provides an example of computation of the BRIDES statistics in evolving networks. We can see that the addition of new nodes and edges to an evolving network can substantially change the distribution of the six types of paths defined in our study. The four heuristic strategies tested in our simulations, called BRIDES (the original strategy), BRIDES_Y, BRIDES_YC and BRIDES_EC, are presented in details in Algorithm 1 below:

Algorithm 1

Given an original undirected network $X = (E_X, N_X)$ and its augmented undirected network $Y = (E_Y, N_Y)$, i.e. network such that $N_X \subset N_Y$, this algorithm calculates the number of Breakthroughs, Roadblocks, Impasses, Detours, Equal paths and Shortcuts (BRIDES) to characterize the evolution of X into Y.

BRIDES

Step 1. Compute the length of the shortest path between all pairs of nodes in network X. Find at most *MaxPathNumber* of simple shortest paths between pairs of original nodes (i,j), (i.e. nodes such that $i \in N_X$ and $j \in N_X$) in network Y, using Dijkstra's algorithm. Store in the list P_{ij} the set of simple shortest paths corresponding to a pair of original nodes (i,j) in network Y.

Step 2. For all pairs of original nodes (i,j), create a list L_{ij} of added nodes k ($k \in N_Y$, $k \notin N_X$) ordered with respect to the closeness of i and j to k. Calculate the distances $d(i,k)$ and $d(j,k)$ in Y using Dijkstra and store at most *MaxPathNumber* of simple shortest paths in P_{ik} and P_{jk}, respectively. Order the list of added nodes L_{ij} according to either the minimum of $\text{Max}(d(i,k), d(j,k))$ (Strategy 1 that provided better overall results in our simulations; the results of this strategy will be presented in the next section) or the minimum of $(d(i,k) + d(j,k))$ (Strategy 2). In order to speed up the algorithm, we can reduce the size of L_{ij} by using the input parameters: *MaxDistance* (the maximum allowed distance from i or from j to k) and/or *MaxNode* (the maximum number of added nodes, k, in this list).

Set the first pair of original nodes (i,j) as the *current pair of nodes*.

Step 3. Do, for the current pair of original nodes (i,j):

If there exists a simple path in P_{ij} that includes at least one added node, update the BRIDES statistics (see Table 1) with the results obtained for the current pair of nodes (i,j), set the next pair (i,j) as the current pair of original nodes and go to the beginning of Step 3; otherwise, go to Step 4.

Step 4. At this point, we have determined that the current pair of original nodes (i,j) is not associated with a Breakthrough, Impasse or Shortcut, and we can now determine whether it should be associated with a Detour, Equal path or Roadblock.

Do, for each node k of the ordered list L_{ij}, starting from the first element of L_{ij}:

Step 4.1. If the *concatenation* of paths$[i,k]$ from P_{ik} and$[j,k]$ from P_{jk} is a simple path, set $d(i,j) = d(i,k) + d(j,k)$ and update the BRIDES statistics (see Table 1) with the results obtained for the pair of nodes (i,j), set the next pair (i,j) as the current pair of original nodes and go to Step 3; otherwise, go to Step 4.2.

Step 4.2. Since there are repeating nodes, except k, on the paths$[i,k]$ and$[j,k]$, temporarily remove them from network Y and recalculate: (1) the shortest path from j to k in the reduced network Y using Dijkstra, storing the result in P_{jk}', and (2) the shortest path from i to k in the reduced network Y using Dijkstra, storing the result in P_{ik}'. Repeat

Step 4.1 with the shortest of concatenations of two paths stored: (1) in P_{ik} and P_{jk}' and (2) in P_{jk} and P_{ik}'; if for both P_{jk}' and P_{ik}' does not return a simple shortest path because no such path exists in the reduced network Y, consider the next node k of L_{ij} in Step 4.1 or go to Step 5 if all element of L_{ij} have been already examined.

Step 5. Classify the path associated with the current pair of nodes (i, j) as a Roadblock.

Step 6. If all the pairs (i,j) have been already examined, print the BRIDES statistics; otherwise, set the next pair (i,j) as the current pair of original nodes and go to Step 3.

Heuristic BRIDES_Y (*BRIDES using Yen's algorithm*)

Replace Steps 3 and 4 above by the following steps:

Step 3'. Do, for the current pair of original nodes (i,j):

Compute the ordered list PY_{ij} of *MaxPathNumber* shortest paths between i and j using Yen's k-shortest path algorithm[22] .

Step 4'. Do, for each path p of the ordered list PY_{ij},:

If p contains at least one added node, update the BRIDES statistics (see Table 1) with the results obtained for the current pair of original nodes (i,j), set the next pair (i,j) as the current pair of original nodes and go to Step 3.

Heuristic BRIDES_YC (*BRIDES using Yen's algorithm and Concatenation of paths*)

Replace Steps 2 and 4 above by the following steps:

Step 2'. For all pairs of original nodes (i,j), create a list L_{ij} of added nodes k ($k \in N_Y$, $k \notin N_X$) ordered with respect to the closeness of i and j to k. Calculate the distances $d(i,k)$ and $d(j,k)$ in Y using Dijkstra. Order the list of added nodes L_{ij} according to the minimum of Max($d(i,k)$, $d(j,k)$) (Strategy 1) or the minimum of ($d(i,k) + d(j,k)$) (Strategy 2). Using Yen's algorithm compute and store at most *MaxPathNumber* of paths in P_{ik} and P_{jk}, respectively. In order to speed up the algorithm, we can reduce the size of L_{ij} by using the input parameters: *MaxDistance* and/or *MaxNode*.

Set the first pair of original nodes (i,j) as the current pair of nodes.

Step 4'. Do, for each node k of the ordered list L_{ij}, starting from the first element of L_{ij}:

Step 4.1. If the *concatenation* of paths[i,k] from P_{ik} and[j,k] from P_{jk} is a simple path, set $d(i,j) = d(i,k) + d(j,k)$ and update the BRIDES statistics (see Table 1) with the results obtained for the pair of nodes (i, j), set the next pair (i,j) as the current pair of original nodes and go to Step 3.

Go to Step 5 if all element of L_{ij} have been already examined.

BRIDES_EC (*BRIDES algorithm based on an Exhaustive Concatenation approach*)

The difference with the original BRIDES algorithm is in Step 4, where we examine all the nodes k of the ordered list L_{ij}, even though a simple path has been found in Step 4.1.

The default values of the parameters *MaxPathNumber*, *MaxDistance* and *MaxNode* in our program are all equal to 100. These default values were also used in our simulation study (see the next section). It is worth noting that in unweighted graphs, *MaxDistance* represents the upper bound of the number of edges on the path between an original and an added node. Obviously, in weighted graphs, this parameter should be specified by the user. The time complexity of Steps 1 and 2 of our original BRIDES algorithm is $O(|N_X| \times MaxPathNumber \times (|E_Y| + |N_Y|$

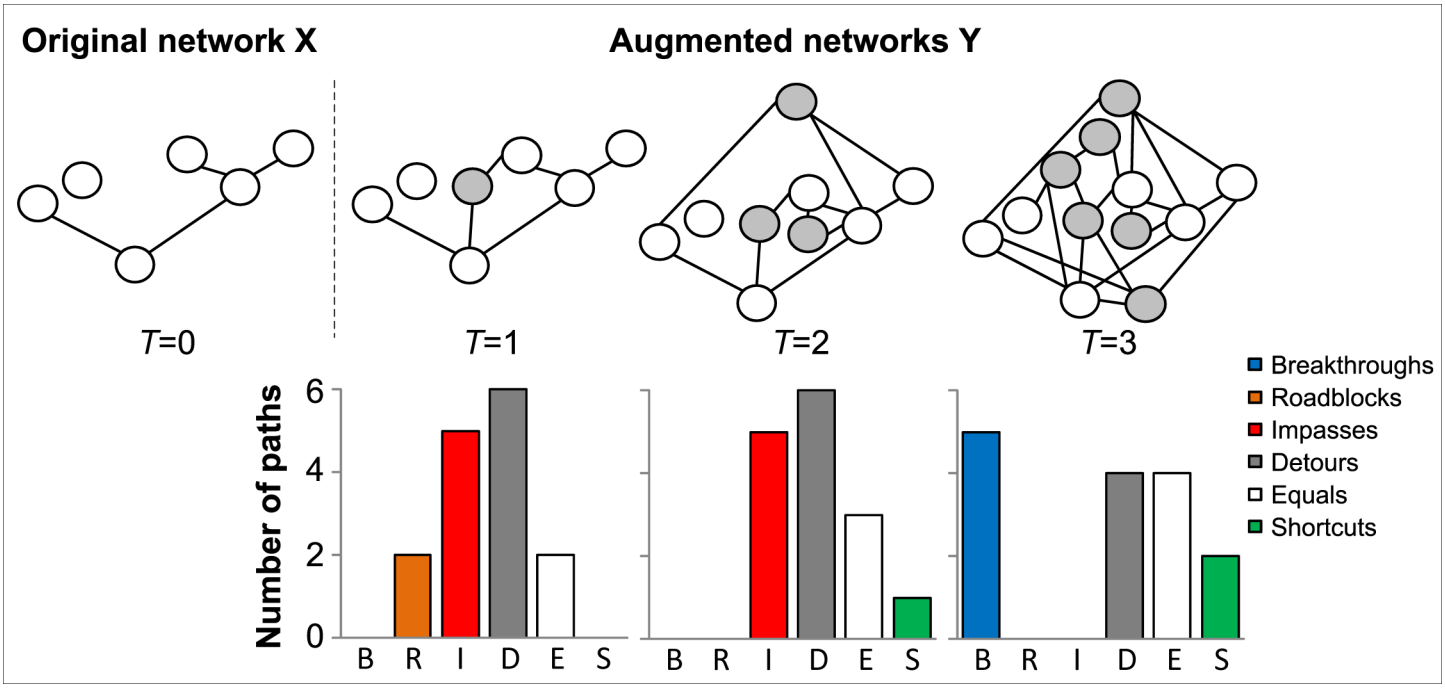


Fig 2. Computation of the BRIDES statistics in evolving networks. Addition of the new nodes (colored in *grey*) and edges to an evolving network changes the distribution of different types of network pathways as time (*T*) progresses. The letters B, R, I, D, E and S at the bottom of the chart stand respectively for Breakthroughs, Roadblocks, Impasses, Detours, Equal paths and Shortcuts.

doi:10.1371/journal.pone.0161474.g002

Table 1. Possible BRIDES outcomes depending on the path type identified by the algorithm.

Simple path between i to j in X	*Simple path between i to j in Y*	*BRIDES statistic*
Impossible	Possible	Breakthrough
Possible	Impossible	Roadblock
Impossible	Impossible	Impasse
Shorter distance	Longer distance	Detour
Equal distance	Equal distance	Equal
Longer distance	Shorter distance	Shortcut

doi:10.1371/journal.pone.0161474.t001

$\log(|N_Y|))$, using asymptotically the fastest known single-source shortest-path version of Dijkstra's algorithm, where $|N_X|$ and $|N_Y|$ are the numbers of nodes in networks X and Y, respectively, and $|E_Y|$ is the number of edges in network Y. The time complexity of Step 4 is $O(|N_X|^2 \times MaxNodes \times (|E_Y| + |N_Y|\log(|N_Y|))$, in the worst case. However, in practice, the runtime of this step is much lower because we rarely execute the internal loop of Step 4 all the *MaxNodes* times. This leads to the total time complexity of BRIDES equal to $O(|N_X| \times (MaxPathNumber + |N_X| \times MaxNodes) \times (|E_Y| + |N_Y|\log(|N_Y|))$. The presented BRIDES algorithm can be applied to analyze undirected graphs with non-negative edge lengths. When negative edge lengths exist in either network X or network Y, the improved version of Bellman–Ford's algorithm [23] could be applied instead of Dijkstra.

We created an R function implementing the BRIDES algorithm using the graph manipulation tools available in the *igraph* package [24]. The R version of BRIDES can be applied for the analysis of small networks (<1,000 nodes). For larger networks (>1,000 nodes), we recommend using the

C++ version of our program, which includes all the four heuristic algorithms, BRIDES, BRIDES_Y, BRIDES_YC and BRIDES_EC, discussed in this paper. Moreover, a parallel OpenMP [25] version of the C++ program is also available (see: https://github.com/etiennelord/BRIDES).

Results

Simulation study

To test the BRIDES algorithm, we carried out a simulation study using three popular network models implemented in the *igraph* package (version 1.0.0) [24]. Precisely, the Erdős–Rényi [26], Barabási–Albert [27] and Watts–Strogatz [28] random graph generation models were considered. The Barabási–Albert model is a well-known approach for generating scale-free (or power-law) networks, while the Erdős–Rényi and Watts–Strogatz models are among the most popular generation models for random graphs that do not exhibit a scale-free degree distribution.

Using each of these three models, we generated 1000 random original networks X with 100 nodes, and then added to them 5, 25, 50 or 100 additional nodes to create augmented networks Y. The simulations were performed using the C++ version of our program executed on a PC computer equipped with an Intel i7-3770 CPU (3.40GHz) and 8Gb of RAM. The four competing heuristic strategies, BRIDES, BRIDES_Y, BRIDES_YC and BRIDES_EC, presented in Algorithm 1 were tested in our simulation study. The accuracy of the competing approaches (see Fig 3A) was calculated as the percentage of correctly labeled path types (i.e. the percentage of true positive Breakthroughs, Roadblocks, Impasses, Detours, Equal paths and Shortcuts) provided by each heuristic. The identification of the correct (i.e. reference) path types was carried out using a brute force procedure based on a depth-first search (DSF) algorithm. Along with the average accuracy, calculated over all generated graphs, we also measured the average runtime (in seconds; see Fig 3B and S1 Table) of each of the four heuristics under comparison.

The results of our simulations suggest that the original BRIDES strategy along with the exhaustive concatenation procedure, BRIDES_EC, were able to provide the correct classification of paths, regardless of the number (i.e. also percentage, in this case) of added nodes (Fig 3A). The two heuristics based on Yen's k-shortest path algorithm (i.e. BRIDES_Y and BRIDES_YC), returned the correct classification of paths in 67% and 90% of cases, respectively, when the number of added nodes was 5. However, the results obtained with BRIDES_Y and BRIDES_YC improved with an increase of the number of added nodes, reaching the accuracy level of 100% for 100 added nodes.

The original BRIDES algorithm was the fastest among the four compared heuristics regardless the number of added nodes (Fig 3B). The worst performance, in terms of running time, was shown by BRIDES_Y, for small numbers of added nodes. This can be explained as follows. On one hand, we restrict our search in Yen's algorithm to only 100 shortest paths. On the other hand, when the proportion of added nodes in Y is very small, all of them can be located far away from both source and destination nodes given as input to Yen's algorithm. Thus, these rare added nodes can never be included in the set of 100 shortest paths returned by BRIDES_Y. This leads to an important decrease in accuracy (Fig 3A) and increase in computational time (Fig 3B). The heuristic BRIDES_YC, which uses both Yen's algorithm and concatenation of paths, performs better in this case since the concatenated paths are guaranteed to contain at least one added node.

It is worth noting that the length of a Detour path identified by the BRIDES algorithm can be longer than the length of the shortest possible Detour existing in the network, but in this work, we are only interested in estimating the distribution of path types, and not in assessing the exact path lengths.

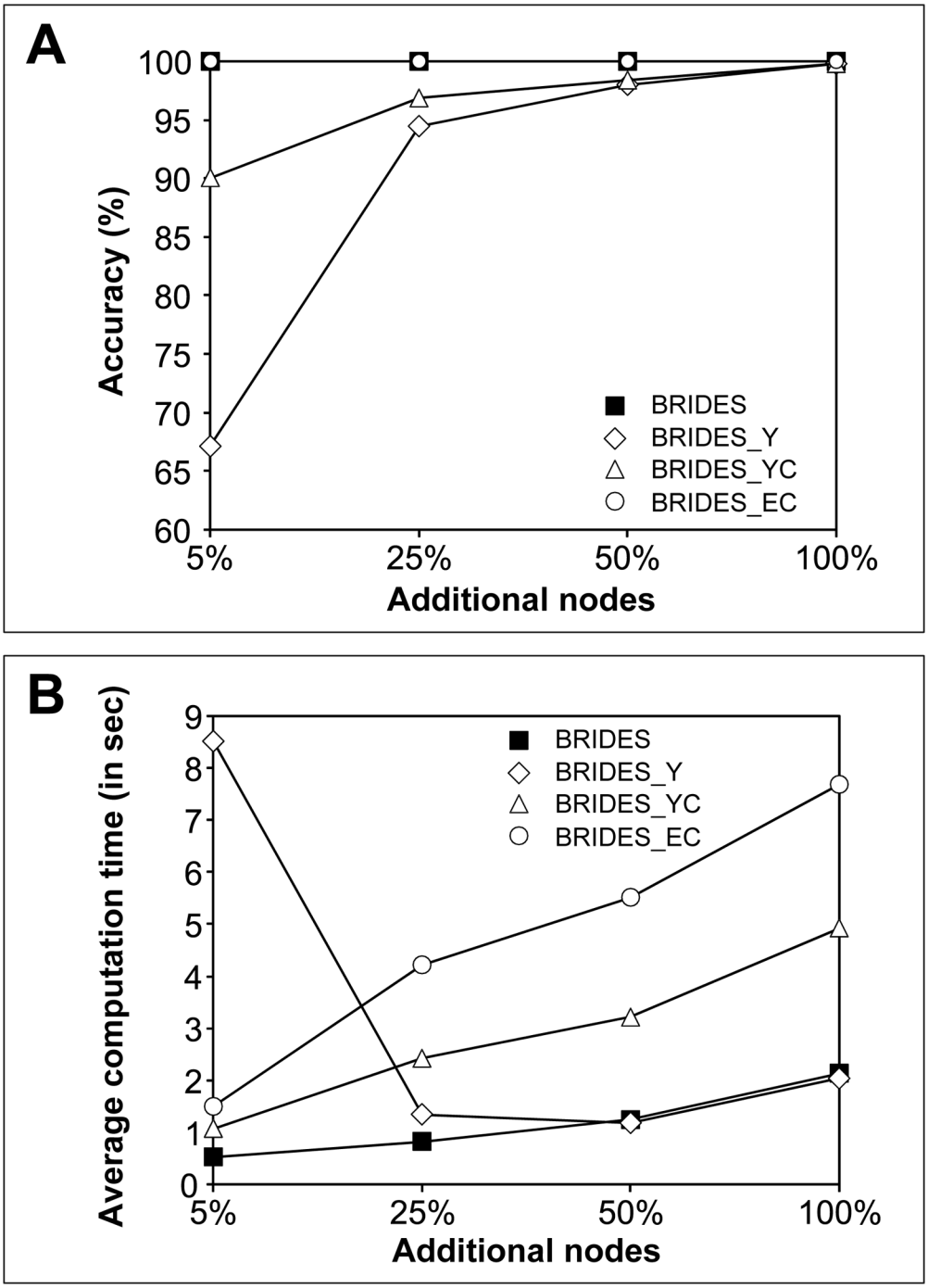


Fig 3. The average accuracy (A) and computational time (B) obtained for the four heuristic strategies, BRIDES, BRIDES_Y, BRIDES_YC and BRIDES_EC, presented in our study. The simulations were carried out with original networks *X* containing 100 nodes and generated using the Erdős–Rényi, Barabási–Albert and Watts–Strogatz random graph generation models. The augmented networks *Y* contained 5, 25, 50 or 100 added nodes and all original nodes of *X*. 1000 graphs were generated for each parameter configuration.

doi:10.1371/journal.pone.0161474.g003

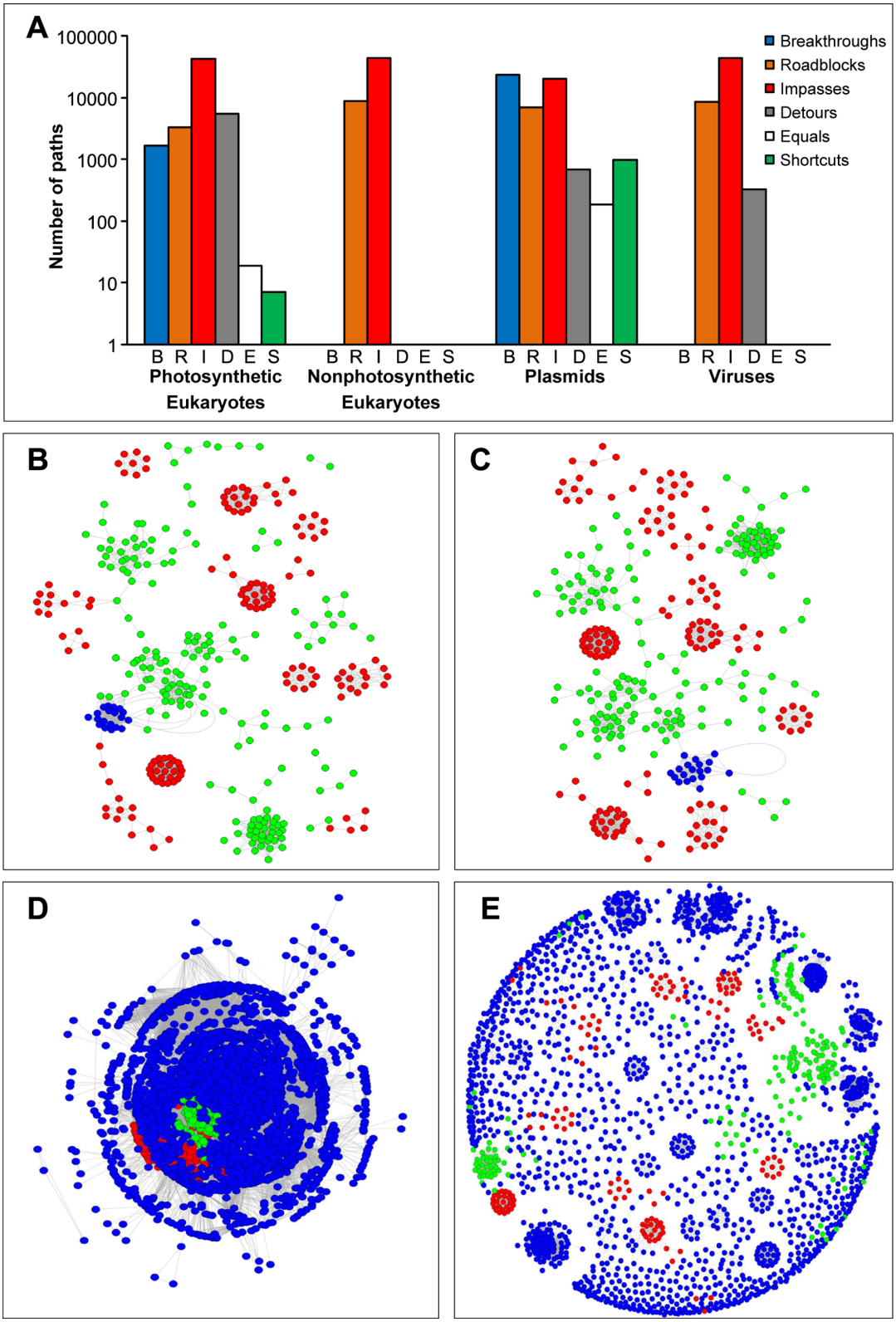


Fig 4. BRIDES statistics for real genome similarity networks at 90% similarity threshold. The BRIDES statistics (A) computed for the original network *X* comprising archaea (in *red*) and bacteria (in *green*), 326 species in total, and the four augmented similarity networks *Y* (the added nodes are in *blue*), including: (B) photosynthetic eukaryotes, (C) nonphotosynthetic eukaryotes, (D) plasmids and (E) viruses.

doi:10.1371/journal.pone.0161474.g004

Table 2. General network and BRIDES statistics for the four real genome similarity networks presented in Fig 4.

	General network statistics					*BRIDES statistics*[a]					
Added species	**Total of nodes**	**Total of edges**	**Average degree**	**Average path length**	**Clustering coefficient**	**B**	**R**	**I**	**D**	**E**	**S**
Eukaryotes photosynt.	345	2,014	11.68	5.97	0.867	1,652	3,329	42,418	5,550	19	7
Eukaryotes nonphotosynt.	345	1,845	10.70	5.29	0.865	0	8,905	44,070	0	0	0
Plasmids	3,848	187,848	97.63	2.90	0.559	23,618	7,057	20,452	689	185	974
Viruses	1,984	12,054	12.15	5.23	0.801	0	8,577	44,070	328	0	0

[a] Computation were carried out using the following input parameters: *MaxDistance* = 100, *MaxNode* = 100 and *MaxPathNumber* = 100.

doi:10.1371/journal.pone.0161474.t002

Application of BRIDES to real biological data

To evaluate the performance of the BRIDES algorithm and calculate the related statistics (Fig 4A) for real networks, we generated four genome similarity networks using a set of 2,094,593 nucleotide sequences of archaea, bacteria, photosynthetic and nonphotosynthetic eukaryotes, plasmids and viruses. Similarity between the nucleotide sequences was determined using *BLAST* [29] with a minimum *e*-value of 10*e*-5. Individual genomes were connected in both original and augmented networks if at least one of their genes shared a homologous sequence (>70% cover, with the 90% similarity threshold). A total of 326 prokaryotes were selected to form the original network *X*, and then complemented with either photosynthetic eukaryotes (Fig 4B), or nonphotosynthetic eukaryotes (Fig 4C), or plasmids (Fig 4D), or viruses (Fig 4E) in the augmented network *Y*.

The BRIDES statistics provided by our algorithm for the four genome similarity networks exhibited different distribution profiles (Fig 4A). Even when the genome networks displayed similar clustering coefficients and comparable average path lengths (Table 2), the BRIDES statistics could be quite different (Fig 4A). For example, the addition of photosynthetic eukaryotes resulted in a large number of Detours, Shortcuts and Equal paths (Fig 4B), whereas the addition of nonphotosynthetic eukaryotes resulted in the complete disappearance of such path types (Fig 4C). This difference can be explained by the fact that nuclear genomes of photosynthetic eukaryotes host gene families from cyanobacterial origin, as a result of gene transfer from their chloroplastic endosymbionts, which are typically absent in nonphotosynthetic eukaryotes. Moreover, the addition of 3552 plasmids (Fig 4D) to the original network *X* led to a similar BRIDES profile as in the case of photosynthetic eukaryotes, while favouring the emergence of Shortcuts. On the contrary, the addition of viruses (Fig 4E) did not introduce any Shortcuts into the augmented similarity network, but led to the increase of the numbers of Roadblocks. The latter result is consistent with findings of Halary and colleagues [8], who identified plasmids as central genetic carriers across prokaryotic genomes, whereas viruses were found to have restricted host ranges for infecting distantly related taxa.

Conclusion

In this paper, we introduced a new fast algorithm and associated software for characterizing different types of paths existing in evolving similarity networks. In particular, our algorithm calculates the number of Breakthroughs, Roadblocks, Impasses, Detours, Equal paths and Shortcuts (BRIDES), which can be present in these networks. Our program, implemented in the C++ and R programming languages, includes four heuristic algorithms for calculating the BRIDES statistics discussed and compared in our study (see Algorithm 1). These statistics can be viewed as a new tool for the characterization and comparison of evolving genome and gene

similarity networks, transcriptional networks [30] or interactome networks [31]. The analysis and comparison of evolving networks can be carried out for different network stringency thresholds. Dijkstra's algorithm used in our method makes it suitable for the analysis of both weighted and unweighted types of networks. Note that our BRIDES heuristic presented here in the case of undirected networks can be easily adapted to the case of directed networks. In the future, it would be interesting to see whether the Uniform Cost Search [32] or A* [33,34] algorithms can be used as an alternative of Dijkstra in order to accelerate the computation of the BRIDES statistics within our method.

Dataset and Source Files

The R and C++ source codes are available from the Github repository (https://github.com/etiennelord/BRIDES/) with a GPL version 3 license. The sample networks (Figs 1 and 2) are located in either the R or C++ source directories. The source code for the simulations (Fig 3) is located in the Simulation directory of the github repository. The genome similarity networks (Fig 4, Table 2) are available in the GenomeNetwork directory.

Supporting Information

S1 Table. Average computational time in seconds (s) obtained for the four heuristics and for different network models. The original network *X* contained 100 nodes and the augmented networks *Y* contained 5, 25, 50 or 100 added nodes. For each model, 1000 networks were created, and 100 path were randomly selected for evaluation. The reported values are average time (in seconds) for the evaluation of each path.
(PDF)

Acknowledgments

We thank Dr. Philippe Lopez for critical discussion on genome similarity networks.

Author Contributions

Conceptualization: EL EB FJL.

Data curation: EL RM.

Formal analysis: EL FJL.

Funding acquisition: EB VM FJL.

Investigation: EL MLC RM.

Methodology: EL MLC EB VM FJL.

Project administration: FJL.

Software: EL MLC RM VM.

Supervision: FJL.

Validation: EL EB FJL.

Visualization: EL.

Writing – original draft: EL MLC FJL.

Writing – review & editing: EL EB VM FJL.

References

1. Cheng S, Karkar S, Bapteste E, Yee N, Falkowski P, Bhattacharya D. Sequence similarity network reveals the imprints of major diversification events in the evolution of microbial life. Front Ecol Evol. 2014; 2: 1–13.
2. Orsini C, Dankulov MM, Colomer-de-Simón P, Jamakovic A, Mahadevan P, Vahdat A, et al. Quantifying randomness in real networks. Nat Commun. 2015; 6: 8627. doi: 10.1038/ncomms9627 PMID: 26482121
3. Lynch MD, Bartram AK, Neufeld JD. Targeted recovery of novel phylogenetic diversity from next-generation sequence data. ISME J. 2012; 6: 2067–2077. doi: 10.1038/ismej.2012.50 PMID: 22791239
4. Wuchty S, Uetz P. Protein-protein interaction networks of *E. coli* and *S. cerevisiae* are similar. Sci Rep. 2014; 4: 7187. doi: 10.1038/srep07187 PMID: 25431098
5. Montoya JM, Pimm SL, Solé RV. Ecological networks and their fragility. Nature. 2006; 442: 259–264. PMID: 16855581
6. Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. Genome Res. 2011; 21: 599–609. doi: 10.1101/gr.115592.110 PMID: 21270172
7. Vidal M, Cusick ME, Barabási AL. Interactome networks and human disease. Cell. 2011; 144: 986–998. doi: 10.1016/j.cell.2011.02.016 PMID: 21414488
8. Halary S, Leigh JW, Cheaib B, Lopez P, Bapteste E. Network analyses structure genetic diversity in independent genetic worlds. Proc Natl Acad Sci USA. 2010; 107: 127–132. doi: 10.1073/pnas.0908978107 PMID: 20007769
9. Forster D, Bittner L, Karkar S, Dunthorn M, Romac S, Audic S, et al. Testing ecological theories with sequence similarity networks: marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms. BMC Biol. 2015; 13: 16. doi: 10.1186/s12915-015-0125-5 PMID: 25762112
10. Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. PloS ONE. 2009; 4: e4345. doi: 10.1371/journal.pone.0004345 PMID: 19190775
11. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014; 11: 333–337. doi: 10.1038/nmeth.2810 PMID: 24464287
12. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. Science. 2002; 298: 824–827. PMID: 12399590
13. Bapteste E, van Iersel L, Janke A, Kelchner S, Kelk S, McInerney JO, et al. Networks: expanding evolutionary thinking. Trends Genet. 2013; 29: 439–441. doi: 10.1016/j.tig.2013.05.007 PMID: 23764187
14. Roberts B. Kroese DP. Estimating the number of s-t paths in a graph. J Graph Algorithms Appl. 2007; 11: 195–214.
15. Martins EQ. Pascoal MM. A new implementation of Yen's ranking loopless paths algorithm. 4OR-Q J Oper Res. 2003; 1: 121–133.
16. Schbath S, Lacroix V, Sagot MF. Assessing the exceptionality of coloured motifs in networks. EURASIP J Bioinform Syst Biol. 2009; 1; 616234.
17. Valiant LG. The complexity of enumeration and reliability problems. SIAM J Comput. 1979; 8: 410–421.
18. Burkard RE, Deineko VG, van Dal R, van der Veen JA, Woeginger GJ. Well-solvable special cases of the traveling salesman problem: a survey. SIAM Rev. 1998; 40: 496–546.
19. Vardhan H, Billenahalli S, Huang W, Razo M, Sivasankaran A, Tang L, et al. Finding a simple path with multiple must-include nodes. IEEE MASCOTS. 2009;1–3.
20. Li WJ, Tsao HSJ, Ulular O. The shortest path with at most/nodes in each of the series/parallel clusters. Networks. 1995; 26: 263–271.
21. Dijkstra EW. A note on two problems in connexion with graphs. Numer Math. 1959; 1: 269–271.
22. Yen JY. An algorithm for finding shortest routes from all source nodes to a given destination in general networks. Q Appl Math. 1970; 27: 526–530.
23. Goldberg AV, Radzik TA. Heuristic improvement of the Bellman-Ford algorithm. Appl Math Lett.1993; 6: 3–6.
24. Csardi G, Nepusz T. The igraph software package for complex network research. Inter J Complex Syst. 2006; 1695: 1–9.
25. OpenMP Forum. Openmp: A proposed industry standard api for shared memory programming. Technical report, Oct. 1997

26. Erdős P, Rényi A. On random graphs. Pub Math Debrecen. 1959; 6: 290–297.
27. Barabási AL, Albert R. Emergence of scaling in random networks. Science. 1999; 286: 509–512. PMID: 10521342
28. Watts DJ, Strogatz SH. Collective dynamics of 'small world' networks. Nature 1998; 393: 440–442. PMID: 9623998
29. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215: 403–410. PMID: 2231712
30. Aylwarda FO, Eppleya JM, Smithb JM, Chavezb FP, Scholinb CA, DeLong EF. Microbial community transcriptional networks are conserved in three domains at ocean basin scales. Proc Natl Acad Sci USA. 2015; 112: 5443–5448. doi: 10.1073/pnas.1502883112 PMID: 25775583
31. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, et al. Uncovering disease-disease relationships through the incomplete interactome. Science. 2015; 347: 1257601. doi: 10.1126/science.1257601 PMID: 25700523
32. Verwer BJH, Verbeek PW, Dekker ST. An efficient uniform cost algorithm applied to distance transforms. IEEE Trans Pattern Anal Mach Intell. 1989; 4: 425–429.
33. Hart PE, Nilsson NJ, Raphael B. A formal basis for the heuristic determination of minimum cost paths. IEEE Trans Syst Sci Cybernetics. 1968; 4: 100–107.
34. Zeng W, Church RL. Finding shortest paths on real road networks: the case for A*. Int J Geogr Inf Sci. 2009; 23: 531–543.

A. Acquisition exogène de nouveaux gènes

Le transfert horizontal de gènes (HGT) est un processus où un organisme va acquérir du matériel génétique provenant d'un autre organisme sans en être son descendant direct. Les HGT sont essentiels dans l'évolution des Bactéries et des Archaea [105], c'est le moyen le plus courant pour acquérir de nouveaux gènes et de nouveaux traits [106,107].

1. Transfert horizontal de gènes chez les procaryotes (Articles IV et V)

Les HGT façonnent les génomes procaryotes. Par exemple, les gènes présents en multiples copies au sein des génomes sont essentiellement dus à des HGT et non à des duplications de gènes comme c'est majoritairement le cas chez les eucaryotes [108]. Une des conséquences directes est que les HGT brouillent le signal phylogénétique des gènes : alors que chez les eucaryotes les gènes possèdent quelques origines différentes, chez les procaryotes les gènes d'un même génome bactérien peuvent provenir d'une myriade de donneurs différents, vivants ou éteints. Par conséquent, reconstruire la généalogie des lignées procaryotes est compliqué et le modèle arborescent est clairement insuffisant pour rendre compte de leurs histoires [109,110]. Une autre conséquence de cette évolution réticulée est que le contenu en gènes des génomes d'une même espèce bactérienne montre de fortes variations reflétant le processus continu d'acquisition et de perte de gènes. Une comparaison du contenu en gènes de 61 génomes *Escherichia coli* montre que seulement 6% des familles de gènes présentes chez *E. coli* sont présentes chez tous les génomes d'*E.coli* [111]. Cette partie conservée du génome est appelée "core" alors que la partie variable est appelée accessoire; enfin l'addition du génome "core" et du génome accessoire correspond au pan-génome [105,112].

Les HGT surviennent plus fréquemment entre individus proches phylogénétiquement qu'entre individus distants [113]. Cette différence peut s'expliquer par le fait que les organismes proches partagent une architecture génomique similaire comme par exemple un contenu en GC proche ou les mêmes biais d'utilisation des codons [113]. Ils ont aussi tendance à être infectés par les mêmes éléments génétiques mobiles qui sont des vecteurs de propagation des gènes dans les génomes [114]. Ceci est vrai aussi pour la conjugaison qui est plus commune entre individus proches phylogénétiquement. Par conséquent, le répertoire en gènes d'une espèce bactérienne, le pan-génome, est une source de diversité et d'adaptabilité pour les organismes de ce groupe puisqu'ils peuvent ainsi partager

les gènes du répertoire [113]. Bien que moins fréquents, les HGT existent entre des organismes distants phylogénétiquement et particulièrement entre organismes partageant le même environnement [115].

Les gènes impliqués dans des fonctions métaboliques sont sur-représentés parmi ceux impliqués dans des HGT, soulignant ainsi une plus grande facilité à être transférés et/ou à être fixés chez les individus receveurs [116]. *A contrario*, les gènes impliqués dans des fonctions informationnelles (c'est-à-dire ceux impliqués dans la transcription, la traduction, la réplication et dans des fonctions semblables) sont plus difficilement transférés et/ou fixés [115]. Les gènes informationnels codent souvent pour des protéines impliquées dans de grosses et complexes machineries cellulaires et donc interagissent avec de nombreuses autres protéines ce qui expliquent leurs difficultés pour se fixer. Cette hypothèse est appelée hypothèse de complexité [117]. Récemment une étude a suggéré que ce n'est pas la fonction des gènes qui limite les transferts mais plutôt leurs fortes connectivités dans les réseaux d'interactions protéines protéines (PPI pour Protein Protein Interaction) qui en est responsable [118].

L'article suivant est issu d'une collaboration avec l'équipe de Yan Boucher. L'étude s'est intéressée aux flux génétiques entre deux groupes de bactéries phylogénétiquement proches et présents dans le même environnement. Bien que les deux groupes de bactéries puissent être considérés comme deux espèces distinctes d'après le pourcentage moyen d'identité nucléotidique, les HGT inter-groupes sont fréquents.

The Dynamics of Genetic Interactions between *Vibrio metoecus* and *Vibrio cholerae*, Two Close Relatives Co-Occurring in the Environment

Fabini D. Orata[1], Paul C. Kirchberger[1], Raphaël Méheust[2], E. Jed Barlow[3], Cheryl L. Tarr[4], and Yan Boucher[1,*]

[1]Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada

[2]Unité Mixte de Recherche 7138, Evolution Paris-Seine, Institut de Biologie Paris-Seine, Université Pierre et Marie Curie, Paris, France

[3]Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada

[4]Enteric Diseases Laboratory Branch, Division of Foodborne, Waterborne, and Environmental Diseases, National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, GA

*Corresponding author: E-mail: yboucher@ualberta.ca.

Accepted: October 2, 2015

Data deposition: The whole-genome sequences generated in this study have been deposited in the DNA Data Bank of Japan (DDBJ), European Molecular Biology Laboratory (EMBL), and GenBank databases under the BioProject accession PRJNA281423. The individual genome accession numbers are listed in supplementary table S1, Supplementary Material online.

Abstract

Vibrio metoecus is the closest relative of *Vibrio cholerae*, the causative agent of the potent diarrheal disease cholera. Although the pathogenic potential of this new species is yet to be studied in depth, it has been co-isolated with *V. cholerae* in coastal waters and found in clinical specimens in the United States. We used these two organisms to investigate the genetic interaction between closely related species in their natural environment. The genomes of 20 *V. cholerae* and 4 *V. metoecus* strains isolated from a brackish coastal pond on the US east coast, as well as 4 clinical *V. metoecus* strains were sequenced and compared with reference strains. Whole genome comparison shows 86–87% average nucleotide identity (ANI) in their core genes between the two species. On the other hand, the chromosomal integron, which occupies approximately 3% of their genomes, shows higher conservation in ANI between species than any other region of their genomes. The ANI of 93–94% observed in this region is not significantly greater within than between species, meaning that it does not follow species boundaries. *Vibrio metoecus* does not encode toxigenic *V. cholerae* major virulence factors, the cholera toxin and toxin-coregulated pilus. However, some of the pathogenicity islands found in pandemic *V. cholerae* were either present in the common ancestor it shares with *V. metoecus*, or acquired by clinical and environmental *V. metoecus* in partial fragments. The virulence factors of *V. cholerae* are therefore both more ancient and more widespread than previously believed. There is high interspecies recombination in the core genome, which has been detected in 24% of the single-copy core genes, including genes involved in pathogenicity. *Vibrio metoecus* was six times more often the recipient of DNA from *V. cholerae* as it was the donor, indicating a strong bias in the direction of gene transfer in the environment.

Key words: *Vibrio metoecus*, *Vibrio cholerae*, horizontal gene transfer, genomic islands, integron, comparative genomics.

Introduction

The genus *Vibrio* constitutes a diverse group of gammaproteobacteria ubiquitous in marine, brackish, and fresh waters. There are currently over 100 species of vibrios that have been described (Gomez-Gil et al. 2014). This includes clinically significant pathogens such as *Vibrio cholerae*, *Vibrio parahaemolyticus*, and *Vibrio vulnificus* among many others. *Vibrio cholerae*, the causative agent of the potent diarrheal disease cholera, is the most notorious of these human pathogens. Cholera remains a major public health concern, with an estimated 1.2–4.3 million cases and 28,000–142,000 deaths every year worldwide (Ali et al. 2012).

A novel *Vibrio* isolate, initially identified as a nonpathogenic environmental variant of *V. cholerae* (Choopun 2004), was

recently revealed to be a distinct species based on comparative genomic analysis (Haley et al. 2010). Additional environmental strains of this species have been isolated since then (Boucher et al. 2011). Also, since 2006, several clinical strains have been recovered from a range of specimen types (blood, stool, ear, and leg wound) and characterized by the Centers for Disease Control and Prevention (CDC, Atlanta, GA). This recently described species, now officially called *V. metoecus* (Kirchberger et al. 2014), is even more closely related to *V. cholerae* than any other known *Vibrio* species based on biochemical and genotypic tests (Boucher et al. 2011; Kirchberger et al. 2014). Previously, the closest known relative of *V. cholerae* was *Vibrio mimicus*, which was first described as a biochemically atypical strain of *V. cholerae* and named after the fact that it "mimicked *V. cholerae*" phenotypically (Davis et al. 1981).

The discovery of a closely related but distinct species which co-occurs with *V. cholerae* in the environment (Boucher et al. 2011) presents a unique opportunity to investigate the dynamics of interspecies interactions at the genetic level. In their environmental reservoir, bacteria can acquire genetic material from other organisms as a result of horizontal gene transfer (HGT; De la Cruz and Davies 2000). HGT plays an important role in the evolution, adaptation, maintenance, and transmission of virulence in bacteria. It can launch nonpathogenic environmental strains into new pathogenic lifestyles if they obtain the right virulence factors. The two major virulence factors that have led to the evolution from nonpathogenic to toxigenic *V. cholerae* are the cholera toxin (CTX), which is responsible for the cholera symptoms (Waldor and Mekalanos 1996), and the toxin-coregulated pilus (TCP), which is necessary for the colonization of the small intestine in the human host (Taylor et al. 1987). These elements are encoded in genomic islands, specifically called pathogenicity islands, and have been acquired horizontally by phage infections (Waldor and Mekalanos 1996; Karaolis et al. 1999). Another genomic island, the integron, is used to capture and disseminate gene cassettes, such as antibiotic resistance genes (Stokes and Hall 1989). Integrons have been identified in a diverse range of bacterial taxa, and are known to play a major role in genome evolution (Mazel 2006; Boucher et al. 2007). As evidenced by multiple HGT events across a wide range of phylogenetic distances, integrons themselves, not only the cassettes they carry, may have been mobilized within and between species throughout their evolutionary history (Boucher et al. 2007). Integrons are ubiquitous among vibrios, but in some species, such as *V. cholerae*, it can occupy up to 3% of the genome and can contain over a hundred gene cassettes with a wide range of biochemical functions (Mazel et al. 1998; Heidelberg et al. 2000).

Here, we investigate the extent of genetic interaction between *V. metoecus* and *V. cholerae* through comparative genomic analysis, with the focus on the genomic islands, known hotspots for HGT (Dobrindt et al. 2004). The co-isolation of both species in the same environment (Boucher et al. 2011) indicates that *V. metoecus* is likely in constant interaction with *V. cholerae*. Our results show that there is a high rate of gene exchange between species, so rapid in the chromosomal integron that this region is indistinguishable between species. Multiple HGT events were also inferred in the core genome, including genes implicated in pathogenicity, with the majority with *V. metoecus* as a recipient of *V. cholerae* genes, suggesting a directional bias in interspecies gene transfer.

Materials and Methods

Bacterial Strains Used

The *V. metoecus* and *V. cholerae* isolates sequenced in this study as well as genome sequences of additional isolates for comparison are listed in supplementary table S1, Supplementary Material online. Environmental strains of *V. metoecus* and *V. cholerae* were isolated from Oyster Pond (Falmouth, MA) on August and September 2009 using previously described methods (Boucher et al. 2011). Isolates were grown overnight at 37 °C in tryptic soy broth (Becton Dickinson, Sparks, MD) with 1% NaCl (BDH, Toronto, ON, Canada). The sequences of the clinical *V. metoecus* strains were determined by the CDC. Additional sequences were obtained from the National Center for Biotechnology Information (Bethesda, MD) GenBank database.

Genomic DNA Extraction and Quantitation

Genomic DNA was extracted from overnight bacterial cultures with the DNeasy Blood and Tissue Kit (QIAGEN, Hilden, Germany). The concentration for each extract was determined using the Quant-iT PicoGreen double-stranded DNA Assay Kit (Molecular Probes, Eugene, OR) and the Synergy H1 microplate reader (BioTek, Winooski, VT).

Genome Sequencing and Assembly

The genomic DNA extracts were sent to the McGill University and Génome Québec Innovation Centre (Montréal, QC, Canada) for sequencing, which was performed using the TrueSeq library preparation kit and the HiSeq PE100 sequencing technology (Illumina, San Diego, CA). The contiguous sequences were assembled de novo with the CLC Genomics Workbench (CLC Bio, Aarhus, Denmark). Functional annotations of the draft genomes were done in RAST v2.0 (Rapid Annotation using Subsystem Technology; Aziz et al. 2008).

Whole Genome Alignment

A circular BLAST (Basic Local Alignment Search Tool) atlas was constructed to visually compare whole genomes. The annotated genome sequences of *V. metoecus* and *V. cholerae* were aligned by BLASTN (Altschul et al. 1990) against a reference, *V. cholerae* N16961 (Heidelberg et al. 2000), using the CGView Comparison Tool (Grant et al. 2012).

Determination of Orthologous Gene Families and Pan-Genome Analysis

Orthologous groups of open-reading frames (ORFs) from all strains of *V. metoecus* and *V. cholerae* were determined by pairwise bidirectional BLASTP using the OrthoMCL pipeline v2.0 (Li et al. 2003) with 30% match cutoff, as proteins sharing at least 30% identity are predicted to fold similarly (Rost 1999). The gene families were assigned into functional categories based on the Clusters of Orthologous Groups of proteins (COG) database (Tatusov et al. 2000). The pan- and core genome profiles for each species were determined with PanGP v1.0.1 (Zhao et al. 2014) using the distance guide algorithm, repeated 100 times. Sample size and amplification coefficient were set to 1,000 and 100, respectively.

Determination of Genomic Islands

The major genomic islands of *V. cholerae* N16961 were identified using IslandViewer (Langille and Brinkman 2009) and confirmed with previously published data (Heidelberg et al. 2000; Chun et al. 2009). To determine whether a putative homolog is present, ORFs in these genomic islands were compared against the ORFs of *V. metoecus* and *V. cholerae* by calculating the BLAST score ratio (BSR) between reference and query ORF (Rasko et al. 2005) using a custom-developed Perl script (National Microbiology Laboratory, Winnipeg, MB, Canada). Only BSR values of at least 0.3 (for 30% amino acid identity) were considered (Rost 1999).

Determination of the Integron Regions

The chromosomal integron regions of *V. metoecus* and *V. cholerae* were recovered by finding the locations of the integron integrase gene *intI4* and the *attI* and *attC* recombination sites, identified with the ISAAC software (Improved Structural Annotation of *attC*; Szamosi 2012). The *intI4* and gene cassette sequences were used to calculate the ANI (Konstantinidis and Tiedje 2005; Goris et al. 2007) between strains (intra- and interspecies) in JSpecies v1.2.1 (Richter and Rosselló-Móra 2009), using the bidirectional best BLAST hits between nucleotides. The ANI of the integron region was compared with the ANI of 1,560 single-copy core ORFs ($\approx$1.42 Mb).

Phylogenetic Analyses

Using the PhyloPhlAn pipeline v0.99 (Segata et al. 2013), 3,978 amino acid positions based on 400 universally conserved bacterial and archaeal proteins were determined. The concatenated alignment was used to construct a core genome maximum-likelihood (ML) phylogenetic tree, with a BLOSUM45 similarity matrix using the Jones-Taylor-Thorton (JTT) + category (CAT) amino acid evolution model optimized for topology/length/rate using the nearest neighbor interchange (NNI) topology search. Robustness of branching was estimated with Shimodaira-Hasegawa-like (SH-like) support values from 1,000 replicates.

Nucleotide sequences within a gene family were aligned with ClustalW v2.1 (Larkin et al. 2007), and an ML tree was constructed using RAxML v8.1.17 (Stamatakis 2014) using the general time reversible (GTR) nucleotide substitution model and gamma distribution pattern. Robustness of branching was estimated with 100 bootstrap replicates. Interspecies gene transfer events were determined and quantified by comparison of tree topologies using the Phangorn package v1.99-11 (Schliep 2011) in R v3.1.2 (R Development Core Team 2014). A tree was partitioned into clades and determined whether the clades were perfect or not. Following the definition by Schliep et al. (2011), we defined a perfect clade as a partition that is both complete and homogeneous for a given taxonomic category (e.g., a clade with all *V. metoecus*, and only *V. metoecus*). At least one gene transfer event was hypothesized if a tree did not show perfect clades for neither *V. metoecus* nor *V. cholerae* (i.e., in a rooted tree, *V. metoecus* and *V. cholerae* are both polyphyletics).

Resulting alignments of the 1,184 single-copy core gene families not exhibiting HGT were concatenated, and alignment columns with at least one gap were removed using Geneious (Kearse et al. 2012). A final alignment with a total length of 771,455 bp was obtained and used to construct a core genome ML phylogenetic tree with RAxML v8.1.17 (Stamatakis 2014), as described above.

Results and Discussion

Vibrio cholerae is widely studied, and the genomes of globally diverse clinical and environmental isolates are available (supplementary table S1, Supplementary Material online). On the other hand, there are currently only two *V. metoecus* genomes available. Strain RC341 was isolated from Chesapeake Bay (MD) in 1998. It was presumptively identified as a variant *V. cholerae* based on 16 S ribosomal RNA gene similarity to *V. cholerae* (Choopun 2004), but was later reclassified into its current species (Haley et al. 2010; Kirchberger et al. 2014). Strain OP3H was isolated in 2006 from Oyster Pond, a brackish pond in Cape Cod, MA. OP3H is considered the type strain of *V. metoecus*, which was recently officially described as a species (Kirchberger et al. 2014). A screen was performed for atypical *V. cholerae* isolates from a historical collection of clinical isolates at the CDC and identified that several of them were, in fact, *V. metoecus* (Boucher et al. 2011). Additional environmental *V. metoecus* strains were isolated in 2009 from Oyster Pond. While examining the population structure and surveying the mobile gene pool of environmental *V. cholerae* in Oyster Pond, Boucher et al. (2011) discovered that both *V. metoecus* and *V. cholerae* co-occur in this location. To gain a better understanding of

the *V. metoecus* species, we sequenced the genomes of four clinical *V. metoecus* strains originating from patients in the United States and an additional four from Oyster Pond. To be able to evaluate genetic interactions between strains of two different species from the same environment, we sequenced an additional 20 genomes of *V. cholerae* isolates from the same Oyster Pond samples (fig. 1).

Vibrio metoecus: The Closest Relative of *V. cholerae*

To obtain a visual comparison of the genomes, provide an overall impression of genome architecture and identify highly conserved and divergent regions, a circular BLAST atlas was constructed (Grant et al. 2012). *Vibrio metoecus* and representative *V. cholerae* genomes were compared by BLASTN alignment of coding sequences (Altschul et al. 1990) against the reference *V. cholerae* N16961, a pandemic strain from Bangladesh isolated in 1971 whose entire genome was sequenced to completion and carefully annotated (Heidelberg et al. 2000). The BLAST atlas shows a clear distinction between species, as sequence identity is higher within a species than between different species for most genes (fig. 2).

On average, *V. metoecus* shares 84% of its ORFs with *V. cholerae*, whereas 89–91% ORFs are shared between strains of the same species (supplementary table S2, Supplementary Material online). In contrast, *V. mimicus*, previously the closest known relative of *V. cholerae*, shares only 64–69% of ORFs with *V. cholerae* (Hasan et al. 2010). It was determined previously that the recommended cutoff point for prokaryotic species delineation by DNA–DNA hybridization (DDH) is 70%, which corresponds to 85% of conserved protein-coding genes for a pair of strains (Goris et al. 2007). These results show clear distinction between the three closely related species based on conserved genes, and *V. metoecus* is a much closer relative to *V. cholerae* than *V. mimicus*.

Another fundamental measure of relatedness between bacterial strains is ANI. This measure was proposed as a modern replacement to the traditional DDH method to determine relatedness of organisms, but still provide equivalent information (i.e., DNA–DNA similarity; Konstantinidis and Tiedje 2005; Goris et al. 2007). The ANI of the core genome is 86–87% between species and 98–100% within species (fig. 3*a*), showing a clear distinction between *V. metoecus* and *V. cholerae*. Two organisms belonging to the same species will have an ANI of at least 95%, corresponding to 70% DDH (Goris et al. 2007), although earlier studies have proposed a 94% cutoff (Konstantinidis and Tiedje 2005). For this reason, we have currently classified the clinical strain 07-2435 as *V. metoecus* as it shows 94% ANI with other *V. metoecus* strains but only 87% ANI with *V. cholerae* (fig. 3*a*).

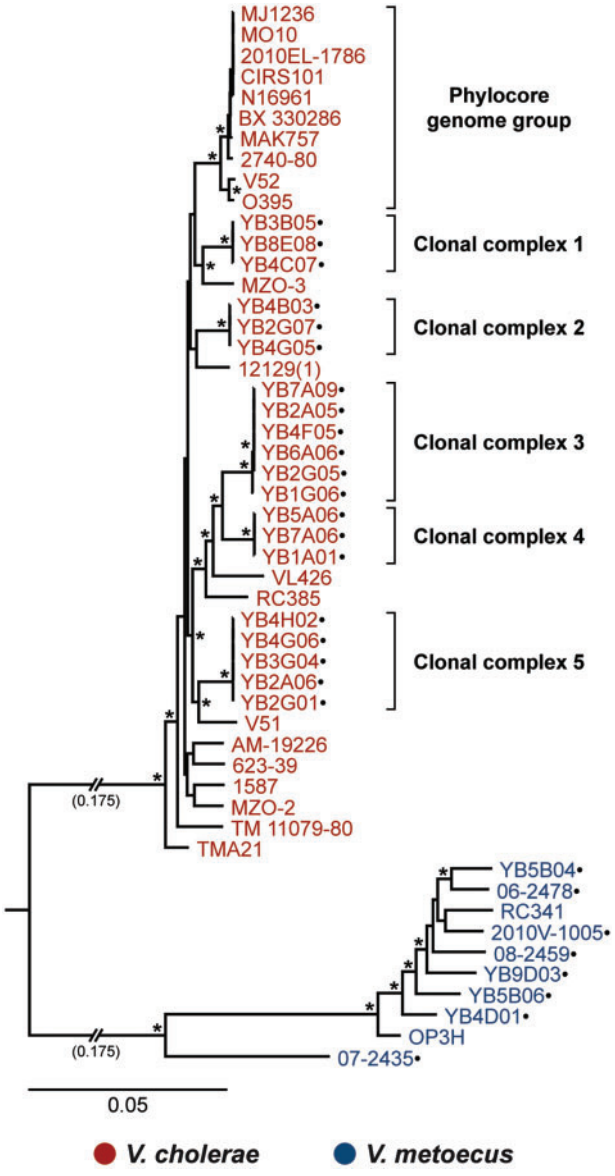


Fig. 1.—The phylogenetic relationship of the *V. metoecus* and *V. cholerae* strains. The ML phylogenetic tree was constructed from the concatenated sequence alignment of single-copy core gene families (771,455 bp). All reliable bootstrap support values are indicated with * and are at least 97% for this tree. The scale bar represents nucleotide substitutions per site. Shortened branch lengths, approximately 3.5× the scale bar (0.175), are indicated. Strains with their genomes sequenced in this study are indicated by dots. Multiple *V. cholerae* strains from Oyster Pond (MA) belong to the same clonal complex.

A Portion of the Genome Escapes the Species Boundary between *V. metoecus* and *V. cholerae*

The BLAST atlas allows for the clear distinction between strains belonging to the *V. cholerae* species and those belonging to the *V. metoecus* species. However, there is a clear and visible exception in one genomic region: The integron. Sequence

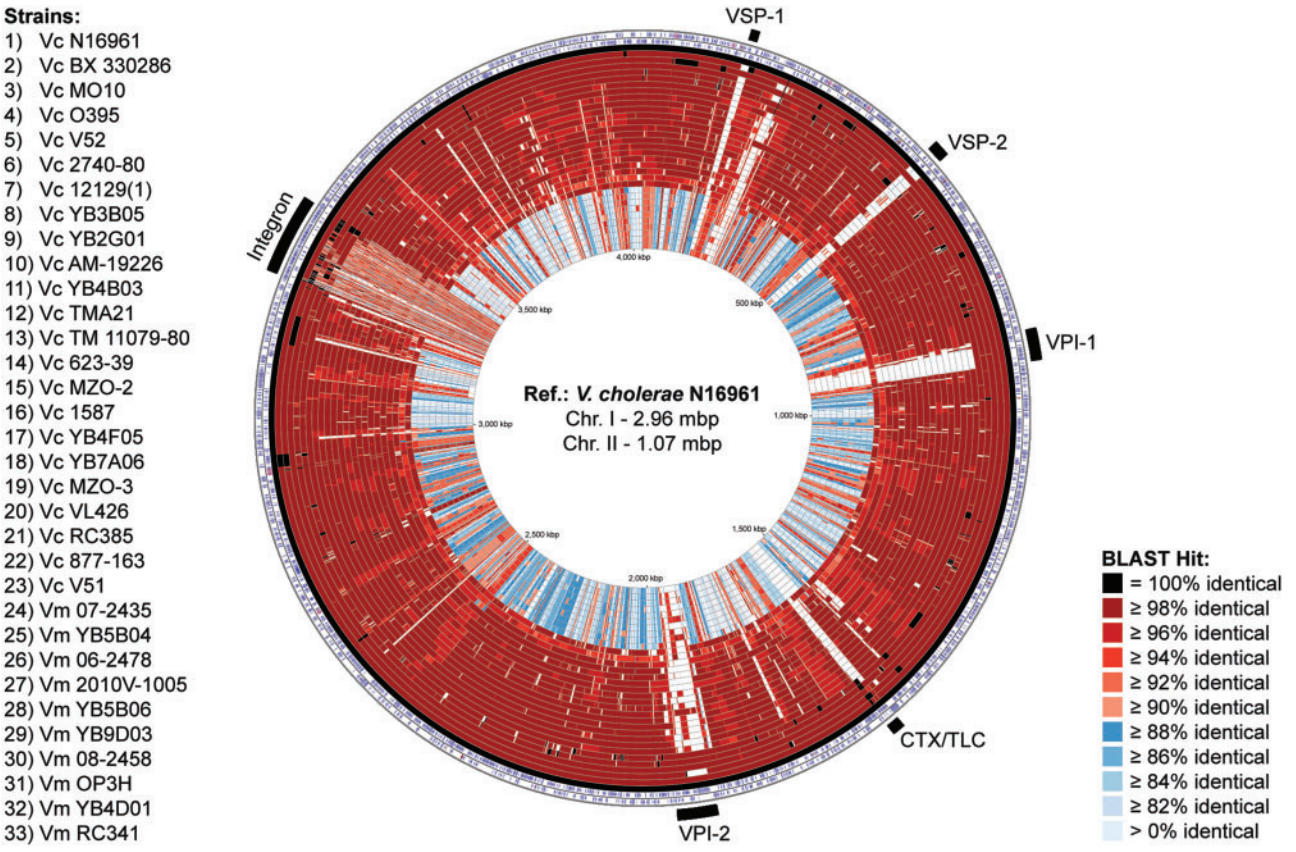


Fig. 2.—The *V. metoecus* (Vm) and *V. cholerae* (Vc) BLAST atlas. The map compares sequenced genomes against the reference (ref.), *V. cholerae* N16961. The two outermost rings show the forward and reverse strand sequence features of the reference. The next 33 rings show regions of sequence similarity detected by BLASTN comparisons between genes of the reference and query genomes. White regions indicate the absence of genes. Outermost black bars indicate the location of the major genomic islands. VSP, *Vibrio* seventh pandemic island; VPI, *Vibrio* pathogenicity island; CTX/TLC, cholera toxin/toxin-linked cryptic; chr., chromosome.

identity of genes found in the integron region does not seem to differ within and between species (fig. 2).

The integron is a region of the genome capable of gene capture and excision (Stokes and Hall 1989) and can occupy up to 3% of the genome in *V. cholerae* (Heidelberg et al. 2000). Although the size of the chromosomal integron region varies between isolates, there is no significant difference in length and number of ORFs between species and between clinical and environmental isolates (supplementary table S3, Supplementary Material online). The ANI of the integron region was determined between pairs of strains and compared with the ANI of the core genome (fig. 3). Although ANI is 86–87% between species and 98–100% within species for the core genome (fig. 3*a*), the integron region displays an average pairwise ANI of 93–94%, both within and between species (fig. 3*b*). Gene cassettes from the 10 *V. metoecus* and 11 *V. cholerae* integron regions were grouped into orthologous gene families, and the occurrence of HGT was quantified for gene families with at least two *V. metoecus* and *V. cholerae* members by the construction of phylogenetic trees. Of the 116 gene families considered, 109 or 94% do not show distinct separation between the two species in a phylogenetic tree. The high number of genes shared between species and their high nucleotide identity are likely the result of frequent interspecies HGT (figs. 2 and 3*b*). A previous study by Boucher et al. (2011) showed that there is indeed a high frequency of gene exchange in the integron region between *V. cholerae* and *V. metoecus*, specifically from the same geographic location (i.e., *V. cholerae* and *V. metoecus* in Oyster Pond) as compared with the same species in different locations (i.e., *V. cholerae* from Bangladesh and the United States). Here, we show that not only is the frequency of interspecies HGT high in the integron, but that its level is such that this region becomes indistinguishable between species.

Although the functions of the majority of integron gene cassettes are unknown (Boucher et al. 2007), many of the known genes are antibiotic resistance genes and are implicated in the evolution of bacteria highly resistant to antibiotics (Collis and Hall 1995; Rowe-Magnus and Mazel 2002). Looking into the predicted functions of the 116 gene families

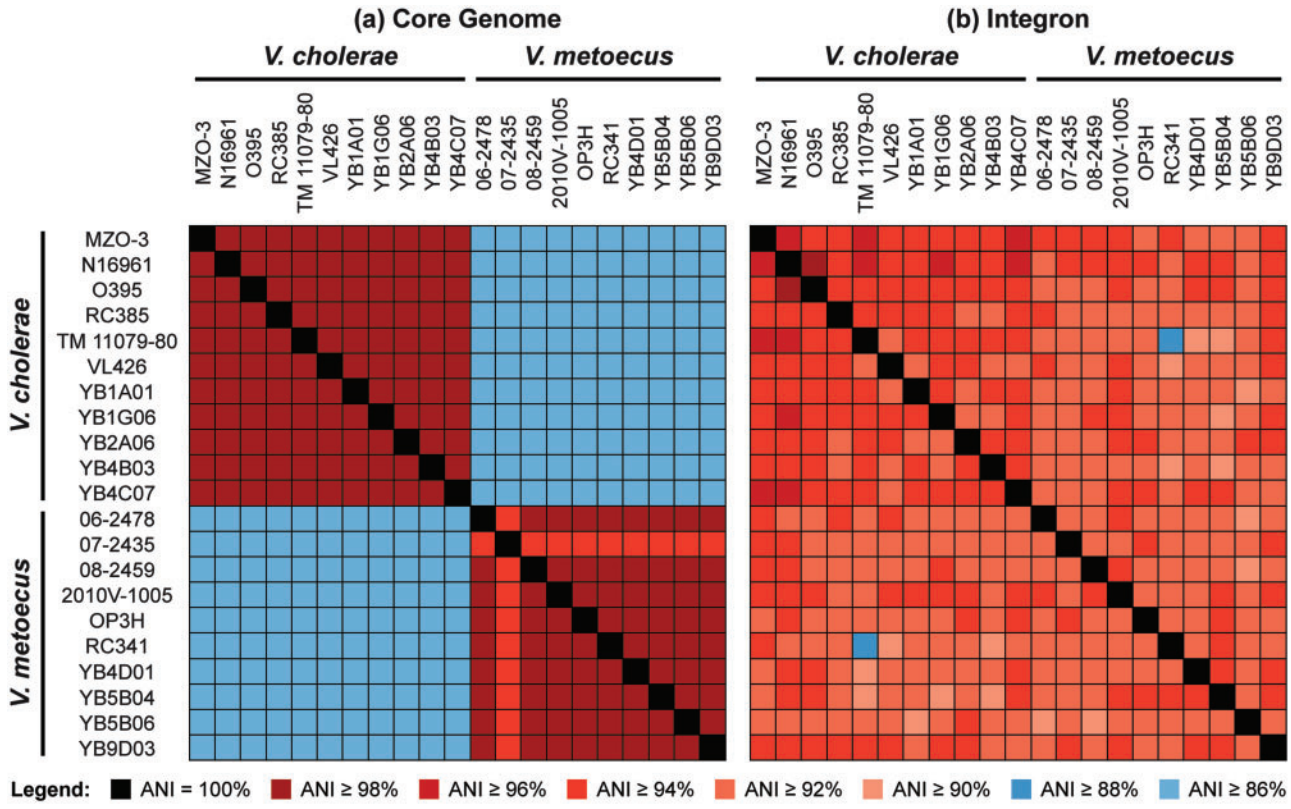


Fig. 3.—ANI of the core genome versus chromosomal integron region of *V. metoecus* and *V. cholerae*. (*a*) Intra- and interspecies pairwise comparison of the 1,560 single-copy core genes (≈1.42 Mb). (*b*) Intra- and interspecies pairwise comparison of the integron gene cassettes.

comprising 1,452 gene cassettes, the majority of which are shared between *V. metoecus* and *V. cholerae*, reveals genes that encode proteins involved in transport and metabolism of various molecules (supplementary fig. S1, Supplementary Material online), suggesting a major contributing function of the integron for host acquisition and distribution of important resources in the environment by bacteria (Koenig et al. 2008). Gene cassettes encoding nicotinamidase-related amidases are present in multiple copies. Nicotinamidase catalyzes the deamination of nicotinamide to produce ammonia and nicotinic acid (Petrack et al. 1965). A key enzyme in many organisms, nicotinamidase has been shown to be important in the proliferation of bacteria pathogenic to mammalian hosts including humans (Purser et al. 2003; Kim et al. 2004). Other genes present are involved in basic cellular functions such as acetyltransferases, involved in posttranslational modifications of ribosomal proteins, the functional significance of which remains unclear but may have regulatory roles (Nesterchuk et al. 2011). Some genes are part of the plasmid stabilization systems, which include the toxin–antitoxin (TA) systems. TA systems are frequently found in gene cassette arrays for the stabilization and prevention of loss of gene cassettes. They also play additional roles in stress response, bacterial persistence, and phage defense (Iqbal et al. 2015).

A Lack of Reciprocity: Directional Gene Flow from *V. cholerae* to *V. metoecus*

To get a quantitative estimate of the amount of HGT between *V. cholerae* and *V. metoecus*, we investigated the amount of interspecies recombination taking place in their core genomes. An ML tree was constructed for each of the 1,947 gene families comprising the *V. metoecus*–*V. cholerae* core genome (fig. 4). The trees were then analyzed for gene transfer events by partitioning them into clades (Schliep 2011). In our analysis, following the definition by Schliep et al. (2011), a gene transfer is hypothesized if a member of one species clusters with members of the other species in a clade, and the tree cannot be partitioned into perfect clades, which must consist of all members from the same species and only of that species. Considering only the single-copy core genes, we have inferred interspecies HGT in 376 of 1,560 genes (24%; supplementary table S4, Supplementary Material online). Our analysis excluded 387 core genes that have duplicates in at least one of the genomes, as it is difficult to reliably assess HGT in genes from large paralogous families (Ge et al. 2005). Using this method, it was possible to determine directionality of HGT, whether from *V. cholerae* to *V. metoecus* or vice versa. HGT was qualified by examining the individual gene trees, and only reliable clustering with at

least 70% bootstrap support was considered (Hillis and Bull 1993). A total of 655 interspecies gene transfer events were detected, with the majority (489 or 75%; $P = 0.0053$) with *V. metoecus* as the recipient (i.e., *V. metoecus* members clustering within the *V. cholerae* clade). On the other hand, we detected 166 (25%) of gene transfer events with *V. cholerae* as the recipient (supplementary table S5, Supplementary Material online).

To investigate whether this bias in directionality of HGT was due to differences in the origin or ecology of strains from one species or the other, we performed the analysis using only environmental strains from Oyster Pond. To ensure equal genetic diversity for both species, we compared the same number of isolates from each species. The 20 *V. cholerae* isolates we sequenced for this study can be grouped into five clonal complexes as determined by multilocus sequence typing of seven housekeeping genes. All the isolates from the same clonal complex cluster together in a core genome phylogenetic tree (fig. 1). They also exhibit 100% ANI only with each other but not with isolates from other clonal complexes (supplementary table S6, Supplementary Material online). Indeed, members of the same clonal complex always cluster together in all the individual gene trees examined (fig. 4). We therefore randomly chose one isolate from each *V. cholerae* clonal complex from Oyster Pond, yielding a final data set of five genomes from each species. A total of 224 interspecies gene transfer events were detected in this environment-specific data set, where 192 (86%; $P = 0.0012$) involved *V. metoecus* as the recipient and only 32 (14%) with *V. cholerae* as the recipient (table 1). One possibility to explain this bias could be that *V. cholerae* genes are more abundant in the environment and therefore more accessible to *V. metoecus*. Indeed, using culture-based methods, *V. cholerae* was ten times more abundant than *V. metoecus* in Oyster Pond. Another possibility is that *V. cholerae* is more refractory to HGT as they contain more barriers to gene uptake, such as restriction-modification systems, or that *V. metoecus* is more permissive, containing more DNA uptake systems (conjugative plasmids, natural competence machinery or phages). However, no significant difference could be found in the number or nature of proteins involved in restriction-modification or DNA uptake systems between *V. metoecus* and *V. cholerae* in our study, although poorly transformable *V. cholerae*, despite having an intact and perfectly functioning DNA uptake system, have been reported (Katz et al. 2013). Additionally, nuclease activity by Dns, Xds, and other DNases can inhibit natural transformation (Blokesch and Schoolnik 2008; Gaasbeek et al. 2009). We also surveyed our *V. metoecus* and *V. cholerae* genomes for predicted DNases and found no significant difference between species.

Despite the directional gene transfer from *V. cholerae* to *V. metoecus*, it seems that the latter might have contributed to the virulence of its more famous relative by HGT. Interspecies recombination was detected in four core genes where at least one clinical *V. cholerae* grouped in the same clade with *V. metoecus* (fig. 4). Interestingly, three of these genes are implicated, whether directly or indirectly, in *V. cholerae* pathogenesis. VC2614 encodes a cyclic adenosine monophosphate regulatory protein, a global regulator of gene expression in *V. cholerae* including CTX and TCP (Skorupski and Taylor 1997). It appears that HGT in this case occurred in the ancestor of the phylocore genome (PG) group, which contains all pandemic strains (fig. 1; Chun et al. 2009), with a clinical *V. metoecus* strain as the possible donor. The new version of this cyclic-AMP regulatory protein was eventually lost in the classical O1 strain (O395). VC2545 encodes an inorganic pyrophosphatase, and its expression in *V. cholerae* may play an important role during human and mouse infection (Lombardo et al. 2007). This transfer was only between clinical *V. metoecus* and classical O1. VCA0925 encodes a dihydroorotase essential for pyrimidine biosynthesis. Biosynthesis of nucleotides is the single most critical metabolic function for growth of pathogenic bacteria in the bloodstream because of scarcity of nucleotide precursors but not other nutrients, and the genes involved serve as potential antibiotic targets for treatments of blood infection (Samant et al. 2008). Here, gene transfer involved not just the PG group of *V. cholerae* but also the environmental strains of clonal complex 5 and 623-39.

Although these interspecies recombination events do not represent novel gene acquisitions, gaining a new allele of a gene can often have important consequences in a pathogen, changing its fitness in the host. This has been demonstrated for single-point mutations in *ompU*, *vpvC*, and *ctxB*. The *ompU* gene encodes for the major outer membrane porin OmpU, generally for the transport of hydrophilic solutes, but has been shown to provide *V. cholerae* resistance to bile acids and antimicrobial peptides in the host (Provenzano et al. 2000; Mathur and Waldor 2004). It is suggested that it can also act as a receptor for phage to infect *V. cholerae* (Seed et al. 2014). The *vpvC* gene encodes for diguanylate cyclase, and the mutation results in a switch from the smooth to rugose phenotype in *V. cholerae* (Beyhan and Yildiz 2007). The single-point mutations in these genes result in a *V. cholerae* that is less susceptible to phage infection, contributing to the evolutionary success of the pathogen (Beyhan and Yildiz 2007; Seed et al. 2014). *Vibrio cholerae* responsible for cholera outbreaks in Bangladesh have changing genotypes of *ctxB*, a subunit of CTX (Waldor and Mekalanos 1996), also caused by a single-point mutation (Rashed et al. 2012). The years 2006 and 2007 saw a dominance of *V. cholerae* with the *ctxB* genotype 1 (*ctxB*1). *Vibrio cholerae* with the *ctxB* genotype 7 (*ctxB*7) outcompeted *ctxB*1 from 2008 to 2012. However, there appears to be a shift back to *ctxB*1 since 2013. The changing *ctxB* genotypes were associated with differing levels of severity of cholera. This also suggests CTX phage-mediated evolution, survival, and dominance of *V. cholerae* (Rashed et al. 2012; Rashid et al. 2015).

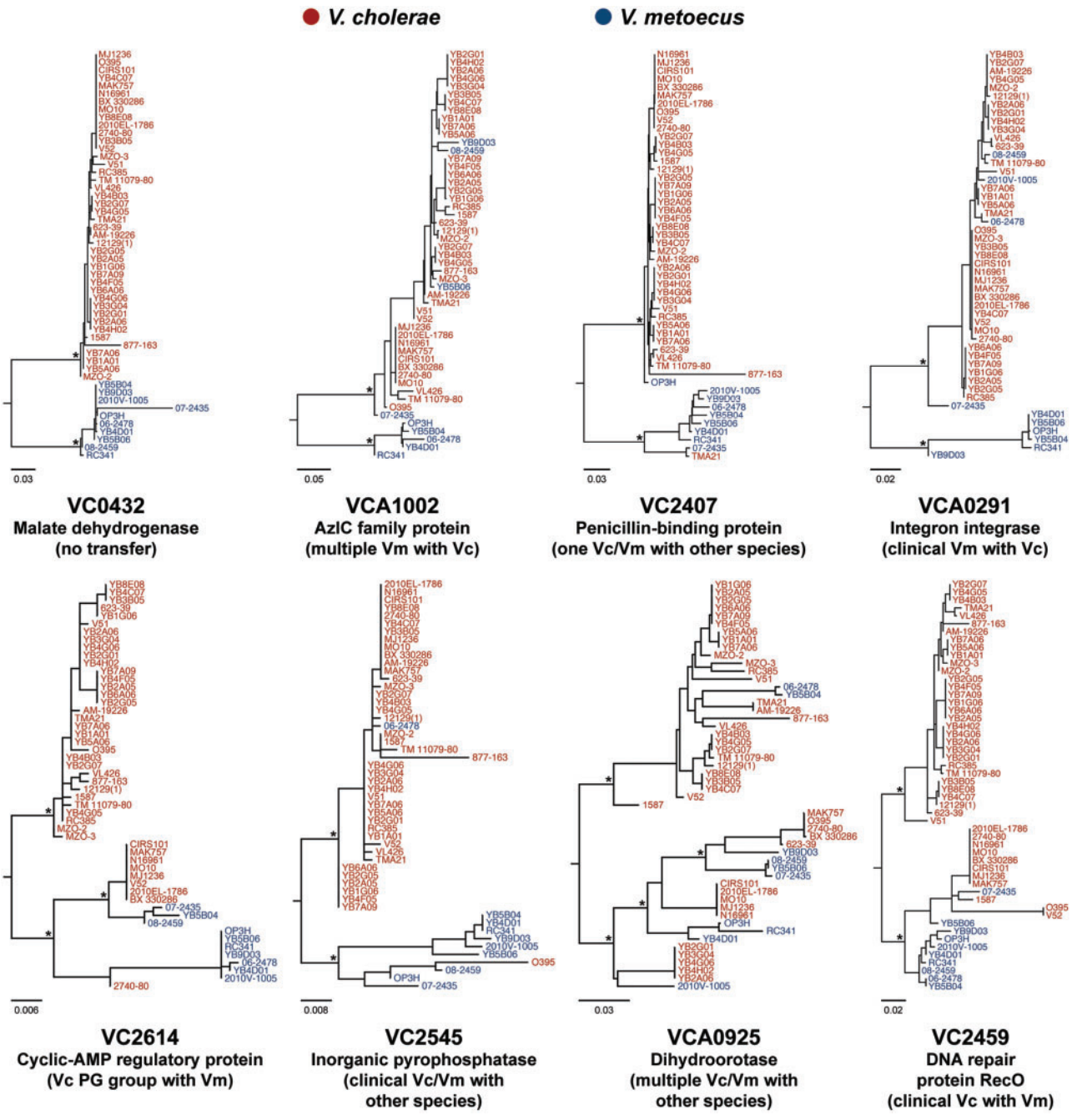


Fig. 4.—Representative HGT between *V. metoecus* (Vm) and *V. cholerae* (Vc). The trees are representative ML phylogenetic trees from 1,560 orthologous families of single-copy core genes showing various examples of transfer events. Bottom trees: Transfers involving at least one clinical *V. cholerae* clustering with *V. metoecus*. Relevant bootstrap support (>70%) is indicated with *. The scale bars represent nucleotide substitutions per site.

Components of Major Pathogenicity Islands Are More Ancient than the *V. cholerae* Species

A BSR map (Rasko et al. 2005) was constructed to show the presence or absence of the genes comprising the major pathogenicity islands in various *V. metoecus* and *V. cholerae* isolates (fig. 5). Using the genes from *V. cholerae* N16961 as reference, BLASTP was used to determine the presence of homologous genes in the other strains (Altschul et al. 1990). The major *V. cholerae* virulence factors, CTX and TCP, which are encoded by pathogenicity islands that have been acquired horizontally by phage infections of the CTXΦ and VPIΦ, respectively (Waldor and Mekalanos 1996; Karaolis et al. 1999), are absent from all clinical and environmental *V. metoecus* (fig. 5a). The absence of CTX and TCP in

Table 1

HGT Count for *Vibrio metoecus* and Representative *Vibrio cholerae* Strains from Oyster Pond (MA) Based on 376 Single-Copy Core Genes with Inferred HGT

Species and Strain	HGT Count	Percent of Total
Vibrio metoecus OP3H	55	25
Vibrio metoecus YB4D01	43	19
Vibrio metoecus YB5B06	37	17
Vibrio metoecus YB5B04	30	13
Vibrio metoecus YB9D03	27	12
	192	**86**
Vibrio cholerae YB2G01 (CC 5)	16	7
Vibrio cholerae YB4F05 (CC 3)	9	4
Vibrio cholerae YB4B03 (CC 2)	4	2
Vibrio cholerae YB7A06 (CC 4)	2	1
Vibrio cholerae YB3B05 (CC 1)	1	0
Total	**32**	**14**

NOTE.—Only one strain from each clonal complex (CC) was included. An HGT event was hypothesized when a strain clustered with members of the other species in a phylogenetic tree, with reliable bootstrap support (>70%). Unequal variance *t*-test, $P = 0.0012$.

V. metoecus is consistent with the absence of reports on a toxigenic *V. metoecus*.

Interestingly, our results show some of the other major pathogenicity islands to be present in some *V. metoecus* and nonpandemic *V. cholerae* strains in fragments and not as a complete presence or absence. This is evident in the *Vibrio* pathogenicity island 2 (VPI-2), which can be divided into four subclusters we call "islets," as indicated in figure 5*a*. These four islets match the previous description of Jermyn and Boyd (2002) for VPI-2: 1) A type-1 restriction-modification system for protection against viral infection, 2) a *nan-nag* cluster for sialic acid metabolism, 3) a Mu phage-like region, and 4) a number of ORFs of unknown function. We hypothesize two scenarios as to the fragmentation of these genomic islands 1) that the islands were obtained as a whole and sections were eventually lost, or 2) that the islands were acquired independently in islets and were accreted into the same region in the genome. Evolution would favor the latter hypothesis, as it is more parsimonious for fewer environmental strains to independently acquire certain islets of the islands rather than a majority of the strains acquiring whole islands and losing most regions eventually (Freeman and Herron 2007). Phylogenetic trees were constructed for the gene families that constitute the four putative islets of VPI-2. Gene trees for islet B, the *nan-nag* cluster, show distinct clustering of *V. metoecus* and *V. cholerae*, suggesting the acquisition of this region by a common ancestor, which diverged and evolved independently after speciation, with more recent isolated HGT events between *V. metoecus* and *V. cholerae* (fig. 5*b* and supplementary fig. S2, Supplementary Material online). A similar pattern of distinct clustering of *V. metoecus* and *V. cholerae* is also observed in islet A, but the latter is only present in O1 El Tor *V. cholerae* and two *V. metoecus* strains (fig. 5*a*), suggesting that it was horizontally transferred between the two species and likely absent from their common ancestor. Furthermore, islet C, the putative Mu phage-like region, is only detected in *V. cholerae* of the PG group and TM 11079-80, an O1 El Tor environmental isolate. This islet is absent in *V. metoecus*, which suggests a more recent acquisition of this region only by certain *V. cholerae*. Finally, islet D is prevalent in the majority of the isolates, whether *V. metoecus* or *V. cholerae*, which do not cluster by species in the phylogeny (fig. 5*b*). This suggests frequent interspecies HGT of its component genes. Taken together, these results support that the VPI-2 island emerged by accretion of smaller islets with different evolutionary histories before reaching the form currently found in *V. cholerae* O1 El Tor or classical pandemic strains. The *nan-nag* cluster (islet B) is likely ancestral, being present before speciation of *V. cholerae* and *V. metoecus*, with islets A and D acquired later by the ancestor of pandemic *V. cholerae* through HGT within or between species and islet C added most recently through HGT from an unknown source.

The *Vibrio* seventh pandemic islands 1 and 2 (VSP-1 and VSP-2, respectively) are genomic islands believed to be present and unique only among the seventh pandemic isolates of *V. cholerae* (Dziejman et al. 2002; O'Shea et al. 2004). These VSPs are hypothesized to provide a fitness advantage to these isolates. However, multiple variants of VSP-2 have been detected in *V. cholerae*, including non-O1/O139 strains, by acquisition and loss of genes at specific loci within a conserved core genomic backbone (Taviani et al. 2010). This core VSP-2 is also present in two *V. metoecus* isolates, the clinical 2010 V-1005 and environmental RC341 (fig. 5*a*), and may have been acquired from *V. cholerae*, as indicated by the great similarity of genes in this region to *V. cholerae* and phylogenetic analysis (fig. 5*b* and supplementary fig. S3, Supplementary Material online). This variant of VSP-2 is stable and present in diverse strains isolated from different times and geographic locations and may be the one circulating among non-O1/O139 isolates (Taviani et al. 2010). VSP-1 is present almost in its entirety in environmental *V. cholerae* VL426 and *V. metoecus* RC341 (fig. 5*a*); similar strains in the environment may serve as reservoirs of VSP-1. There is no correlation between the presence of VSP-1 and VSP-2 in non-O1/O139 *V. cholerae*, indicating that both islands were acquired independently in different HGT events by seventh pandemic *V. cholerae* (Taviani et al. 2010). The presence of both of the entire VSP-1 and the core of VSP-2 in *V. metoecus* strains indicate interspecies movement of pathogenicity islands, suggesting that interspecies transfer can contribute to the evolution of pathogenic variants.

Fundamental Genetic Differences between *V. metoecus* and *V. cholerae*

To determine genetic differences between *V. cholerae* and *V. metoecus* and the unique gene content of each species,

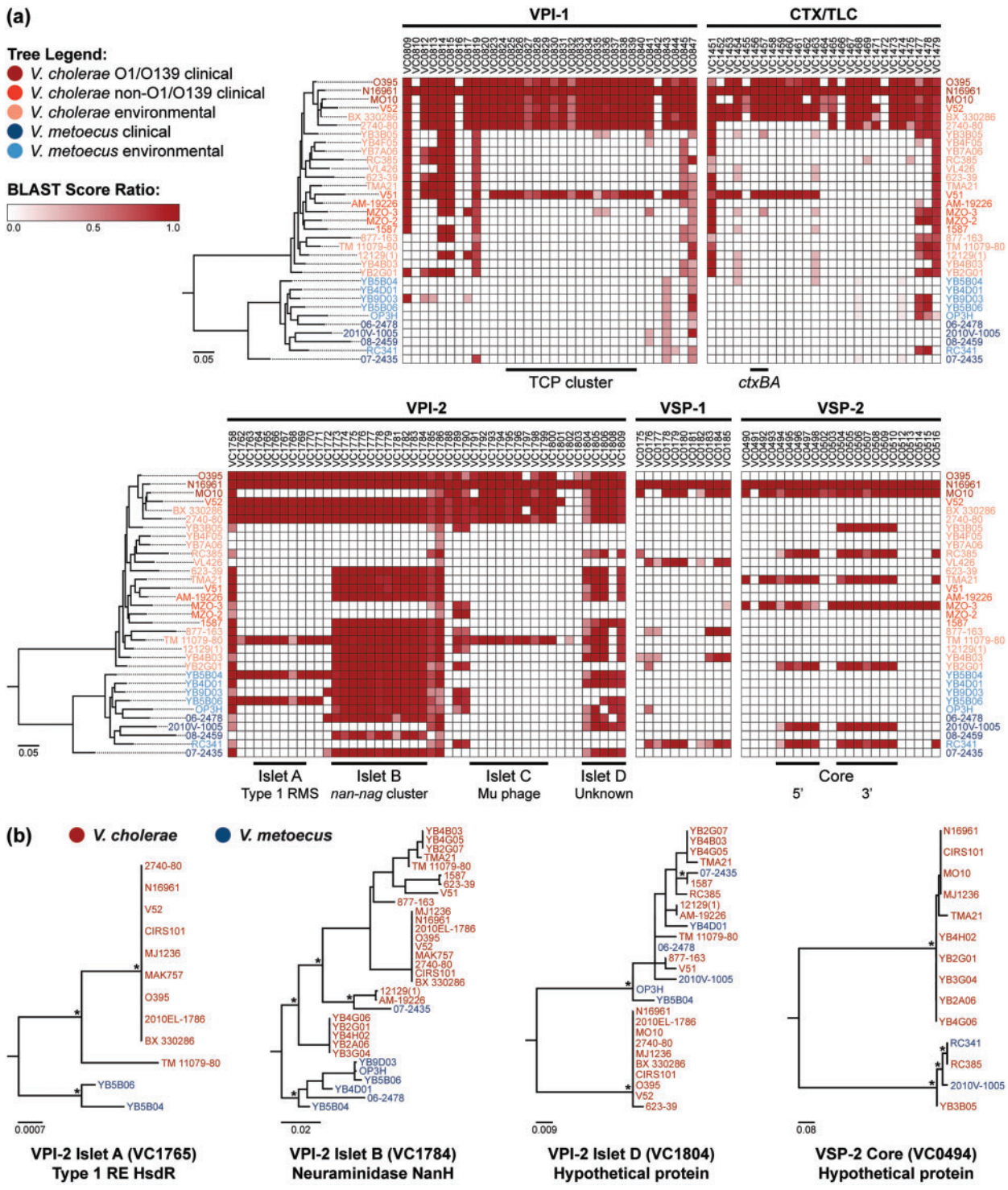


Fig. 5.—Virulence factors present in *V. metoecus* and *V. cholerae*. (*a*) The phylogenetic relationship of the *V. metoecus* and *V. cholerae* strains is shown on the left of each BSR map. The ML phylogenetic tree was constructed using 3,978 amino acid positions based on 400 universally conserved bacterial and archaeal proteins. The scale bars represent amino acid substitutions per site. The columns on the BSR maps show genes (locus tags) from genomic islands VPI-1, CTX/TLC, VPI-2, VSP-1, and VSP-2 of the reference, *V. cholerae* N16961. The black bars at the bottom of the BSR maps indicate the TCP cluster of VPI-1, *ctxAB* of CTX/TLC, islets of VPI-2, and core regions of VSP-2. The gradient bar shows the BSRs and their corresponding colors, with white regions indicating the absence of genes. Only BSR values of at least 0.3 were included. VPI, *Vibrio* pathogenicity island; CTX/TLC, cholera toxin/toxin-linked cryptic; VSP, *Vibrio* seventh pandemic island; RMS, restriction-modification system. (*b*) Representative ML phylogenetic trees of orthologous gene families of the VPI-2 islets and the VSP-2 core. Relevant bootstrap support (>70%) is indicated with *. The scale bars represent nucleotide substitutions per site. RE, restriction endonuclease.

we first compiled their pan- and core genomes (supplementary fig. S4, Supplementary Material online). The pan-genome is the entire gene repertoire of a bacterial species, whereas the core genome comprises genes shared by all the strains (Tettelin et al. 2005, Vernikos et al. 2015). ORFs from both species were assigned to orthologous groups based on sequence similarity, yielding pan- and core genomes containing 5,613 and 2,089 gene families, respectively, based on the 42 *V. cholerae* genomes used in this study (supplementary fig. S4*a*, Supplementary Material online). This differs from the previous estimate of Chun et al. (2009), who determined the *V. cholerae* core genome to contain 2,432 gene families based on 23 strains, a higher core genome size than we obtained from our data set. The reduced core genome size is expected as the number of shared genes decreases with the addition of each new genome (Tettelin et al. 2005). It also depends on the degree of relatedness of the organisms. A study on 32 *Vibrionaceae* genomes, including 18 representative *V. cholerae*, established a core genome of only 1,000 gene families (Vesth et al. 2010). The *V. metoecus* pan- and core genomes constitute 4,298 and 2,872 gene families, respectively, based on the ten genomes currently available (supplementary fig. S4*b*, Supplementary Material online). The difference in pan- and core genome sizes of *V. cholerae* and *V. metoecus* can be explained by the significant difference in the number of genomes used. We expect the pan- and core genomes of *V. metoecus* to ultimately reach sizes similar to that of *V. cholerae* when genomes of additional strains become available.

As a newly described species, very little is currently known about the biology of *V. metoecus* and what sets it apart genetically from *V. cholerae*. From the combined pan-genome of both species, orthologous gene families present in various groups of strains were determined: Families unique to *V. metoecus* and *V. cholerae*, or unique to clinical and environmental strains (supplementary fig. S5, Supplementary Material online). Function was predicted for each gene family based on the COG database (supplementary fig. S6, Supplementary Material online). *Vibrio metoecus* contains more unique gene families than *V. cholerae* that are involved in carbohydrate transport and metabolism (supplementary fig. S6*a*, Supplementary Material online). In the species description study by Kirchberger et al. (2014), it was determined that although the majority of biochemical and growth characteristics of *V. metoecus* resemble *V. cholerae*, the former was mainly differentiated from the latter for its ability to utilize the complex sugars D-glucuronic acid and *N*-acetyl-D-galactosamine. Indeed, multiple β-galactosidase/β-glucuronidase enzymes for the breakdown of D-glucuronic acid (Louis and Doré 2014) were present in our *V. metoecus*-specific COG data set, but not in *V. cholerae*. Multiple hexosaminidases for the hydrolysis of terminal *N*-acetyl-D-hexosamine (Magnelli et al. 2012) were also detected in *V. metoecus*, which supports the phenotype observed by Kirchberger et al. (2014). Additionally, genes unique for clinical *V. metoecus* and clinical *V. cholerae* were identified (supplementary fig. S6*b*, Supplementary Material online). Clinical *V. cholerae* have more genes encoding proteins involved in replication, recombination, and repair (mostly transposases), and signal transduction, such as the GGDEF family protein. Transposases in pathogenicity islands can contribute to the instability and mobilization of virulence genes (Schmidt and Hensel 2004). The GGDEF family protein is critical in biofilm formation (García et al. 2004) and is highly induced in *V. cholerae* during infection in humans and mice (Lombardo et al. 2007). As expected, genes of the CTX and TCP clusters were not found in our clinical *V. cholerae*-specific data set because they are not unique to clinical strains, but are also present in some environmental ones (fig. 5*a*). Among the genes uniquely found in clinical *V. metoecus* is a putative *mdaB* (modulator of drug activity B) gene. The *mdaB* gene has been shown to play an important role in oxidative stress resistance and host colonization in *Helicobacter pylori* (Wang and Maier 2004), and may also contribute to the fitness of clinical *V. metoecus* in the host.

Conclusion

The discovery of *V. metoecus*, the closest known relative of *V. cholerae*, presents an opportunity to study the HGT events between species and the role this might play in the evolution of pathogenesis. In contrast to the core genome, which is distinctly more similar between members of the same species, the chromosomal integron region, occupying approximately 3% of *V. cholerae* and *V. metoecus* genomes, represents a pool of genes which is freely exchanged between these two species. This genomic region displays no greater similarity within than between species. Genomic islands encoding pathogenicity factors, known to play a role in pandemic *V. cholerae* virulence, are also occasionally found in *V. metoecus*, either completely or in part. This includes VPI-2, found in most pandemic *V. cholerae*, as well as the VSP islands, previously believed to be specific to *V. cholerae* strains from the seventh pandemic. VPI-2 and VSP-2 seem to have assembled over time by accretion of smaller units, which we call islets. Some islets, such as the *nan-nag* cluster of the VPI-2 (islet B) for sialic acid metabolism, have been stable over time and were present in the common ancestor of *V. metoecus* and *V. cholerae*. Other islets, such as islet A (restriction-modification system) and islet D (unknown function) of VPI-2, the core of VSP-2, or the entire VSP-1 island seem to move frequently between *V. metoecus* and *V. cholerae* and are not restricted to pandemic strains.

The most striking finding is that even the core genome of *V. cholerae* is susceptible to frequent interspecies recombination with *V. metoecus*. Twenty-four percent of the genes found in all *V. cholerae* and *V. metoecus* had experienced interspecies recombination. There also seems to be a

directional bias to these recombination events. In Oyster Pond, in particular, *V. metoecus* is the recipient of genes six times more than *V. cholerae*. The cause of this bias is unclear, but it does not seem to be restricted to a single environment, as all *V. metoecus* are recipients of more interspecies DNA transfers than any of the *V. cholerae* strains investigated. One possibility is that *V. cholerae* is more abundant in most environments than *V. metoecus* and there is, therefore, simply more of its DNA available for uptake. Indeed, in this study, *V. cholerae* was isolated ten times more frequently than *V. metoecus* from Oyster Pond, which is consistent with the observed HGT bias. However, this explanation is very tentative and requires more evidence, as this study is the first one to isolate *V. cholerae* and *V. metoecus* quantitatively from the same site, and this was done using a culture-based method. This relative abundance would not necessarily be obtained with more accurate culture-free quantitative methods. Also, HGT could be biased because of differences in phage abundance/susceptibility, presence of DNA uptake systems, or restriction-modification systems. Nonetheless, this is, to our knowledge, the first quantitative report of HGT bias for bacteria in the natural environment and has fundamental implications for understanding the evolution of microbial populations.

Supplementary Material

Supplementary figures S1–S6 and tables S1–S6 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors are grateful for the assistance of Tania Nasreen, Paul Stothard (University of Alberta), Lee Katz, Mike Frace, Maryann Turnsek (Centers for Disease Control and Prevention), Éric Bapteste (Université Pierre et Marie Curie), Gary Van Domselaar, and Aaron Petkau (National Microbiology Laboratory). They appreciate the helpful discussions with Rebecca Case, Stefan Pukatzki, and David Wishart (University of Alberta). This work was supported by the Natural Sciences and Engineering Research Council of Canada, the Canadian Institute for Advanced Research, the Canadian Foundation for Innovation (to Y.B.), and the Alberta Innovates—Technology Futures (to F.D.O. and P.C.K.).

Literature Cited

Ali M, et al. 2012. The global burden of cholera. Bull World Health Organ. 90:209–218A.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Aziz RK, et al. 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics 9:75.

Beyhan S, Yildiz FH. 2007. Smooth to rugose phase variation in *Vibrio cholerae* can be mediated by a single nucleotide change that targets c-di-GMP signalling pathway. Mol Microbiol. 63:995–1007.

Blokesch M, Schoolnik GK. 2008. The extracellular nuclease Dns and its role in natural transformation of *Vibrio cholerae*. J Bacteriol. 190:7232–7240.

Boucher Y, et al. 2011. Local mobile gene pools rapidly cross species boundaries to create endemicity within global *Vibrio cholerae* populations. MBio 2:e00335–10.

Boucher Y, Labbate M, Koenig JE, Stokes HW. 2007. Integrons: mobilizable platforms that promote genetic diversity in bacteria. Trends Microbiol. 15:301–309.

Choopun N. 2004. The population structure of *Vibrio cholerae* in Chesapeake Bay [Ph.D. thesis]. [College Park (MD)]: University of Maryland.

Chun J, et al. 2009. Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. Proc Natl Acad Sci U S A. 106:15442–15447.

Collis CM, Hall RM. 1995. Expression of antibiotic resistance genes in the integrated cassettes of integrons. Antimicrob Agents Chemother. 39:155–162.

Davis BR, et al. 1981. Characterization of biochemically atypical *Vibrio cholerae* strains and designation of a new pathogenic species, *Vibrio mimicus*. J Clin Microbiol. 14:631–639.

De la Cruz F, Davies J. 2000. Horizontal gene transfer and the origin of species: lessons from bacteria. Trends Microbiol. 8:128–133.

Dobrindt U, Hochhut B, Hentschel U, Hacker J. 2004. Genomic islands in pathogenic and environmental microorganisms. Nat Rev Microbiol. 2:414–424.

Dziejman M, et al. 2002. Comparative genomic analysis of *Vibrio cholerae*: genes that correlate with cholera endemic and pandemic disease. Proc Natl Acad Sci U S A. 99:1556–1561.

Freeman S, Herron JC. 2007. Evolutionary analysis. San Francisco (CA): Pearson Benjamin Cummings.

Gaasbeek EJ, et al. 2009. A DNase encoded by integrated element CJIE1 inhibits natural transformation of *Campylobacter jejuni*. J Bacteriol. 191:2296–2306.

García B, et al. 2004. Role of the GGDEF protein family in *Salmonella* cellulose biosynthesis and biofilm formation. Mol Microbiol. 54:264–277.

Ge F, Wang LS, Kim J. 2005. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. PLoS Biol. 3:e316.

Gomez-Gil B, et al. 2014. The family *Vibrionaceae*. In: Rosenberg E, DeLong EF, Lory S, Stackebrandt E, Thompson F, editors. The prokaryotes—gammaproteobacteria. Berlin (Germany): Springer. p. 659–747.

Goris J, et al. 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol. 57:81–91.

Grant JR, Arantes AS, Stothard P. 2012. Comparing thousands of circular genomes using the CGView Comparison Tool. BMC Genomics 13:202.

Haley BJ, et al. 2010. Comparative genomic analysis reveals evidence of two novel *Vibrio* species closely related to *V. cholerae*. BMC Microbiol. 10:154.

Hasan NA, et al. 2010. Comparative genomics of clinical and environmental *Vibrio mimicus*. Proc Natl Acad Sci U S A. 107:21134–21139.

Heidelberg JF, et al. 2000. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. Nature 406:477–483.

Hillis DM, Bull JJ. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst Biol. 42:182–192.

Iqbal N, Guérout AM, Krin E, Le Roux F, Mazel D. 2015. Comprehensive functional analysis of the 18 *Vibrio cholerae* N16961 toxin-antitoxin

systems substantiates their role in stabilizing the superintegron. J Bacteriol. 197:2150–2159.

Jermyn WS, Boyd EF. 2002. Characterization of a novel *Vibrio* pathogenicity island (VPI-2) encoding neuraminidase (*nanH*) among toxigenic *Vibrio cholerae* isolates. Microbiology 148:3681–3693.

Karaolis DK, Somara S, Maneval DR Jr, Johnson JA, Kaper JB. 1999. A bacteriophage encoding a pathogenicity island, a type-IV pilus and a phage receptor in cholera bacteria. Nature 399:375–379.

Katz LS, et al. 2013. Evolutionary dynamics of *Vibrio cholerae* O1 following a single-source introduction to Haiti. MBio 4:e00398–13.

Kearse M, et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–1649.

Kim S, et al. 2004. *Brucella abortus* nicotinamidase (PncA) contributes to its intracellular replication and infectivity in mice. FEMS Microbiol Lett. 234:289–295.

Kirchberger PC, et al. 2014. *Vibrio metoecus* sp. nov., a close relative of *Vibrio cholerae* isolated from coastal brackish ponds and clinical specimens. Int J Syst Evol Microbiol. 64:3208–3214.

Koenig JE, et al. 2008. Integron-associated gene cassettes in Halifax Harbour: assessment of a mobile gene pool in marine sediments. Environ Microbiol. 10:1024–1038.

Konstantinidis KT, Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A. 102:2567–2572.

Langille MG, Brinkman FS. 2009. IslandViewer: an integrated interface for computational identification and visualization of genomic islands. Bioinformatics 25:664–665.

Larkin MA, et al. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948.

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

Lombardo MJ, et al. 2007. An *in vivo* expression technology screen for *Vibrio cholerae* genes expressed in human volunteers. Proc Natl Acad Sci U S A. 104:18229–18234.

Louis P, Doré J. 2014. Functional metagenomics of human intestinal microbiome β-glucuronidase activity. In: Nelson KE, editor. Encyclopedia of metagenomics. New York: Springer. p. 1–8.

Magnelli P, Bielik A, Guthrie E. 2012. Identification and characterization of protein glycosylation using specific endo- and exoglycosidases. Methods Mol Biol. 801:189–211.

Mathur J, Waldor MK. 2004. The *Vibrio cholerae* ToxR-regulated porin OmpU confers resistance to antimicrobial peptides. Infect Immun. 72:3577–3583.

Mazel D. 2006. Integrons: agents of bacterial evolution. Nat Rev Microbiol. 4:608–620.

Mazel D, Dychinco B, Webb VA, Davies J. 1998. A distinctive class of integron in the *Vibrio cholerae* genome. Science 280:605–608.

Nesterchuk MV, Sergiev PV, Dontsova OA. 2011. Posttranslational modifications of ribosomal proteins in *Escherichia coli*. Acta Naturae 3:22–33.

O'Shea YA, et al. 2004. The *Vibrio* seventh pandemic island-II is a 26.9 kb genomic island present in *Vibrio cholerae* El Tor and O139 serogroup isolates that shows homology to a 43.4 kb genomic island in *V. vulnificus*. Microbiology 150:4053–4063.

Petrack B, Greengard P, Craston A, Sheppy F. 1965. Nicotinamide deamidase from mammalian liver. J Biol Chem. 240:1725–1730.

Provenzano D, Schuhmacher DA, Barker JL, Klose KE. 2000. The virulence regulatory protein ToxR mediates enhanced bile resistance in *Vibrio cholerae* and other pathogenic *Vibrio* species. Infect Immun. 68:1491–1497.

Purser JE, et al. 2003. A plasmid-encoded nicotinamidase (PncA) is essential for infectivity of *Borrelia burgdorferi* in a mammalian host. Mol Microbiol. 48:753–764.

R Development Core Team. 2014. R: a language and environment for statistical computing. Version 3.1.2. Vienna (Austria): R Foundation for Statistical Computing.

Rashed SM, et al. 2012. Genetic characteristics of drug-resistant *Vibrio cholerae* O1 causing endemic cholera in Dhaka, 2006-2011. J Med Microbiol. 61:1736–1745.

Rashid MU, et al. 2015. *ctxB*1 outcompetes *ctxB*7 in *Vibrio cholerae* O1, Bangladesh. J Med Microbiol. Advance Access published October 19, 2015; doi: 10.1099/jmm.0.000190.

Rasko DA, Myers GS, Ravel J. 2005. Visualization of comparative genomic analyses by BLAST score ratio. BMC Bioinformatics 6:2.

Richter M, Rosselló-Móra R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A. 106:19126–19131.

Rost B. 1999. Twilight zone of protein sequence alignments. Protein Eng. 12:85–94.

Rowe-Magnus DA, Mazel D. 2002. The role of integrons in antibiotic resistance gene capture. Int J Med Microbiol. 292:115–125.

Samant S, et al. 2008. Nucleotide biosynthesis is critical for growth of bacteria in human blood. PLoS Pathog. 4:e37.

Schliep K, Lopez P, Lapointe FJ, Bapteste É. 2011. Harvesting evolutionary signals in a forest of prokaryotic gene trees. Mol Biol Evol. 28:1393–1405.

Schliep KP. 2011. Phangorn: phylogenetic analysis in R. Bioinformatics 27:592–593.

Schmidt H, Hensel M. 2004. Pathogenicity islands in bacterial pathogenesis. Clin Microbiol Rev. 17:14–56.

Seed KD, et al. 2014. Evolutionary consequences of intra-patient phage predation on microbial populations. Elife 3:e03497.

Segata N, Börnigen D, Morgan XC, Huttenhower C. 2013. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat Commun. 4:2304.

Skorupski K, Taylor RK. 1997. Cyclic AMP and its receptor protein negatively regulate the coordinate expression of cholera toxin and toxin-coregulated pilus in *Vibrio cholerae*. Proc Natl Acad Sci U S A. 94:265–270.

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

Stokes HW, Hall RM. 1989. A novel family of potentially mobile DNA elements encoding site-specific gene-integration functions: integrons. Mol Microbiol. 3:1669–1683.

Szamosi JC. 2012. ISAAC: an improved structural annotation of *attC* and an initial application thereof [M.Sc. thesis]. [Hamilton (ON)]: McMaster University.

Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 28:33–36.

Taviani E, et al. 2010. Discovery of novel *Vibrio cholerae* VSP-II genomic islands using comparative genomic analysis. FEMS Microbiol Lett. 308:130–137.

Taylor RK, Miller VL, Furlong DB, Mekalanos JJ. 1987. Use of *phoA* gene fusions to identify a pilus colonization factor coordinately regulated with cholera toxin. Proc Natl Acad Sci U S A. 84:2833–2837.

Tettelin H, et al. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome." Proc Natl Acad Sci U S A. 102:13950–13955.

Vernikos G, Medini D, Riley DR, Tettelin H. 2015. Ten years of pan-genome analyses. Curr Opin Microbiol. 23:148–154.
Vesth T, et al. 2010. On the origins of a *Vibrio* species. Microb Ecol. 59:1–13.
Waldor MK, Mekalanos JJ. 1996. Lysogenic conversion by a filamentous phage encoding cholera toxin. Science 272: 1910–1914.
Wang G, Maier RJ. 2004. An NADPH quinone reductase of *Helicobacter pylori* plays an important role in oxidative stress resistance and host colonization. Infect Immun. 72:1391–1396.
Zhao Y, et al. 2014. PanGP: a tool for quickly analyzing bacterial pan-genome profile. Bioinformatics 30:1297–1299.

Associate editor: Rachel O'Neill

Une grande partie des échanges génétiques entre organismes cellulaires est véhiculée par des éléments génétiques mobiles (MGE) comme des plasmides et des virus [114]. Par conséquent, beaucoup de gènes procaryotes sont susceptibles d'être portés par des MGE. Ces gènes peuvent être maintenus dans la population à cause du fort taux de HGT et malgré l'absence de sélection. De plus, des études ont montré le rôle des MGE dans l'adaptation et la coopération des communautés bactériennes [34,119] ou encore comme potentiel réservoir de diversité [34,120]. Par exemple, Modi et ses collègues ont montré que le phageome associé au microbiote de souris traitées avec des antibiotiques était enrichi en gènes de résistance aux antibiotiques et que les HGT étaient plus fréquents que chez des souris non traitées. Ces études sont en accord avec l'hypothèse des biens publics génétiques [121] qui suggère que certains gènes sont accessibles à l'ensemble des individus d'une communauté. L'ensemble de ces études montre que les cellules procaryotes et les éléments génétiques mobiles sont deux mondes qui s'interpénètrent et que ces derniers sont indispensables à l'évolution des procaryotes. Or, à notre connaissance, aucune étude ne s'est intéressée à la proportion de gènes procaryotes externalisés dans les MGE. Dans l'étude suivante, nous avons utilisé les graphes bipartis pour étudier les flux génétiques entre 8000 génomes de MGE et 400 génomes de procaryotes. Nos résultats montrent que, malgré le faible échantillonnage en MGE, une proportion non négligeable de gènes procaryotes, essentiellement liés au métabolisme, est présente dans le mobilome. Cette étude confirme le rôle primordial des MGE dans l'évolution des procaryotes comme agents de transferts de gènes et suggèrent leurs rôles comme réservoir de diversité et pour l'adaptation des communautés microbiennes.

Dynamics of gene flow in a bipartite web of life

Submitted

E. Corel†, R. Méheust†, J. O. McInerney‡, P. Lopez†, E. Bapteste†*

†Sorbonne Universités, Université Pierre et Marie Curie, Institut de Biologie Paris-Seine, Centre National de la Recherche Scientifique, Unité Mixte de Recherche 7138 Evolution Paris-Seine, 7 quai St Bernard, 75005 Paris, France.

‡Chair in Evolutionary Biology, Michael Smith Building, The University of Manchester, Oxford Rd, Manchester M13 9PL, United Kingdom.

*Corresponding author.

Introductory paragraph:

Complex microbial gene flows affect how we understand virology, microbiology, medical sciences, genetic modification and evolutionary biology. Phylogenies only provide a narrow view of these gene flows: plasmids and viruses, lacking core genes, cannot be attached to cellular life on phylogenetic trees. Using bipartite graphs that connect thousands of gene families with thousands of related and unrelated genomes, we can show that biological evolution is a modular process only partially constrained by cells and vertical inheritance. Gene families are recycled by lateral gene transfer, and evolution abundantly copies gene families between completely unrelated genomes, *i.e.* viruses, plasmids and prokaryotes. In particular among Bacteria, a process of 'gene externalization' takes place where genes are copied to mobile elements, mainly driven by gene function. Bipartite graphs give us a view of vertical and horizontal gene flow beyond classic taxonomy on a single infinitely-expandable graph that goes beyond the cellular Web of Life.

Main Text.

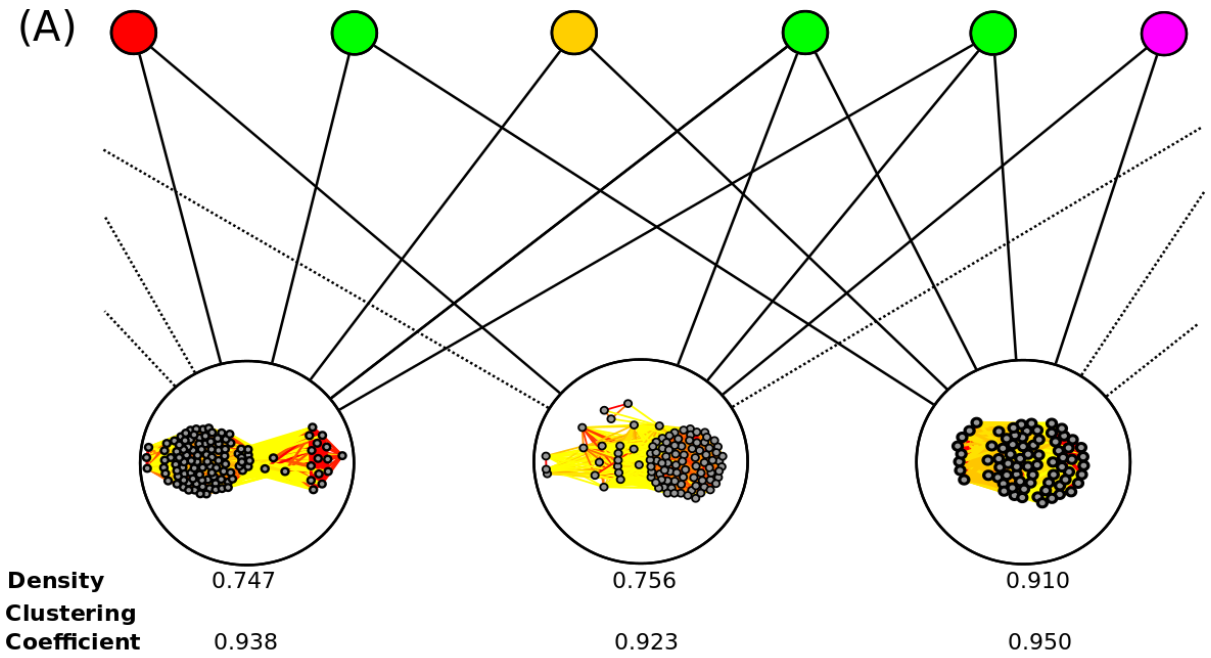
Lateral gene transfer (LGT) from donor to host genome profoundly shapes prokaryotic genome content. It also explains the complex, non-tree like, and at times adaptive, distribution of traits in micro-organisms, while simultaneously hindering prokaryotic systematics. The notion of a web of life, a revolution in evolutionary biology, largely stems from the realization of this complexity [1], though debates continue over the extent of the web of life's reticulations [2–6]. These reticulations represent genetic sharings [7] and occasional mergings [8] between cellular lineages. Outside the web of cellular life, though playing a very central role in its construction and maintenance, are the most abundant DNA carriers on Earth - mobile genetic elements (MGEs), such as viruses and plasmids. Despite this numerical abundance, MGEs lack a core genome and often share no homologies with other MGEs and cellular life-forms. Phylogenies are based on analyses of commonalities and differences and therefore have a limited utility when trying to construct a complete map of the web of life. Consequently, most of the evolving entities on Earth are routinely excluded from descriptions of gene flow dynamics and this surely hinders our understanding of gene movement in the environment and over time. Thus, it is desirable to have a model or structure that can analyze and display gene movements between genomes. Networks are proving useful in this context [1].

Sequence Sharing Network (SSN) approaches have been successfully introduced in order to describe the global structure of the web of life and the spider webs of genetic sharing between MGEs [9,10]. SSNs have also been used to test hypotheses about the phylogenetic or environmental drivers of genomic diversity [11–13] and the selective advantages of introgressed genes [14]. SSNs [15] have the advantage of including all key evolutionary players in a common framework and can offer a global description of gene sharing. However, by adopting a bipartite graph approach we can go further in our description of gene sharing.

Bipartite graphs consist of nodes of two fundamentally different kinds. These nodes are connected by edges such that two nodes on either end of an edge are never of the same kind. Bipartite graphs have been successfully used to explore gene-disease relationships [16], the evolution of malaria parasites [17], recipe ingredient datasets [18] and are used extensively in social media network analysis [19]. In the context of gene flow, bipartite graphs can reveal, for instance, which exact groups of genes are shared exclusively by certain exact groups of genomes, instead of simply showing what groups of genomes share some genes.

We developed a bipartite graph model consisting of gene family nodes on one side, and genome nodes on the other side, with edges connecting genomes with the gene families contained in those genomes (Figure 1, panel A). We modeled a large selection of genomes using this structure and examined the data in order to² statistically and graphically analyze the all-encompassing web of life. Bipartite graphs can be decomposed, and to a certain extent summarized, using numerous measures and approaches [20,21]. However, in this study, we examined two simple, but biologically important patterns - twins and articulation points (Fig. 1 panel B). “Twins” are defined as sets of gene family nodes that have identical connectivity to genome nodes. Articulation points are gene family nodes that are connected to parts of the graph that otherwise share no gene family nodes. Our analyses of these patterns led us to five major observations. First, gene transmission is mainly structured by host type. Therefore, we recover many prokaryotic (cellular) and MGE (acellular) kinds with exclusive gene contents. Second, within these kinds, gene transmission is largely reticulated. Third, many genes are found both on cellular genomes and MGEs, a process that we call 'gene externalization'. Gene externalization between cellular genomes and mobile elements introduces extra-genomic copies of genes in microbial communities by a different process than gene duplication, and contributes to the spread of genetic public goods [22] across these communities. Fourth, the web of life is composed of phylogenetically distinct yet genetically intertwined structured sets of

agents (*i.e.* taxonomically and typologically consistent), permeated by transposases. Fifth, despite their highly mosaic, dynamic nature, plasmid genomes can further be classified into groups of exclusive gene sharing. However, this gene sharing represents only a limited portion of their genomes. These insights can only come from these kinds of networks, which we suggest can be infinitely expanded for all future genetic material to provide a detailed view of the web of life.



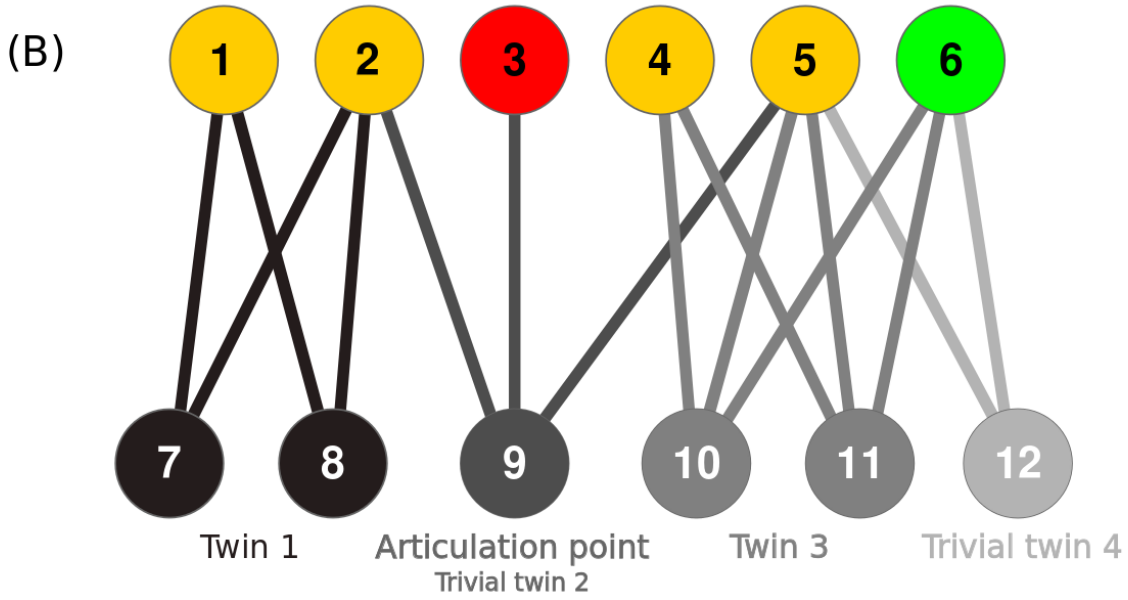


Fig. 1. Construction of the bipartite graphs and identification of the twins and articulation points.

(A) Construction of the genome-gene families bipartite graphs. Top nodes represent genomes of cells and mobile genetic elements. Bottom nodes represent gene families: we display the corresponding connected component of the sequence similarity network (see Methods), edge color (from yellow to red) indicates increasing % ID. Density and clustering coefficient are proxies for the divergence of the family.

(B) Bottom twins and articulation points: bottom nodes forming a twin class and their incident edges are drawn in the same shade of gray. Nodes 7 and 8 have the same neighbors (node 1 and 2) thus form twin class 1. Twins 2 and 4 are trivial since they contain only one node. Node 9 is an articulation point since its removal disconnects the graph.

Results

We initially constructed bipartite graphs from our dataset of 382 prokaryotic genomes and 8,099 mobile element genomes, where the "top" nodes correspond to the genomes and the "bottom" nodes correspond to gene families, defined at various stringencies (Figure 1, panel A). The stringency parameters allowed us to focus, for example, on recent gene family transmissions (*i.e.* when two sequences could be aligned over ≥ 80% of their mutual length, and were ≥ 95% identical in sequence (95% ID for short)). Varying stringency parameters allowed us to consider a variety of evolutionary time scales (see Material and Methods). An undirected edge connecting a top and a bottom node indicated that a member of a gene family was found in a genome. These graphs were explicitly multilevel, showing genes and genomes, and multi-agent, showing cells and MGEs, thus simultaneously informative about gene family evolution and their distributions across a broad range of genomes. The distribution of homologs across taxa should not be understood as the detection of direct gene exchanges,

since intermediate unknown players are very likely (see for instance those discovered lately in Hug *el al.* [23]).

For each network at a given stringency threshold, we first enumerated all its connected components (CCs), *i.e.* all sets of nodes for which there is always an interconnecting path. These CCs represent groups of genomes associated with an exclusive pool of gene families, *i.e.* a gene family found in a CC is by definition absent in any other CC. The recovery of multiple CCs (522 in the graph at ≥ 95% ID and 156 in the graph at ≥ 30%ID) is consistent with the genetic worlds identified by Halary *et al.* [9], albeit with a now much larger dataset and a different network approach. The discrete nature of this graph indicates a discontinuity in gene transmission between genomes belonging to different CCs. This barrier may reflect phylogenetic isolation, ecological isolation, the use of an alternative genetic code or quite simply the non-exhaustive dataset of genes and genomes at our disposal. For example, the CC in Figure S1 illustrates the case of the Spiroplasma phages, which are characterized by the alternative use of the codon UGA to encode Tryptophan instead of "STOP" (*i.e.* the Mycoplasma/Spiroplasma code). The taxonomic homogeneity of this CC suggests that these phages have been exclusively sharing a unique pool of genes, in effect privatizing these genes [22] for their own lineage. Moreover, our bipartite graphs were characterized by the presence of a giant CC (gCC), encompassing 6362 (*i.e.* 80.1%) genomes and 80136 (99%) gene families (at ≥ 30% ID) (Supplementary Table 1). This single gCC include genomes that have no homologous genes in common, yet participate in a giant network of sharings. The ability to reconstruct such a pattern is a significant advantage associated with the use of bipartite graphs.

To understand the co-inheritance of gene families, and also to provide us with a tool for understanding phenotype evolution when phenotypes are not associated with a single monophyletic taxonomic group, we analyzed each CC, including the gCC, at a more fine-

grained level, by enumerating all the twins of bottom nodes (BT) within these connected components (Figure 1, panel A). Twins are nodes with identical sets of neighbors in a graph. BTs represent gene families that are exclusively present in exactly the same set of genomes. Therefore, a BT defines a group of genes that are likely co-transmitted, vertically and/or horizontally, within a club of genomes (themselves not necessarily closely related). Being in the same BT means having the same pattern of inheritance from a common ancestor or *via* LGT, compatible for a subset of genes with some functional linkage. For example, we discovered a BT shared between 3 ruminal bacteria (2 distinct species of Firmicutes and 1 Fibrobacter) and one plasmid genome in our bipartite graph constructed at≥ 90% ID stringency (Supplementary Figure 2). This BT consisted of 3 conserved gene families – a predicted membrane protein, a signal transduction histidine kinase, and a Na^+-driven multidrug efflux pump.

Detecting individual or sets of gene families shared by many genomes with otherwise totally distinct gene contents (at a given similarity threshold) is essential to track long-distance gene transmission across the web of life. This detection is best achieved *via* graph compression. We reduced the bipartite graph by grouping together BT nodes into bottom super-nodes. Notably, the result of this graph reduction is unique and robust, *i.e.* it does not depend from the order in which twins are merged. This merging produced a quotient (*i.e.* BT-free) bipartite graph with no loss of information due to this compression. It is then trivial to enumerate all bottom articulation points (BAPs) in such reduced graphs, *i.e.* all nodes whose removal would increase the number of connected components. Although strictly topological, the notion of BAPs could in principle help to detect public genetic goods [22], *i.e.* genetic material that is being shared by taxonomically distant genomes, which possibly benefit from the properties they confer, for some other reason than genealogy (*i.e.* genes coding for environmental adaptation or others hitch-hiking with them...). We report for instance the case

of the 3'-phosphoadenosine 5'-phosphosulfate sulfotransferase (PAPS reductase)/FAD synthetase gene family, shared by the gram-positive bacterium *Ruminococcus bromii* and the gram-negative bacterium *Fibrobacter succinogenes*, forming a BAP in our graph at ≥ 90% ID. This gene family encodes an enzyme with the rare ability to store two electrons without the need for cofactor or prosthetic groups, which likely enhances the success rate of transfer for this gene family in the rumen (Supplementary Fig. 2). At a stringency of ≥ 90% ID, 56 BAP nodes (out of 811 BAP nodes) encompass transposases which, as the graph suggests, possess the capability to move across distantly related genomes.

Simple graph patterns in a gene family-genome network are already sufficient to provide abundant biological information. Detecting recurrent patterns in the compressed bipartite graphs provide novel knowledge about gene transmission in this dataset of 382 prokaryotic genomes and 8099 mobile elements (viruses and plasmids). More precisely, CC and BT analyses (Supplementary Tables 1, 2, 3 and 4) suggest several rules of gene transmission.

First, the vast majority of CCs and BTs were composed of genomes consistent by type of hosts, for all stringency thresholds. For example, at ≥ 90% ID, 94.26% of the CCs, and 92.01% of BTs showed gene sharing between genomes of the same type (*i.e.* either exclusively cellular, exclusively viral or exclusively plasmid). In addition, not only were the vast majority of CCs and BTs consistent with genome type, but the constituent genomes were also generally taxonomically consistent (Supplementary Tables 1 and 2), for all stringency thresholds. For example, at ≥ 90% ID, 78,5 % of the CCs, and 99% of the BTs that contain prokaryotes (*i.e.* 19% of all BTs) showed gene sharing among members of the same phylum (as defined independently by the NCBI taxonomy). Overall, this strong taxonomic signal reflects the fact that distantly related genomes have rather different gene contents, which is

consistent with the relatively independent evolution of various kinds of cellular organisms in the web of life [9].

Second, within these major taxonomic types, evolution seems largely reticulated. We verified this by computing, for each group of closely related genomes in this dataset, the number of gene families that are shared by all members of this group and exclusively by them. The size of these Exclusively Shared Gene families (ESG) amounts to the percentage of BTs (*i.e.* sets of exclusive gene families) found in all genomes belonging to the taxonomic group. This measure would be 100% if exclusive gene families were present in all members of the group in the dataset, *e.g.* in all 15 Methanococci. The size of these ESG is typically small (Suppl. Fig. 3 and 4): at most 35 % for the 15 Methanococci, and typically less than 7%/10%/15% for the 48 Thermoprotei, 16 Cyanobacteria, and 25 Halobacteria contained in our dataset. Thus, most BT are not associated with all genomes of a taxonomic group, consistent with a high turnover of gene families, at least of exclusive core genes, within prokaryotic genomes.

Third, since our networks encoded exact information about which genomes shared which gene families, we were able to quantify the extent of 'gene externalization', that is, of sharing between cellular genomes and MGE genomes (*e.g.* when a given family is connected to two genomes of different kinds). Gene externalization differs from LGT between cellular genomes, although it can contribute to LGT when a gene from a cellular genome is copied to a MGE and from that MGE to another cellular genome. The difference between gene externalization and LGT means that rules relating to gene externalization may differ from rules relating to LGT. In particular, gene externalization may be random and at a high rate, which would not be visible from LGT analyses, if the host recipient cell selects against the residency of some of the externalized genes (*i.e.* for example, informational genes may be

more externalized than transferred). We observed an impressive proportion of externalized genes in the web of life (Figure 2, Suppl. Table 5).

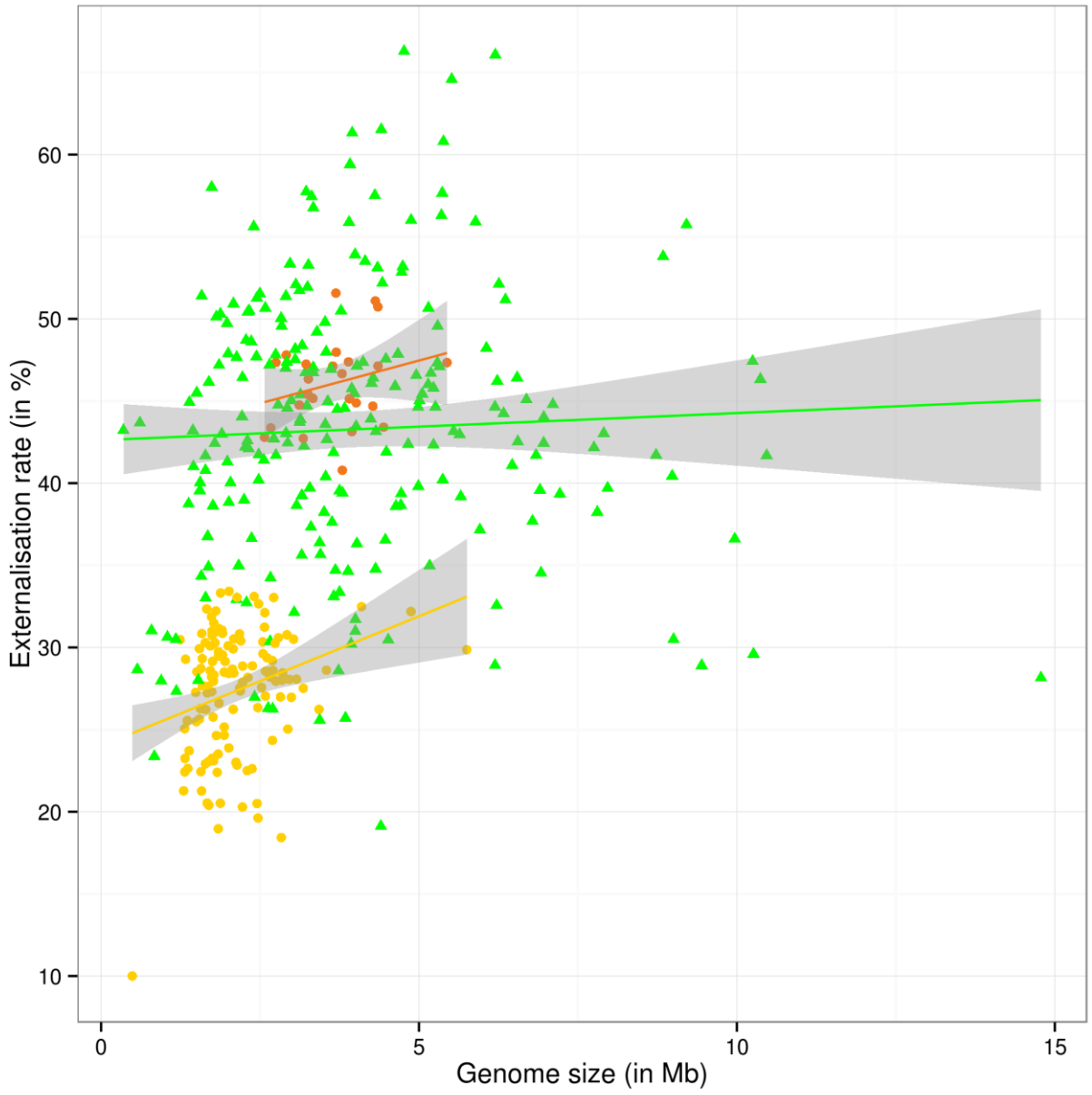


Fig. 2. Percentage of gene externalization at ≥ 30% ID for the 382 prokaryotes in our dataset. The externalization rate of Bacteria (blue triangles) is significantly higher (Student *t*-test, p-value < 10^{-16}) than for Archaea (green dots), to the notable exception of Haloarchaea (red dots). It is moreover largely uncorrelated with genome size (regression lines with shaded confidence interval at 95%).

For our dataset, Bacteria generally have higher externalization than Archaea (Significant *t*-test for ≥ 30%, 40%, 50% and 60% ID). A notable exception to this rule are the Haloarchaea, which is likely explained by their chimeric nature (Suppl. Table 5). For example, genomes of *Starkeya novella DSM 506, Rhodospirillum rubrum ATC 11170* and *Mesorhizobium australicum WSM 2073* have a very high proportion of externalized genes (higher than 60% at ≥ 30% ID, see Figure 2), and so do Haloarchaea (higher than 40% at ≥ 30% ID).We also followed the dynamics of gene externalization for networks of decreasing stringency (assuming a molecular clock, from the most recent to the most ancient externalization events) in order to identify some of externalization rules and to test whether gene externalization is random with respect to gene function.

Gene externalization is not random with respect to gene function. The distributions of the COG categories associated with externalized genes were markedly different, and these differences persisted at different similarity thresholds (Figure 3). The 'L' category was abundant amongst recently externalized genes, suggesting that transposases were amongst the recent gene families that have moved across different types of host genome. However, genes from that 'L' category do not tend to accumulate in their externalized form in genomes, as evidenced by the way their proportion dropped in graphs with lower %ID. This indicates that these transposases do not persist in their host genomes. By contrast, the proportion of externalized genes from the 'M' (membrane biogenesis) and 'T' (Signal transduction) categories was smaller for recent events than for events considered over a longer time period (*i.e.* genes from these categories tended to accumulate progressively in genomes as externalized). Other functional categories, such as 'E' (Amino-acid metabolism and transport) and 'P' (Inorganic ion metabolism and transport) presented more complex distributions.

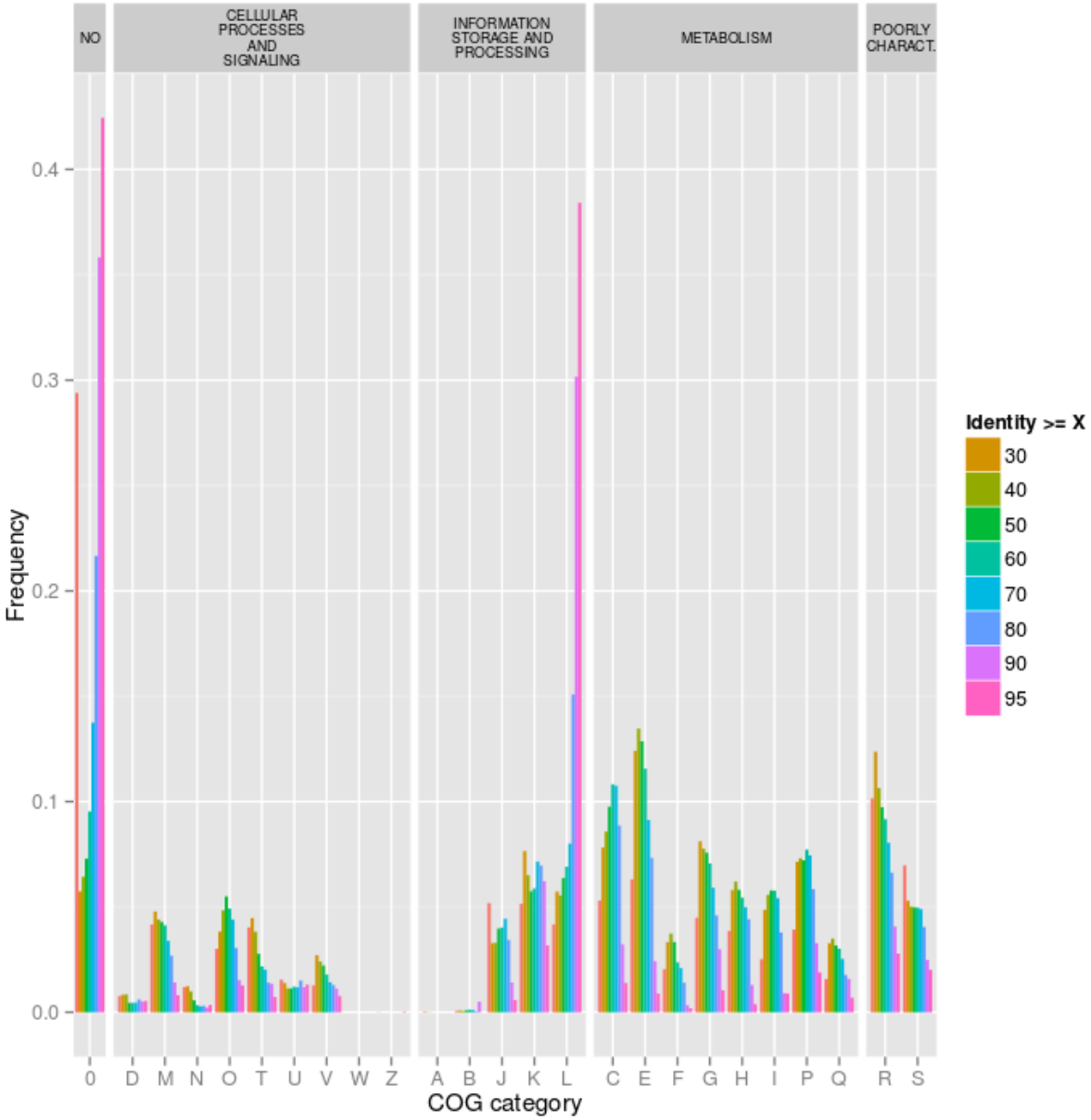


Fig. 3. Distribution of functional categories among externalized genes.

Color bars represent the percentage of a given COG category among externalized genes above a given identity threshold (according to the color code on the right side of the figure).

On the upper bar, COG categories are grouped by large functional groups (including 'Poorly characterized', which includes COG categories R and S). The 'No'/0 class (on the left) refers to the genes for which no COG category was attributed). Note the very conspicuous peak for the 'L' category at thresholds ≥ 90% and 95% ID.

Due to gene externalization, the web of life appears, in its prokaryotic parts, as a collection of largely disconnected, isolated, prokaryotic strands, affected by introgression between close relatives, doubled by spider-webs of mobile elements. One should therefore not be misled by the taxonomical consistency of CCs and BTs. Phylogenetically consistent prokaryotic groups are typically subjected to and sustained by processes that are not simple vertical descent with modification. Gene externalization, from cells to mobile genetic elements, and from mobile genetic elements to cells, may contribute to the high turnover of genes in genomes and their patchy distribution in prokaryotic lineages. The high levels of gene externalization that we report indicate that, collectively, genomes of MGE contain most (possibly all) gene families from several individual bacterial genomes, which are present, dispersed through fragmented copies, in the unrelated genomes of viruses and plasmids. We predict that as more MGE genomes are sequenced, the percentage of externalized genes per prokaryotic genome will increase, albeit at different rates for different biological functions.

Finally, gene externalization can play a direct role in the mobilization of public genetic goods. Our graph also provided evidence for this process, presenting some BTs that correspond to the sharing of gene families between genomes, with potential adaptive content [24]. In the graph at ≥ 90%ID, our very discrete sampling of genomes contained 20 BT distributed on genomes from different phyla. In the graph at ≥ 30%ID, including more ancient sharing events, there were 12,864 BT (*i.e.* 30,9% of all BT) grouping genomes from different phyla. Such taxonomically heterogeneous BTs point to candidate genetic public goods, transferred over large phylogenetic distances, *i.e.* since these sequences are used by phylogenetically heterogeneous hosts. For example, Twin 7227 is a gene family involved in cell wall - peptidoglycan - lysis. The protein is found in viruses and bacteria and is important in degrading the cell wall - either for the purposes of infecting a bacterium or for cell division. This kind of "cell puncturing device" is likely to enhance horizontal transfer. Twin 3034 is the

LexA protein, which in purified form acts as a repressor of RecA and itself. This protein can function to reduce the level of recombination and SOS-mediated response from an organism. The SOS response is triggered by DNA damage, as is RecA. Therefore the function of this twin seems to be to repress recombination and to stop DNA repair processes which might prevent the integration of a sequence into a genome. Other interesting examples stem from these analyses. Twin 7401 at ≥ 90% ID corresponds to a particular prokaryotic compartment involved in the carbon fixation from atmospheric CO_2 called the carboxysome [25], shared by taxonomically divergent bacteria (2 Cyanobacteria and 2 Gammaproteobacteria). The carboxysome is also present as twin 69 (under a sufficiently divergent form as to make a different gene family): this time it is even an articulation point linking 1 Bacteroidetes, 1 Chloroflexi and 1 Actinobacterium. We also find conspicuous plant nodule associated genes: twin 1436 is a nitrogenase subunit NifH forming a twin for a club of three nodule associated Alphaproteobacteria, and twin 7710 is an articulation point, with a dehydrogenase function, between one Acidobacterium (*Candidatus* Solibacter usitatus) and 2 nodule Alphaproteobacteria (*Methylocella silvestris* and *Mesorhizobium australicum*). The removal of the articulation point neatly separates the three according to taxonomy, and seems ecologically driven since this Acidobacterium actually lives in soil [26] and has a large number of genes associated with MGEs [27]. Public goods however are not the only genes that can be shared so broadly. Twin 13016 is a toxin-antitoxin system, a famous “addiction” system. Both genes are needed in the genome in order to function. In general, the toxin is long-lived and the antitoxin is short-lived, and keeps the cell safe from the toxin by binding to it. When the genes are removed from the cell, then the short-lived antitoxin breaks down, leaving the toxin to kill the cell. This mechanism removes cells that have been cured of the toxin-antitoxin system, providing an advantage to those cells that have both genes.

We also observed the diffusion of other so-called 'selfish' genes. In general, transposases were broadly distributed over MGE and cellular genomes, as expected according to *e.g.* Aziz and al. [28]. In the overall graph at ≥ 90% ID, 4.78% of the CCs and 8.03% of the BTs were annotated as containing a transposase, respectively. Interestingly, transposases were over-represented in BTs mixing different types of genomes, because some transposases travel across different host genomes. Thus, transposases were found in 7.21% (888 out of 12,321) of the BTs joining the same type of genomes but in 17.94% (192 out of 1,070) of the BT joining different types of genomes (*e.g.* any combination of virus, plasmids or cellular genomes). Likewise, transposases were found in 3.04% (14 out of 460) of the CCs joining the same type of genomes, but in 21.43% (6 out 28) of the CCs that joined different types of genomes. In about 1/6 of these BTs with heterogeneous phyla, other gene families hitch-hiked with these transposases. Thus transposases are actively travelling across the web of life, but they do not organize it (*i.e.* removing annotated transposases from the analyses does not substantially change the topology of the bipartite graph).

Introgressive evolution has likewise shaped mobile genetic elements – as can be seen first in the sharings of very similar genes between viruses and plasmids within the mobilome network (*i.e.* 8 CCs (out of 488) and 107 BTs (out of 13,391) mixing viruses and plasmids at ≥ 90%ID). The impact of introgression is particularly clear in the exclusive gene sharing between plasmid genomes (Suppl. Fig. 5 and 6).

Discussion

In bipartite graphs, BT and BAP can be used to extract information of gene and genome evolution, gene externalization and transmission. Because these processes produce detectable imprints in these networks, it is possible to detect and analyze them in order to represent the structure of the web of life, and gain insight about its dynamics. We report a

disconnected web of life with major prokaryotic kinds, largely but not absolutely isolated from one another in terms of gene sharing, surrounded by spider-webs of mobile genetic elements that carry extra-genomic gene copies. Moreover, large portions of the web of life are regularly infiltrated at high speed by transposases. Overall, biological evolution appears to be a modular process, which, in addition to lateral gene transfer between cellular life forms, recycles gene families and copies them between unrelated genomes. Bipartite graphs, describing the dynamics of gene families across related and unrelated genomes, are a powerful novel way to categorize this process of microbial life beyond classic taxonomy and customary genomic analyses. Beyond coding gene families, bipartite graphs be further generalized to small RNA families-genomes networks, and gene families-metagenome networks, and applied to even larger datasets to keep up with the impressive accumulation of molecular sequences. Ultimately, gene families-genomes-metagenomes tripartite graphs constitute an exciting horizon for expanded multilevel analyses. It would be fascinating to obtain an even bigger and more accurate picture of the dynamics of biological complexity.

Materials and methods

Data collection. We have downloaded all complete genomes for viruses, plasmids and Archaea available as of Nov. 2013 from the NCBI, as well as one complete genome from each bacterial family, in order to compensate for the sampling bias towards Bacteria in the available genomic databases. In this way, we obtained 230 Bacteria, 152 Archaea, 4350 plasmids and 3749 viruses (see Supplementary Excel files).

BLAST sequence analysis. We ran a BLAST all-against-all (`blastp` version 26, E-value 10^{-5}) on all the resulting 1,151,260 protein coding sequences. We filtered the returned hits by keeping only the best hit between two sequences, whenever the corresponding matching length covered at least 80% of both sequences.

Bipartite graph generation. For a given % ID, we constructed gene families as the non-trivial (*i.e.* of cardinality at least 2) connected components of the following graph: nodes are protein-coding sequences, and an edge is drawn between two nodes if the sequences have ≥ 80% best reciprocal cover and the returned identity percentage is at least the required % ID (Fig. 1 panel A). We constructed the gene families-genomes bipartite graphs by taking as top nodes the genomes and as bottom nodes the gene families, an edge connecting a genome node to a gene family node if the genome contained at least one member of the gene family (Fig. 1 panel B). We further simplified this bipartite graph by removing all bottom nodes having degree 1. We detected BTs and BAP using custom Python computer code (available upon request).

Plasmid heatmaps. The heatmaps for the plasmid genomes were constructed by applying the same procedure to the dataset consisting only of plasmid protein-coding genes. The underlying matrix $M=(M_{ij})$ is indexed by plasmids as rows and twins as columns. The element $M_{ij}=t_i/s_j$ is equal to the ratio of the number t_i of gene families comprising twin i over the number s_j of gene families having one representative in plasmid j (used as a proxy of the size of the plasmid j), and 0 if twin i is not present in plasmid j. The hierarchical clustering was performed on the matrix whose entry (i,j) is 1 if twin i is present in plasmid j and 0 otherwise, by using R's `hclust` procedure with the Euclidean metric and Ward's D2 agglomerative algorithm.

References

1. Corel, E., Lopez, P., Méheust, R. & Bapteste, E. Network-Thinking: Graphs to Analyze Microbial Complexity and Evolution. *Trends Microbiol.* (2016). doi:10.1016/j.tim.2015.12.003
2. Kunin, V., Goldovsky, L., Darzentas, N. & Ouzounis, C. A. The net of life:

reconstructing the microbial phylogenetic network. *Genome Res.* **15,** 954–959 (2005).

3. Nelson-Sathi, S. *et al.* Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* **517,** 77–80 (2015).
4. Dagan, T. & Martin, W. The tree of one percent. *Genome Biol.* **7,** 118 (2006).
5. Smillie, C. S. *et al.* Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480,** 241–4 (2011).
6. Beiko, R. G., Harlow, T. J. & Ragan, M. A. Highways of gene sharing in prokaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 14332–14337 (2005).
7. Skippington, E. & Ragan, M. A. Lateral genetic transfer and the construction of genetic exchange communities. *FEMS Microbiol. Rev.* **35,** 707–735 (2011).
8. Nelson-Sathi, S. *et al.* Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proc. Natl. Acad. Sci.* **109,** 20537–20542 (2012).
9. Halary, S., Leigh, J. W., Cheaib, B., Lopez, P. & Bapteste, E. Network analyses structure genetic diversity in independent genetic worlds. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 127–32 (2010).
10. Alvarez-Ponce, D., Lopez, P., Bapteste, E. & McInerney, J. O. Gene similarity networks provide tools for understanding eukaryote origins and evolution. *Proc. Natl. Acad. Sci.* **110,** E1594–E1603 (2013).
11. Kloesges, T., Popa, O., Martin, W. & Dagan, T. Networks of Gene Sharing among 329 Proteobacterial Genomes Reveal Differences in Lateral Gene Transfer Frequency at Different Phylogenetic Depths. *Mol. Biol. Evol.* **28,** 1057–1074 (2011).
12. Forster, D. *et al.* Testing ecological theories with sequence similarity networks: marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms. *ISME J.* **13,** 1–16 (2014).

13. Cheng, S. *et al.* Sequence similarity network reveals the imprints of major diversification events in the evolution of microbial life. *Front. Ecol. Evol.* **2,** 1–13 (2014).

14. Bapteste, E. The origins of microbial adaptations: How introgressive descent, egalitarian evolutionary transitions and expanded kin selection shape the network of life. *Front. Microbiol.* **5,** 1–4 (2014).

15. Atkinson, H. J., Morris, J. H., Ferrin, T. E. & Babbitt, P. C. Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies. *PLoS One* **4,** e4345–e4345 (2009).

16. Hwang, T. *et al.* Robust and efficient identification of biomarkers by classifying features on graphs. *Bioinformatics* **24,** 2023–2029 (2008).

17. Larremore, D. B., Clauset, A. & Buckee, C. O. A network approach to analyzing highly recombinant malaria parasite genes. *PLoS Comput. Biol.* **9,** e1003268–e1003268 (2013).

18. Ahn, Y.-Y., Ahnert, S. E., Bagrow, J. P. & Barabási, A.-L. Flavor network and the principles of food pairing. *Sci. Rep.* **1,** 196 (2011).

19. Murata, T. Detecting Communities from Bipartite Networks Based on Bipartite Modularities. in *2009 International Conference on Computational Science and Engineering* **4,** 50–57 (IEEE, 2009).

20. Barber, M. J. Modularity and community detection in bipartite networks. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.* **76,** 66102 (2007).

21. Alzahrani, T. & Horadam, K. J. *Complex Systems and Networks*. *Understanding Complex Systems* **73,** (Springer Berlin Heidelberg, 2016).

22. McInerney, J. O., Pisani, D., Bapteste, E. & O'Connell, M. J. The public goods hypothesis for the evolution of life on Earth. *Biol. Direct* **6,** 41 (2011).

23. Hug, L. A. *et al.* A new view of the tree of life. *Nat. Microbiol.* **1,** 16048 (2016).

24. Karcagi, I. *et al.* Indispensability of horizontally transferred genes and its impact on bacterial genome streamlining. *Mol. Biol. Evol.* (2016). doi:10.1093/molbev/msw009

25. Yeates, T. O., Kerfeld, C. A., Heinhorst, S., Cannon, G. C. & Shively, J. M. Protein-based organelles in bacteria: carboxysomes and related microcompartments. *Nat. Rev. Microbiol.* **6,** 681–691 (2008).

26. Challacombe, J. F. *et al.* Biological Consequences of Ancient Gene Acquisition and Duplication in the Large Genome of Candidatus Solibacter usitatus Ellin6076. *PLoS One* **6,** e24882–e24882 (2011).

27. Challacombe, J. & Kuske, C. Mobile genetic elements in the bacterial phylum Acidobacteria. *Mob. Genet. Elements* **2,** 179–183 (2012).

28. Aziz, R. K., Breitbart, M. & Edwards, R. A. Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res.* **38,** 4207–4217 (2010).

29. Smillie, C., Garcillan-Barcia, M. P., Francia, M. V, Rocha, E. P. C. & de la Cruz, F. Mobility of Plasmids. *Microbiol. Mol. Biol. Rev.* **74,** 434–452 (2010).

30. Lanza, V. F. *et al.* The Plasmidome of Firmicutes: Impact on the Emergence and the Spread of Resistance to Antimicrobials. *Microbiol. Spectr.* **3,** PLAS–0039–2014 (2015).

31. Lima-Mendez, G., Van Helden, J., Toussaint, A. & Leplae, R. Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol. Biol. Evol.* **25,** 762–777 (2008).

Acknowledgements We thank D. Bhattacharya and H. Le Guyader for reading the manuscript and critical comments. E.C. and E.B. are funded by FP7/2007-2013 Grant Agreement #615274.

Author contributions E.C., P.L. and E.B. designed the study, R.M. collected and processed the data, E.C. performed the analyses, E.B. wrote the paper, all authors discussed the results and commented on the manuscript.

Additional information The complete sequence data and bipartite graphs are available as a tarball at the following URL: `http://www.evol-net.fr/index.php/fr/downloads`. Reprints and permissions information is available at `www.nature.com/reprints`. The authors declare no competing financial interests. Correspondence and request for materials should be addressed to E.C. (eduardo.corel@upmc.fr).

2. Transfert de gènes horizontaux chez les eucaryotes

a) Acquisitions épisodiques par EGT

Alors que l'importance des HGT dans l'évolution des procaryotes est bien établie, les HGT ont eu des rôles différents chez les eucaryotes [122]. Chez les procaryotes, les HGT entre organismes sont fréquents et continus [123], c'est le processus majoritaire pour acquérir de nouveaux gènes. Le contraire est observé chez les eucaryotes. Bien qu'il soit sans doute encore trop tôt pour l'affirmer, la majorité des eucaryotes ne semble pas posséder de pan-génomes, ou alors dans une moindre mesure comme chez *Emiliania huxleyi* et *Micromonas* [124,125], alors que chez les procaryotes le pan-génome a été érigé au rang de paradigme [126]. Chez les eucaryotes, les familles multigéniques sont majoritairement dues à de véritables duplications de gènes et rarement à des HGT [127,128].

En somme, les HGT ont un impact moindre sur l'évolution des génomes eucaryotes, du moins comme processus continu puisque de nombreux gènes présents chez des organismes eucaryotes possèdent des homologues procaryotes et ont été hérités d'eux. La Figure 12 montre que les familles eucaryotes héritées de procaryotes sont généralement présentes chez plusieurs supergroupes eucaryotes et sont rarement spécifiques de lignées eucaryotes. *A contrario,* les familles ne possédant pas d'homologues chez des procaryotes et qui sont interprétées comme des familles de gènes nouvellement créées chez les eucaryotes montrent une distribution différente (Figure 13) : la majorité des familles sont spécifiques de lignées et très peu de familles sont conservées dans plusieurs lignées. Les familles spécifiques des eucaryotes semblent donc avoir été acquises de façon continue alors que les familles eucaryotes héritées des procaryotes montrent un patron d'acquisition épisodique et massif qui coïncide avec les acquisitions des ancêtres de la mitochondrie et du chloroplaste. Comme expliqué précédemment (voir les sections II.B.3 et II.C.1), le processus d'intégration des deux endosymbiontes a impliqué de nombreux EGT du génome des endosymbiontes vers le génome nucléaire [70]. Les nombreuses familles présentes uniquement chez des eucaryotes photosynthétiques possèdent de nombreux homologues chez les cyanobactéries (Figure 12) et proviennent d'EGT du génome chloroplastique vers le génome nucléaire dans le cas des *Archaeaplastida* et même d'EGT du génome du nucléomorphe vers le génome nucléaire dans le cas des organismes ayant acquis secondairement la photosynthèse [73,81]. Le cas des EGT mitochondriaux est plus complexe : seule une faible partie des gènes d'origine bactérienne montre une origine claire associée à la lignée des alpha-protéobactéries. Cette observation a

poussé les chercheurs à proposer des hypothèses pour expliquer cette absence de signal phylogénétique cohérent. Parmi ces hypothèses, on trouve le mosaicisme hérité [129], le remplacement de gènes ou encore l'acquisition de HGT par différents donneurs ou par plusieurs partenaires symbiotiques [66].

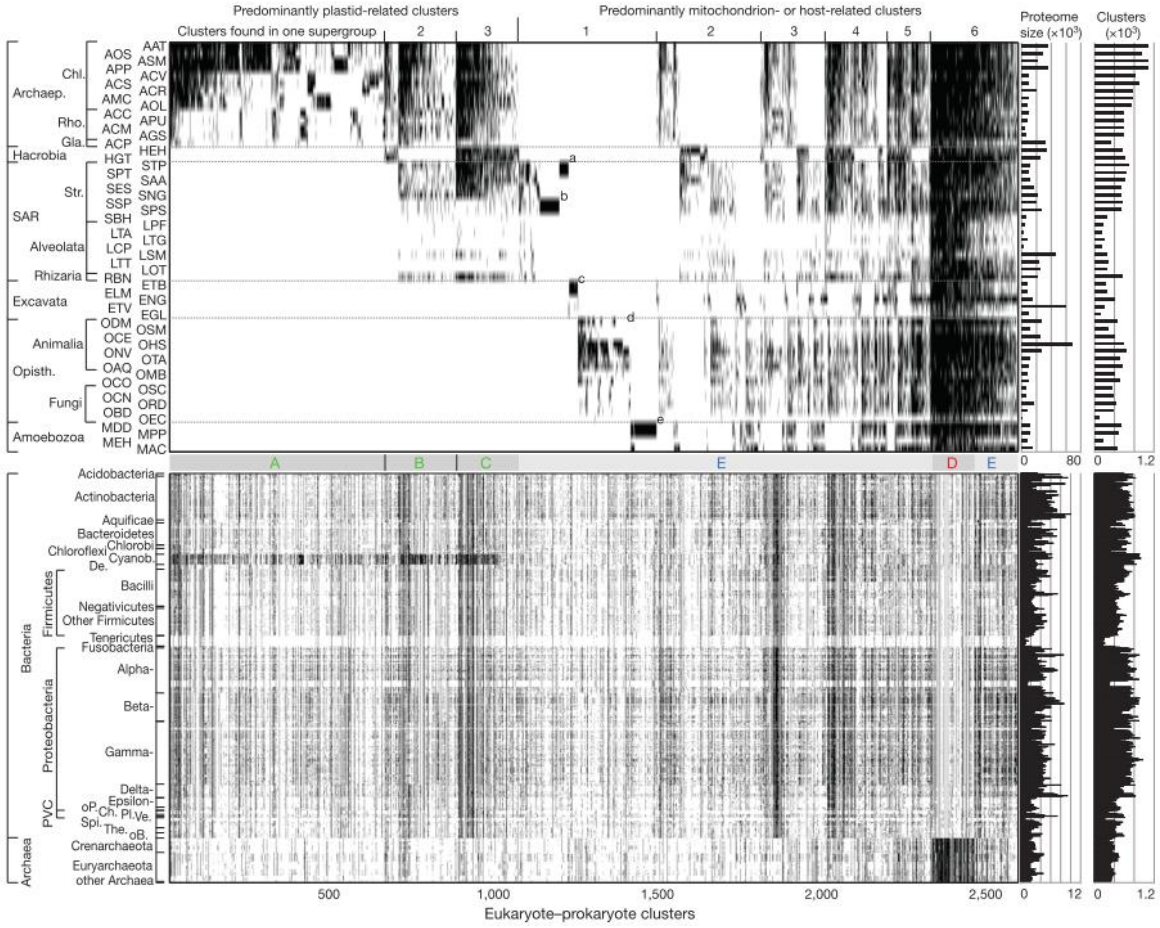


Figure 12. Distribution des familles de gènes présentes dans les génomes eucaryotes et possédant des homologues chez les procaryotes (source : [70]**)**

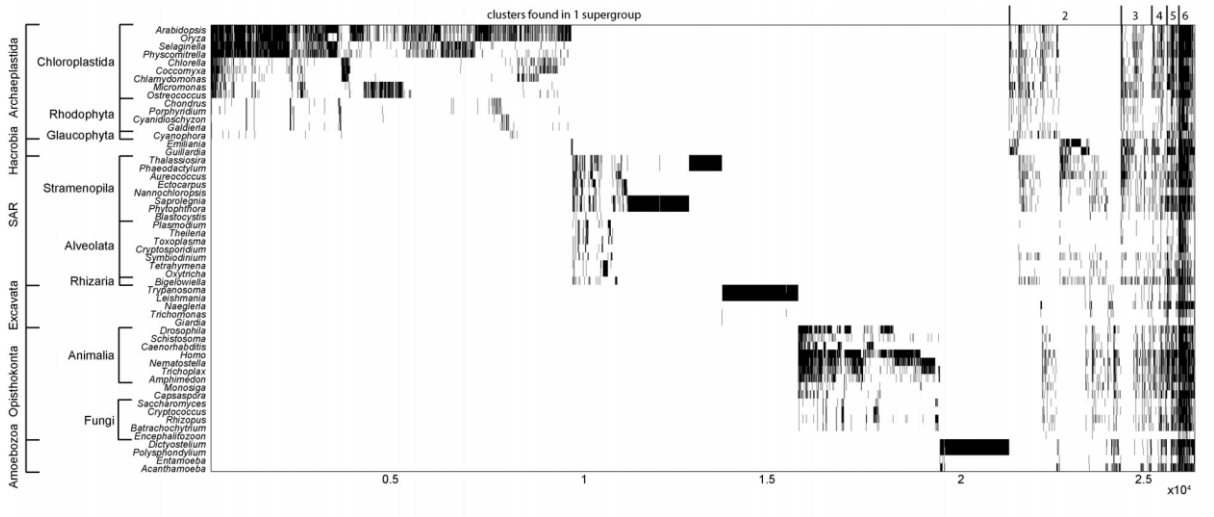


Figure 13. Distribution des familles de gènes présentes dans les génomes eucaryotes et ne possédant pas d'homologues chez les procaryotes (source : [70]**)**

b) Acquisitions continues de gènes par HGT

En dehors des EGT, le transfert horizontal de gènes comme processus continu d'acquisition de gènes est beaucoup plus contesté. Plusieurs transferts de gènes bactériens vers certaines lignées eucaryotes sont évidents comme le transfert d'une partie du génome de *Wolbachia* dans le chromosome X de la bruche chinoise [130], les 22 gènes bactériens transférés vers le génome de la cochenille dans la symbiose emboitée entre une cochenille et deux bactéries [29] ou encore l'acquisition de deux gènes d'origine virale ayant un rôle dans l'évolution du placenta [131]. Les transferts de gènes entre un symbionte bactérien (ou viral) et son hôte eucaryote sont assez bien reconnus mais leurs importances quantitatives restent assez dérisoires. Une étude conservative chez les eucaryotes unicellulaires et parasitiques estime que seulement 1% des gènes des génomes des parasites ont été affectés par des HGT [132,133].

Plus controversés encore sont les HGT ayant eu lieu à la racine de grands groupes eucaryotes ou comme processus d'acquisition continue de nouveaux gènes [70,134]. Plusieurs études ont suggéré l'importance des HGT chez le groupe des Fungi [135], dans l'émergence du pouvoir pathogène des Oomycètes [136–139] ou encore dans la transition des plantes d'un environnement aquatique vers un environnement terrestre [140]. Ces études sont souvent controversées car quelques dizaines de gènes sont détectés comme provenant de HGT mais aucun symbionte pouvant expliquer ces acquisitions n'est connu. Par exemple, plusieurs études suggèrent que l'intégration du chloroplaste chez les eucaryotes photosynthétiques s'est faite grâce à l'aide d'un troisième partenaire symbiotique : une bactérie intracellulaire du

groupe des *Chlamydiae* [141]. Plusieurs dizaines de gènes présents chez les eucaryotes photosynthétiques branchent dans des arbres phylogénétiques avec le groupe des *Chlamydiae* et ils ont été suggérés comme provenant d'un troisième partenaire symbiotique [142–145]. Or des analyses récentes suggèrent le contraire [134,146]. Le mystère reste donc entier [147] mais il souligne surtout la difficulté de détecter des HGT chez les eucaryotes lorsque ces événements sont anciens et non massifs.

B. Acquisition autogène de nouveaux gènes chez les eucaryotes

1. Plusieurs mécanismes pour créer de nouveaux gènes

La faible fréquence des HGT chez les eucaryotes en comparaison avec les procaryotes ne signifie pas qu'aucun gène n'est apparu depuis l'eucaryogenèse. D'ailleurs, la majeure partie des gènes présents dans les génomes eucaryotes sont spécifiques de la lignée eucaryote (Figure 13) [148,149]. Ces nouveaux gènes n'ont pas nécessité l'apport de matériel génétique extérieur, ils ont utilisé le matériel génétique existant dans le génome. Par exemple en dupliquant des gènes préexistants [150] (Figure 14.b), en transformant une région non codante en région codante (création *de novo* de gène) [151] (Figure 14.g) ou encore en recombinant des morceaux génétiques distincts donnant naissance à des gènes composites [152]. Ces derniers sont particulièrement intéressants dans le cadre de cette thèse car ils impliquent des phénomènes non arborescents.

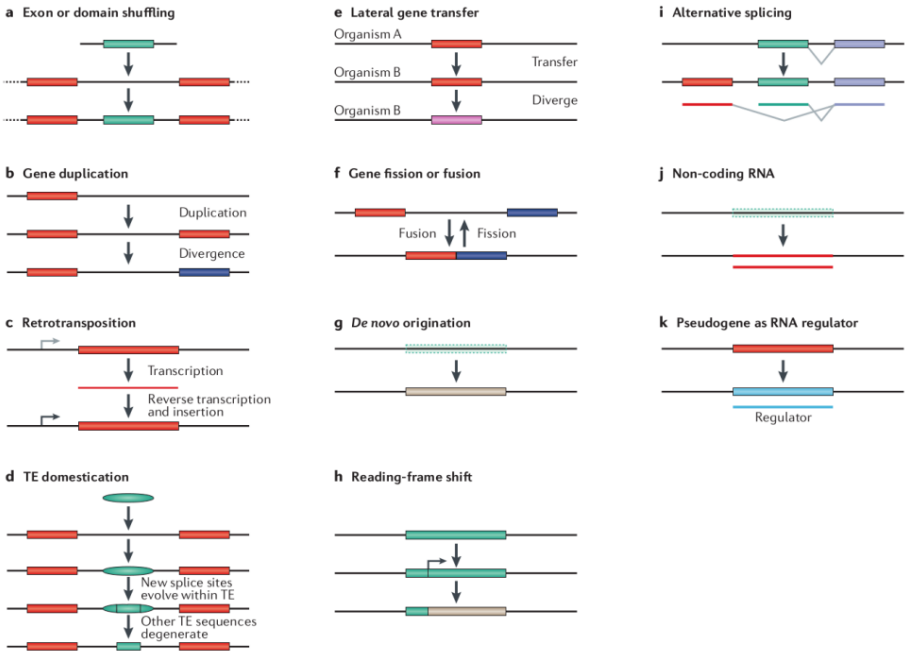


Figure 14. Plusieurs mécanismes de création de nouveaux gènes (source : [153]**)**

2. Les gènes composites

On appelle gènes composites les gènes créés avec au moins deux fragments génétiques appartenant à des familles de gènes différentes. De nombreux termes ont été utilisés pour désigner ce type de gènes, gènes chimériques [154], gènes de fusion [155], gènes codant pour des protéines multi-domaines [156] ou encore gènes composites [157]. Pour plus de clarté, nous utiliserons par la suite le terme de gène composite ou gène chimérique. Les fragments composant un gène composite sont appelés les composantes du gène composite (Figure 15).

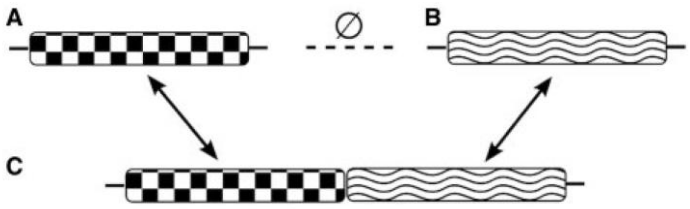


Figure 15. Un gène composite (C) et ses deux composantes (A et B). A et B ne sont pas similaires (source : [158]**)**

a) Mécanismes de création de gènes chimériques

Il ne semble pas exister de mécanisme type pour la création de gènes chimériques, les mécanismes sont nombreux et variés : le brassage d'exons et de domaines [159] (Figure

14.a), la fusion de gènes par suppression d'un codon stop et du signal de terminaison de la transcription [160] (Figure 14.f), l'ajout *de novo* d'une nouvelle région [161], l'insertion dans un gène d'un élément génétique mobile [162] (Figure 14.d) ou d'une "rétrocopie" de gène [163], une altération dans le processus d'épissage alternatif entre gènes récemment dupliqués a aussi été observée [164]. En plus de ces créations permanentes de gènes, il a été récemment montré que de l'épissage entre des exons de gènes adjacents pouvait aboutir à la création de nouvelles unités de transcriptions chimériques et temporaires [103]. Tous ces mécanismes permettent de bricoler [165] de nouvelles structures de protéines, parfois grâce à une combinaison de plusieurs de ces mécanismes comme pour le gène *jingwey*.

b) L'exemple du gène Jingwey chez la drosophile

Au début des années 1990, un des premiers gènes chimériques fut décrit chez des espèces de drosophiles africaines [166]. La récence de ce gène ainsi que le bon échantillonnage des espèces de drosophiles en fait sans doute le gène chimérique le mieux caractérisé [167] (Figure 16).

Le gène *jingwey* fut d'abord caractérisé comme une "rétrocopie" du gène *Adh* qui code pour une alcool déshydrogénase avant d'être re-caractérisé comme un authentique nouveau gène chimérique chez des drosophiles africaines [166]. En effet, le gène *jingwey* possède une extension exonique en 5' par rapport au gène *Adh* qui fut apportée par le gène *yellow emperor* [168]. L'ancêtre commun de *Drosophila yakuba* et *D. teissieri* possédait une copie du gène *yellow-emperor* (*ymp*) et une copie du gène *Adh*. Le gene *ymp* fut dupliqué en deux copies, une appelée *ymp* et une autre appelé *yande* (*ynd*). Alors que la copie *ymp* conserva sa fonction, la copie *ynd* fut à l'origine de la création du gène chimérique *jingwey*. Peu de temps avant la spéciation, l'ARNm de l'*Adh* a subi une retrotransposition dans le troisième intron du gène *yande* puis la "rétrocopie" a recombiné avec les trois premiers exons du gène *yande* aboutissant à la création du gène *jingwey*. L'insertion du retrotransposon entraina la dégénérescence des 9 exons terminaux du gène *yande* (Figure 16).

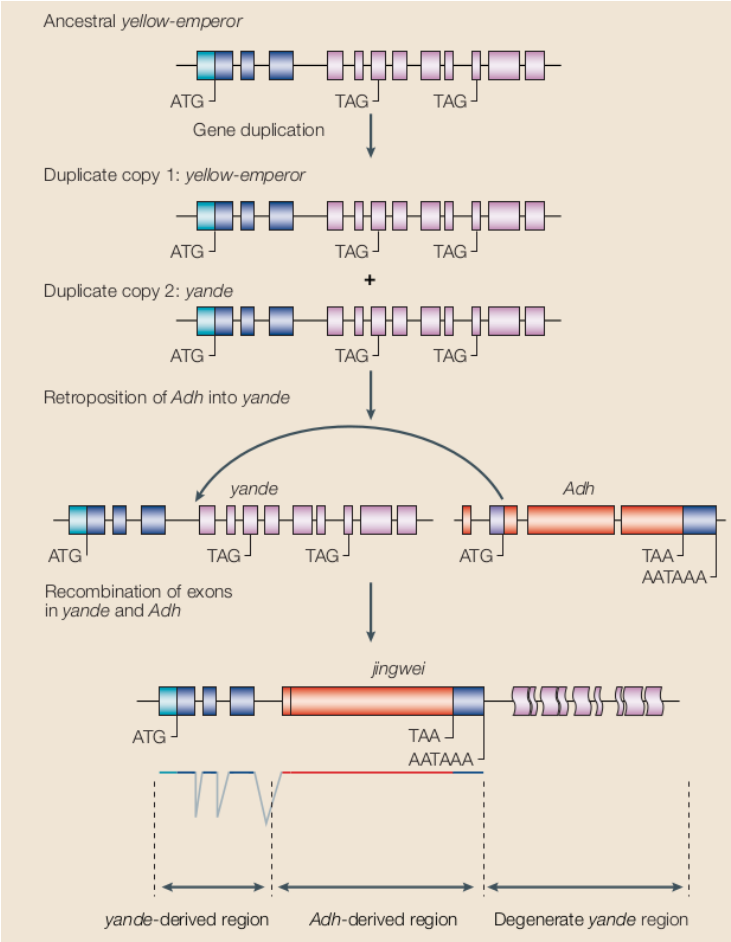


Figure 16. Mécanisme de création du gène *jingwey* (source : [167]**)**

L'exemple du gène *jingwey* montre que l'apparition d'un gène chimérique n'implique pas forcement la perte des gènes qui ont permis sa formation conservant ainsi leurs fonctions. Il montre aussi que l'apparition de gènes chimériques permet de créer de nouvelles fonctions sous sélection positive [169]. Dans le cas du gène *jingwey*, les trois premiers exons provenant du gène *yande* forment une sous-unité essentielle ayant un rôle fonctionnel dans le métabolisme hormonal et la synthèse de phéromones [169].

c) Gènes composites et convergence

De manière extrêmement intéressante, on trouve deux autres gènes chimériques dérivés du gène *Adh* chez d'autres espèces de drosophiles : *Adh-Finnegan* et *Adh-Twain*. Ces gènes sont apparus indépendamment du gène *jingwey* et sont dérivés de la fusion du gène *Adh* et de l'extrémité 5' d'un autre gène [170]. Ces résultats montrent que les mêmes composantes peuvent être recyclées dans de nouveaux gènes et parfois aboutir à des convergences, c'est-à-dire à la création indépendante de gènes identiques [171].

d) Gènes composites, innovations fortes et faibles

Les gènes chimériques sont des innovations en termes de séquences dans le sens où ils ne possèdent pas d'homologues. En revanche, l'apport fonctionnel de ces nouveaux gènes pour la cellule est plus contrasté. Certaines de ces innovations géniques permettent la création de nouvelles fonctions dans la cellule; de nombreux cas ont été rapportés chez la drosophile en plus du gène *jingwey* [172–174] et une grande partie sont sous sélection positive [173] et/ou codent pour de nouvelles fonctions. D'autres cas de création de gènes chimériques au pouvoir adaptatif ont été recensés chez les fougères [175–177], l'homme [154,178], les primates [171,179], les ciliés [180] ou encore chez les champignons par exemple [181]. Ces innovations peuvent être considérées comme fortes car elles apportent des propriétés émergentes à leurs porteurs. D'autres semblent plutôt améliorer des associations existantes [182] et peuvent être considérées comme des innovations faibles. C'est le cas des gènes fusionnant des gènes codant pour des protéines en interaction dans les réseaux PPI [183–186] ou encore des gènes codant pour des protéines impliquées dans les mêmes voies métaboliques [187,188].

IV. Les gènes symbiogénétiques : création de gènes chimériques chez les organismes composites à partir de fragments génétiques des partenaires symbiotiques (Articles VI, VII et VIII)

L'impact des endosymbioses de la mitochondrie et du chloroplaste est souvent résumé à l'acquisition de nouveaux compartiments cellulaires et de nouveaux gènes. Or le niveau d'intrication va au-delà du mosaïcisme génomique. Au niveau structural, le stigma des dinoflagellés du groupe des *Warnowiaceae*, un organite photorécepteur similaire à un œil [189], est un assemblage d'éléments structuraux de la mitochondrie et du chloroplaste [190] (Figure 17). Les voies métaboliques possèdent souvent des enzymes de différentes origines [191–193]. La plupart des machines macromoléculaires eucaryotes héritées des partenaires symbiotiques comme le système ubiquitine/protéasome [194], le système chaperonne [195], le système d'interférence de l'ARN [196], le complexe I de la chaine respiratoire [197] ou encore la machinerie transcriptionnelle [198] se sont complexifiés dans le sens où de nouveaux composants d'origines parfois différentes ont été ajoutés [199].

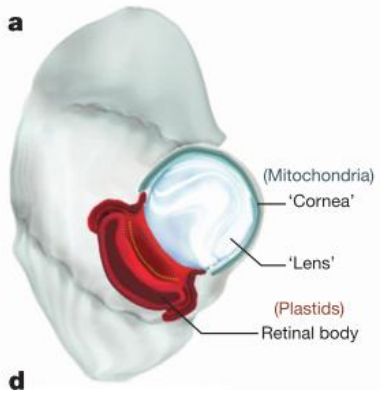


Figure 17. Schéma du stigma avec l'origine des différentes structures chez *Nematodinium*. (source : [190]**)**

Dans le cadre de ma thèse, je me suis particulièrement intéressé au lien entre les transitions égalitaires et l'évolution au niveau subgénique.

La réunion au sein d'un même génome de gènes ayant des origines phylogénétiques distinctes a-t-elle encouragé la création de nouveaux gènes en combinant des fragments génétiques d'origines différentes ?

Ces nouveaux gènes composites dérivant de fragments génétiques de différents partenaires symbiotiques ont été appelés gènes « symbiogénétiques » (S-gènes). Par exemple, dans le cas des eucaryotes photosynthétiques, existe-t-il des gènes composites exclusifs aux

eucaryotes photosynthétiques possédant au moins une composante d'origine cyanobactérienne (et donc provenant sans doute de l'endosymbionte photosynthétique) ? (Figure 18).

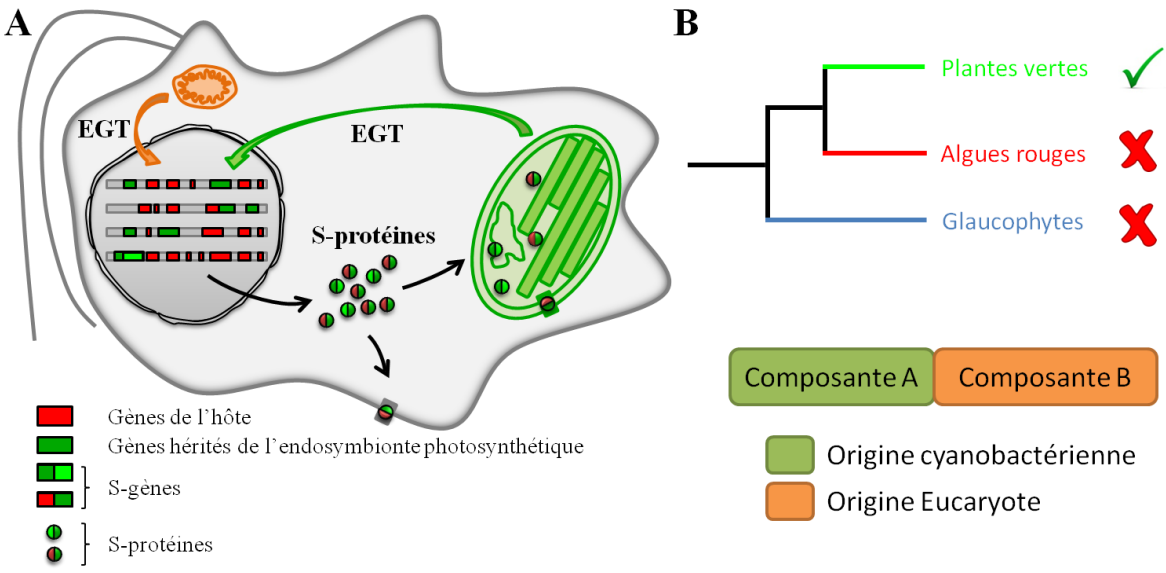


Figure 18. Origine des gènes composites chez les eucaryotes photosynthétiques (figure adaptée de : [200]**)**

L'existence ou l'absence de tels gènes pose des questions intéressantes. Si leur existence est avérée, quels sont leurs rôles dans la cellule photosynthétique ? Ont-ils un rôle dans l'intégration de l'endosymbionte ? Confèrent-ils des propriétés émergentes à l'holobionte au-delà de l'intégration de l'endosymbionte ? Peut-on trouver des règles d'assemblage suivant l'origine des composantes comme par exemple une plus grande occurrence de S-gènes composés de fragments d'origine cyanobactérienne ? *A contrario*, Une absence de S-gènes poserait la question des contraintes empêchant leur évolution malgré le rapprochement physique de ces différents matériaux génétiques depuis au moins plusieurs centaines de millions d'années.

L'ensemble de ces questions a été étudié chez d'autres organismes composites issus de transitions évolutives majeures comme les eucaryotes et les halobactéries, groupe d'Archaea qui possèdent une forte proportion de gènes d'origine bactérienne [98].

Protein networks identify novel symbiogenetic genes resulting from plastid endosymbiosis

Raphaël Méheust[a], Ehud Zelzion[b], Debashish Bhattacharya[b], Philippe Lopez[a], and Eric Bapteste[a,1]

[a]Unité Mixte de Recherche 7138 Evolution Paris Seine, Institut de Biologie Paris Seine, Université Pierre et Marie Curie, Centre National de la Recherche Scientifique, Sorbonne Universités, 75005 Paris, France; and [b]Department of Ecology, Evolution, and Natural Resources, Rutgers University, New Brunswick, NJ 08901

Edited by John M. Archibald, Dalhousie University, Halifax, Canada, and accepted by the Editorial Board February 14, 2016 (received for review September 8, 2015)

The integration of foreign genetic information is central to the evolution of eukaryotes, as has been demonstrated for the origin of the Calvin cycle and of the heme and carotenoid biosynthesis pathways in algae and plants. For photosynthetic lineages, this coordination involved three genomes of divergent phylogenetic origins (the nucleus, plastid, and mitochondrion). Major hurdles overcome by the ancestor of these lineages were harnessing the oxygen-evolving organelle, optimizing the use of light, and stabilizing the partnership between the plastid endosymbiont and host through retargeting of proteins to the nascent organelle. Here we used protein similarity networks that can disentangle reticulate gene histories to explore how these significant challenges were met. We discovered a previously hidden component of algal and plant nuclear genomes that originated from the plastid endosymbiont: symbiogenetic genes (S genes). These composite proteins, exclusive to photosynthetic eukaryotes, encode a cyanobacterium-derived domain fused to one of cyanobacterial or another prokaryotic origin and have emerged multiple, independent times during evolution. Transcriptome data demonstrate the existence and expression of S genes across a wide swath of algae and plants, and functional data indicate their involvement in tolerance to oxidative stress, phototropism, and adaptation to nitrogen limitation. Our research demonstrates the "recycling" of genetic information by photosynthetic eukaryotes to generate novel composite genes, many of which function in plastid maintenance.

gene fusion | endosymbiosis | photosynthesis | eukaryote evolution | novel gene origin

The genomes of the proteobacterium-derived mitochondrion and the cyanobacterium-derived plastid have undergone significant genome reduction due to outright gene loss or transfer to the nuclear genome (1, 2). Organelle gene loss by transfer to the nucleus is known as endosymbiotic gene transfer [EGT (a special form of horizontal gene transfer; HGT)] and has resulted in chimeric host nuclear genomes with, in the case of plastids, from ca. 200 to several thousand intact endosymbiont genes being relocated (3) (Fig. 1*A*). Plastid EGT has a long evolutionary history, extending back over a billion years in the case of primary plastid origin in the Archaeplastida (glaucophytes, red and green algae, and their sister group, plants) and several hundred million years for secondary plastids in groups such as diatoms, haptophytes, and dinoflagellates (4). A common fate for many nuclear-encoded organelle-derived proteins is to be targeted back to the compartment of origin via channels [i.e., translocons at the outer- and inner-envelope membrane of plastids and mitochondria (Toc/Tic and Tom/Tim, respectively)] to carry out organelle functions (5). Identification of EGT candidates generally relies on phylogenetic methods that use simultaneous alignment of colinear proteins sharing significant sequence similarity over all, or most, of their lengths to reconstruct the tree and its constituent branch lengths. An alternative approach is network methods that rely on reconstruction of both full and partial (i.e., protein domain; Fig. 1*B*) gene relationships using pairwise protein similarity values. These methods allow detection of reticulate sequence evolution, such as the fusion of domains derived from heterologous proteins (6–10). Here we used networks to ask the following two questions: (*i*) Did the Archaeplastida plastid endosymbiont contribute gene fragments to symbiogenetic genes (S genes) that are detectable in algal and plant nuclear genomes? (*ii*) If so, are these S genes expressed, and what putative functions did the novel domain combinations confer to the host lineage? These questions are motivated by the knowledge that although fundamental to the origin of complex life forms such as plants and animals, plastid endosymbiosis wrought significant challenges for the first algal lineages. These resulted from light harvesting, which can capture excess energy that must be dissipated, and oxygen evolution, which leads to the formation of reactive oxygen species (ROS) that need to be detoxified (11, 12).

Results and Discussion

We identified 67 families of expressed nuclear-encoded S genes (Fig. 2). These families are distributed in 349 algae and plants. Four S-gene families were likely present in the Archaeplastida ancestor, 11 S-gene families are shared by the red and the green lineages, and 28 S-gene families are found both in primary and secondary photosynthetic lineages, demonstrating their ancient origins and functional relevance (Fig. 3 and Fig. S1). The 55 S-gene candidates we focused on here are predicted to be plastid-targeted (Table S1), and at least 23 of these function in redox regulation and light and stress responses (Fig. 2).

Significance

Endosymbiotic gene transfer from the plastid genome to the nucleus comprises the most significant source of horizontal gene transfer in photosynthetic eukaryotes. We investigated genomic data at the infragenic level to determine whether the cyanobacterial endosymbiont also contributed gene fragments (i.e., domains) to create novel nuclear-encoded proteins. We found 67 such gene families that are expressed as RNA and widely distributed among plants and algae. At least 23 genes are putatively involved in redox regulation and light response, namely the maintenance of a photodynamic organelle. Our results add a new layer of complexity to plastid integration and point to the role of fused proteins as key players in this process.

Author contributions: R.M., P.L., and E.B. designed research; R.M. and E.Z. performed research; R.M. performed detection of S genes; E.Z. analyzed RNA-sequencing data; R.M., D.B., P.L., and E.B. analyzed data; and R.M., D.B., P.L., and E.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. J.M.A. is a guest editor invited by the Editorial Board.

Freely available online through the PNAS open access option.

Data deposition: The FASTA sequences of the S genes reported in this paper are available at www.evol-net.fr/downloads/S-genes.zip.

[1]To whom correspondence should be addressed. Email: epbapteste@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1517551113/-/DCSupplemental.

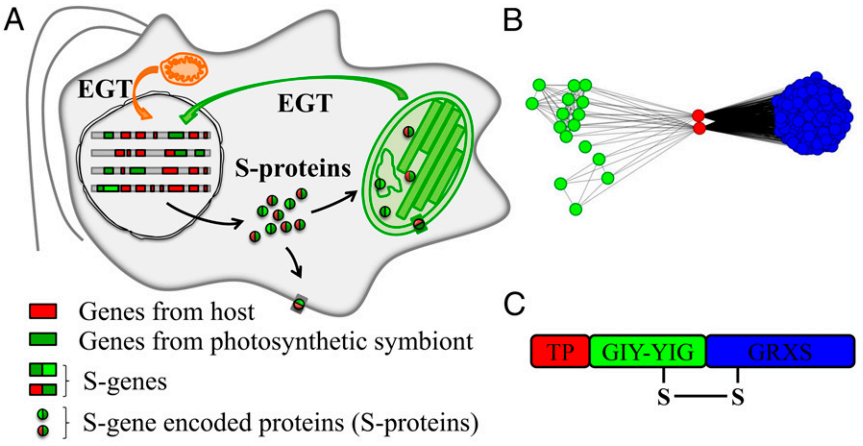


Fig. 1. Origin of composite genes in algae and plants. (*A*) The role of plastid endosymbiosis in providing the genetic toolkit for S-gene origin. (*B*) Network analysis of the *AtGRXS16* (family 14) S gene in *A. thaliana*. The red nodes identify the S genes; green and blue nodes are the components from GIY–YIG and GRXS domains, respectively, that gave rise to S genes through gene fusion. (*C*) Domain structure of AtGRXS16. An intramolecular disulfide bond can be formed between the two domains. TP, transit peptide.

Evidence That S Genes Are Not Assembly Artifacts. It is conceivable that the union of two unrelated protein domains that we report here as S genes could potentially be explained by misassembly of genomic or transcriptomic reads, an expected outcome of the analysis of large datasets. Given this concern, we used several approaches to validate the existence of S genes. The first was to collect RNA-seq (sequencing) data that could be used to map to coding sequences (CDSs) and genomic sequences of S genes. If the RNA reads mapped uniformly across the CDS or genomic DNA with no loss of coverage at the domain junctions, then we had evidence the coding region was authentic. We did this procedure for two taxa, a green alga *Picochlorum* and the model plant *Arabidopsis thaliana*. In the first case, we downloaded transcriptome reads from *Picochlorum* SE3 [National Center for Biotechnology Information (NCBI) BioProject accession no. PRJNA245752] and mapped these to the CDSs of S genes from its closely related sister species *Picochlorum oklahomensis* and *Picochlorum* RCC944 (Table S2). These results showed that for nine shared homologs, transcriptome coverage across the CDSs was nearly 100% and uniform across the domain junctions (Fig. S2). These results strongly support the existence of these S genes. Furthermore, we used PCR with genomic DNA from *Picochlorum* SE3 for five S genes to validate that they were intact fragments. These results are shown in Fig. S3, and sequencing of the nearly complete CDS fragments showed identity to the genomic region encoding the S gene. Mapping of RNA-seq reads to *A. thaliana* S gene-encoding genomic regions (i.e., exons and introns; Table S2) also showed robust and uniform mapping to the exonic regions (Fig. S2), again supporting the existence of intact S genes in this well-annotated genome.

We also checked whether S genes may result from gene misannotation (i.e., the annotation of two separate gene sequences as a single gene, or misincorporation of an exon from two overlapping genes into a gene annotation). We found evidence that 23 S-gene families have at least one gene with all domains being positioned in the same exon, thereby arguing against possible misincorporation of exon information (Table S3). Finally, although we did not validate every S gene cited in this study, we are buoyed by the fact that all families are found in at least one genome and one transcriptome, with many occurring in >10 taxa (Fig. 2 and Fig. S1), making it highly unlikely that these data are explained by artifacts due to misassembly. Although it is difficult to reconstruct robust and resolved domain phylogenies due to their small size, we assessed whether S genes may have been misannotated by reconstructing complete S-gene trees. For example, the phylogeny of an anciently derived S gene (family 31) limited to Viridiplantae is shown in Fig. S4 and supports the existence of this composite sequence in the green lineage ancestor. This tree is in agreement with the accepted relationship of green lineages, thereby showing no evidence of a complex history but rather persistence of the gene family across species. These results are summarized in Fig. 2, which also reports the number of transcriptomes and genomes of distinct organisms in which homologs of S genes were found. Because some transcriptomes are derived from phagotrophic protists (in particular, heterotrophic dinophytes such as *Oxyrrhis*), there is a risk of prey contamination (i.e., the S gene might derive from prey DNA). Therefore, identifying the S gene in multiple transcriptomes from a given taxonomic group provides stronger support for the presence of the S gene in that group.

S Genes Involved in Redox Regulation. Many S-gene families play a role in redox regulation, including family 14, which contains AtGRXS16, a plastid-localized protein in *A. thaliana* (Fig. 2 and Fig. S2). This gene family is widely distributed in Viridiplantae (green algae and plants), and may also be present in a small number of other species (Fig. S1). *AtGRXS16* is composed of two fused domains that do not exist together elsewhere in the tree of life. This S-gene family encodes an N-terminal GIY–YIG (GlyIleTyr–TyrIleGly) endonuclease fold of cyanobacterial origin and a C-terminal CGFS-type monothiol GRXS (glutaredoxin; disulfide oxidoreductase) of bacterial (yet noncyanobacterial) origin that are negatively regulated by the formation of an intramolecular disulfide bond (Fig. 1*C*). This association allows ROS scavenging via the GRXS domain coupled with the ability of the GIY–YIG endonuclease to repair oxidative stress-induced DNA double-strand breaks in plant plastid genomes (13). Consequently, this anciently derived S gene plays an important role in coordinating redox regulation and DNA repair in response to ROS (13). Consistent with these observations are RNA-seq data (14) that show a ca. threefold up-regulation of *AtGRXS16* ($P < 0.01$) in *A. thaliana* seedlings in light versus dark conditions (*S-Gene Expression Analysis*).

Domains in S genes can be reused for redox regulation, as illustrated by family 4. This gene is found in the red, green, and secondary plastid-derived lineages and is composed of two fused domains. The N-terminal region again encodes a GIY–YIG endonuclease fold of cyanobacterial origin, whereas the C terminus encodes a NifU domain of cyanobacterial origin that is involved in iron–sulfur (Fe–S) cluster assembly (15). Bioinformatic evidence was found for plastid targeting of this protein (Fig. 2).

Another S gene involved in redox regulation that is widely distributed in Viridiplantae is family 19 (Fig. 2). This modular gene (*SufE3*) encodes quinolate synthetase and defines a novel combination of two biochemically interacting domains: a SufE domain of cyanobacterial origin and a NadA domain of (noncyanobacterial) prokaryotic origin (16). The quinolate activity of the NadA domain relies on a highly oxygen-sensitive (4Fe–4S) cluster, whose formation depends on a cysteine residue present in its novel genetic partner, the SufE domain, which is involved in the long-term competence of the enzyme (16). Because this nuclear-encoded quinolate synthetase is plastid-localized (17), it is likely to be exposed to high levels of oxidative stress. The SufE domain has been proposed to continuously repair/reconstitute

Family	Domains (gene name)	Assumed function	Taxonomic distribution										Evidence						
			Glaucophyta	*Rhodophyta*	*Chlorophyta*	*Streptophyta*	*Cryptophyta*	*Haptophyta*	*Cercozoa*	*Ochrophyta*	*Bacillariophyta*	*Dinophyta*	*Euglenozoa*	Dataset genomes	Other genomes	Transcriptomes	Published evicence	RNA-seq	RT-PCR	Read mapping	Same exon
1	**hydrolase†*** + LPLAT† (*ELT* or *PES*)	maintenance of membrane integrity	1	6	12	59		7		5	22	12	1	7	63	59					
2	**NTF2-like†*** + SOUL† (*SOUL3*)	eyespot related in *C.reinhardtii*		5	12	54	1	10	2	10	26	2		11	57	55					
3	**PAPS reductase** + thioredoxin	sulfur assimilation		1	5	44	1	10	2	5	32	9		5	46	62					
4	**GIY-YIG†*** + **NifU†**	DNA repair + redox	1	3	8	1		1		5	8	4		6	3	22					
5	**SufE*** + BolA (*SufE1*)	redox		5	6	35		2		8	14	2		5	36	34					
6	**photolyase** + **hydrolase†***	DNA repair + light response		2	5	45		1		2		1		2	45	9					
7	**DUF3593†** + **DUF2499†***	-		3	4			6		5	16	5		2	3	37					
8	**3-dehydroquinate synthase** + O-methyltransferase	shikimate biosynthesis		1					1	4	9	5	1	1	3	17					
9	**serine protease** + LIM zinc finger domain	-	1						1	2	1	13		3	0	15					
10	PPIase† + **rhodanese†** (*PIN3*)	auxin efflux in *A.thaliana*		6	7	21		3		3				4	25	11					
11	**psbD†** + **psbC†**	photosynthesis		1	2	1		2			1			1	1	5					
12	**DUF760†*** + **DUF760†***	-			1	4		4		3		1		1	4	8					
13	**tsf†*** + **tsf†***	translation			3	1			2			10	1	4	3	10					
14	**GIY-YIG†*** + GRXS (*AtGRXS16*)	DNA repair + redox		1	3	45						1		2	45	3					
15	**peroxiredoxine** + thioredoxin†	redox		1	1					5		1		1	1	6					
16	**atpB†** + **atpE†**	ATP synthase		2		4		1			2			1	4	4					
17	PPIase + **tRNA-i(6)A37 methylthiotransferase†**	tRNA transferase		4						4	9	1		1	1	17					
18	**ferredoxin†*** + tetratricopeptide repeat	redox			5	43				1	1			3	43	4					
19	**SufE†*** + NadA (*SufE3*)	quinolate synthetase			14	51				5			1	4	55	13					
20	DUF2256 + **unknown domain 3†**	-			2			3			11	4		3	2	16					
21	glycosyltransferase + **glycosyltransferase***	-						1		1		1	1	1	0	3					
22	**phytochrome†** + PAS domain	photosensory signaling protein	1			127	1							1	130	2					
23	TPR repeat + RING + **ATP-dependant protease**	-		1	3	43								1	43	3					
24	DnaJ + **ferredoxin†***	redox			4	5						1		3	5	2					
25	**RNA methyltransferase†** + 2OG-Fe(II) oxygenase	-			3			2				1		1	1	5					
26	CobU† + **DUF4346†**	cobalamin biosynthesis					1	1			15			2	1	14					
27	**acetyltransferase†** + methyltransferase	-					1	6				1		1	0	7					
28	**tic-20†** + calmodulin	translocation								7	29	14		3	3	47					
29	**allophycocyanin beta subunit†*** + **allophycocyanin beta subunit†***	photosynthesis		1			1							1	0	1					
30	RPS13 + **RPS11** + **RNA polymerase alpha subunit†**	transcription/translation		1							1			1	0	1					
31	**RIBR†** + **DUF1768** (*PyrR*)	riboflavin biosynthesis			11	55								6	57	3					
32	**DUF2499†*** + unknown domain 1	-			1						3			1	0	3					
33	**PDZ** + FKBP	periplasmic protease			7							1		2	3	3					
34	**ferredoxin nitrite reductase†** + oxido-reductase + rubredoxin	nitrogen assimilation			7							1		2	2	4					
35	NTF2-like + **unknown domain 2†**	-						5			1			1	1	4					
36	**DUF2930†*** + **DUF2930†***	-							1			4		1	0	4					
37	**NTF2-like†*** + **lipase†**	-								4		8		1	0	13					
38	**ftsH†*** + **ftsH†***	maintenance of photosynthesis									1	10		1	0	10					
39	**phycobilisome linker polypeptide duplication†***	photosynthesis	2											1	0	1					
40	**SufE†*** + tRNA 5-methylaminomethyl-2-thiouridylate methyltransferase	tRNA transferase		2										1	0	2					
41	glutamylcyclotransferase + **hydrolase†*** + LPLAT† (*ELT* or *PES*)	maintenance of membrane integrity		6										3	1	2					
42	DnaJ† + **phycocyanobilin lyase†**	light response + redox		9										3	1	5					
43	**O-methyltransferase*** + deoxyadenosine/deoxycytidine kinase	-		2										1	0	1					
44	acyltransferase + **glyoxalase†**	Lactoylglutathione lyase		2										1	0	1					
45	**DUF89** + Fructose-1,6-bisphosphatase†	Calvin cycle enzyme			4									1	2	1					
46	ABC transporter + **DUF2246**	ABC transporter			3									1	0	2					
47	S1† + S1† + **tsf*** + **tsf†*** (*PETs*)	translation			3									2	0	1					
48	carbohydrate Binding Module 48 + **sucrose phosphatase**	carbohydrate metabolism			2									1	0	1					
49	**9-cis-epoxycarotenoid dioxygenase** + glutathione S-transferase†	abscisic acid biosynthesis + redox			8									2	3	3					
50	**fasciclin** + chlorophile a-b binding protein	surface protein			7									2	2	3					
51	**peroxiredoxin-like†** + **PAP-fibrillin†**	-			3									1	0	2					
52	**excinuclease B†** + **excinuclease C**	DNA repair + light response			3									1	1	1					
53	ankyrin + **DEAD/DEAH box helicase** + **DSHCT†**	-			2									1	0	1					
54	**transcription activator TenA** + HMP-P kinase† + TMP-Ppase†	thiamin metabolism			2									1	0	1					
55	**glycosyltransferase*** + SNARE	-			4									1	1	2					
56	**DUF393** + polyphosphate glucokinase† + EF-hand + thioredoxin	-			5									1	2	2					
57	**ribokinase** + kinase + **CHAT domain**	pentose phosphate pathway						4						1	0	3					
58	**arsC transcriptional regulator†** + arsM	-						5						1	0	4					
59	**SNARE** + **2-polyprenyl-6-methoxyphenol hydroxylase**	FAD-dependent oxidoreductase						2						1	0	1					
60	**phosphatase** + EF-hand + **CBS**	-									9			1	0	8					
61	G6PD† + **6PGD**	pentose phosphate pathway									7			1	0	6					
62	**SufE†*** + cysteine--tRNA synthetase†	tRNA synthetase										10		1	0	9					
63	methyltransferase + **cytochrome b6/f complex, subunit V†**	-										2		1	0	1					
64	bacteriorhodopsin-like + PAS + **transduction signal**	photosensory signaling protein										16		1	0	15					
65	**CobW** + TNF receptor superfamily + EPS_sugar_tfrase	-										11		1	0	10					
66	**ferredoxin†*** + ferritin†	iron storage										5		1	0	4					
67	DUF2358 + **NTF2-like†*** + **hydrolase***	-										1		1	0	1					

Fig. 2. Sixty-seven S-gene families identified in our study. Domains in bold originated from Cyanobacteria. Plastid-localized protein families (i.e., families with at least one protein predicted to be plastid-targeted according to ChloroP and ASAFind) are shaded in gray. *, domain of cyanobacterial origin occurring more than once per S gene. †, highly confident domain of cyanobacterial or prokaryotic (noncyanobacterial) origin (Table S4).

the Fe–S cluster in the NadA domain of the quinolate synthetase to maintain a functional protein (18). *SufE3* is ubiquitously expressed in all major plant organs and is embryo-lethal when knocked out in *A. thaliana* (18). In this model plant, there is ca. twofold gene (*At5g50210*) up-regulation ($P < 0.01$) in light versus dark conditions (14) (*S-Gene Expression Analysis*).

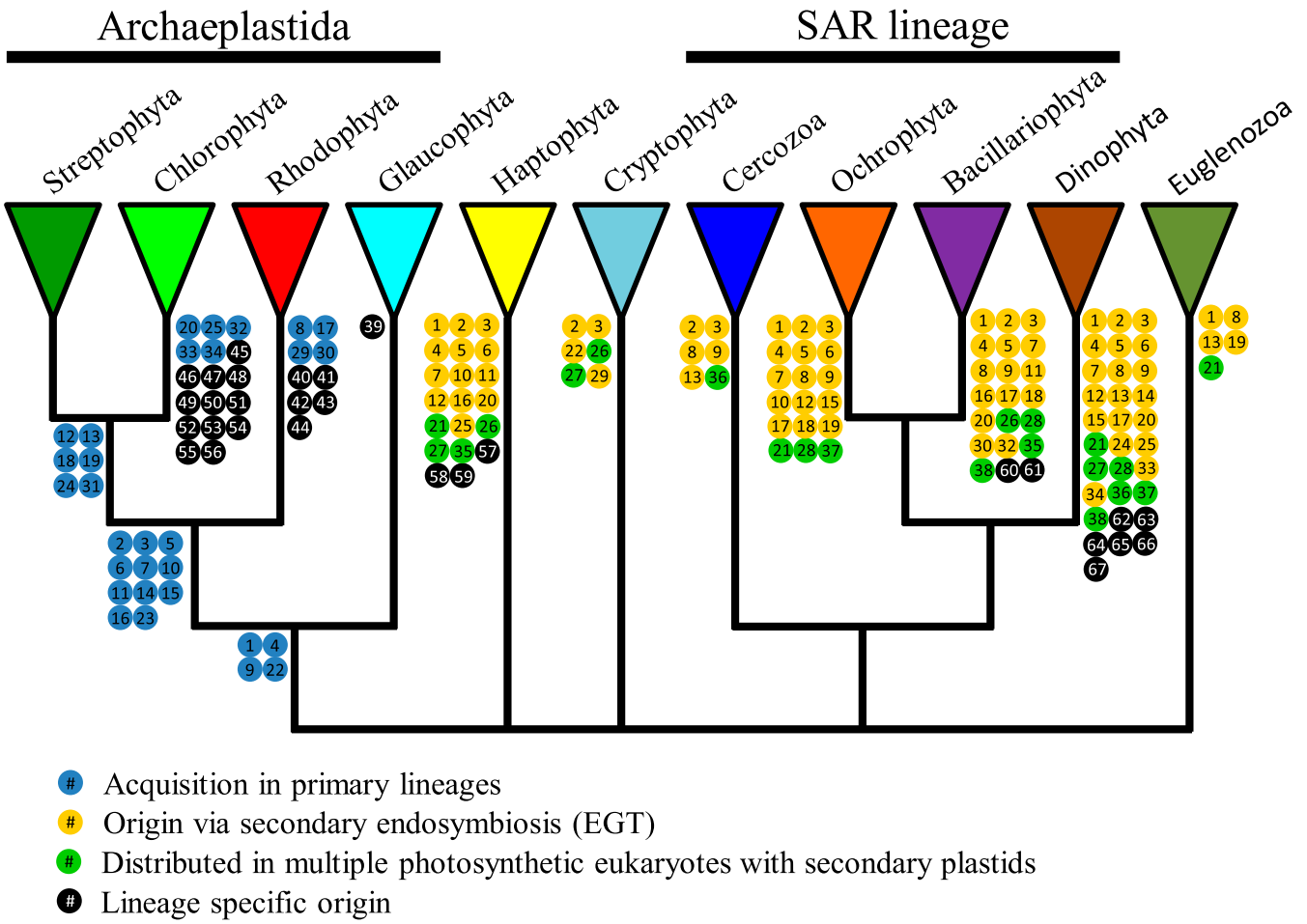


Fig. 3. Putative nuclear gene-based phylogeny of photosynthetic eukaryotes, showing the distribution of the 67 S-gene families we report. SAR, Stramenopiles-Alveolates-Rhizaria.

Another fascinating example is family 49 (Fig. 2), which is restricted to prasinophyte green algae and encodes two cyanobacterium-derived domains. The N-terminal region is a 9-*cis*-epoxycarotenoid dioxygenase (RPE65) domain involved in the production of abscisic acid from xanthophyll precursors (19), whereas the C terminus contains a glutathione S-transferase (GST) domain, which in plants plays a major role in reducing oxidative stress damage. Whereas responses to oxidative stress appear to be central to S-gene evolution, we also find examples of their roles in coordinating algal responses to light direction to optimize photosynthesis and growth.

S Genes Involved in Light Responses. S-gene family 2 (Fig. 2) defines the well-studied *AtHBP5* gene in *A. thaliana* and *SOUL3* in *Chlamydomonas reinhardtii*. This gene fusion is composed of an N-terminal region of cyanobacterial origin and a C-terminal region of prokaryotic derivation, and is present in the red and green lineages within Archaeplastida as well as in secondary plastid-containing algae. The heme-binding protein in *A. thaliana* (AtHBP5) is localized in plastoglobules, where it is likely involved in chlorophyll degradation (20). SOUL3 is localized to the plastid eyespot of *C. reinhardtii* (21) and, when knocked-out, the eyespot is reduced in size and its location is altered, negatively impacting phototaxis (21). AtHBP5 and SOUL3, which facilitate a coordinated response to light of the photosynthetic cell, produce an analogous phenotype to the communal phototropism of the well-known prokaryotic consortium *Chlorochromatium aggregatum* (22). In the latter case, cross-talk between photosynthetic epibiontic bacteria is transferred to a central motile, brown bacterium, thereby moving the collective to a location where epibionts can most efficiently perform photosynthesis (22).

Another family of S genes, family 10, is involved in phototropism and gravitropism (Fig. 2). This gene is composed of two domains, a peptidyl prolyl isomerase (PPIase) and a rhodanese superfamily domain, with the former of (noncyanobacterial) prokaryotic origin and the latter of cyanobacterial provenance. This S gene encodes a widely distributed PPIase in plants, red algae, haptophytes, and stramenopiles that is likely to be plastid-targeted (Fig. 2). In *A. thaliana*, this developmental protein (known as PIN3) is localized to the plasma membrane and reallocates auxin, affecting phototropism and gravitotropism of young sprouts (23).

S Genes Involved in Endosymbiont Stabilization. Achieving genetic integration also required innovations to stabilize the endosymbiont in the host cell. S genes were involved in this function as well, with some playing a role in scavenging organelle degradation products during abiotic stress. Family 1 (Fig. 2) encodes a plastid-localized composite protein in *A. thaliana* that contains two domains [e.g., an esterases/lipases/thioesterases (ELT) or phytyl ester synthase (PES) domain, and a hydrolase domain of cyanobacterial origin]. This protein is widely distributed in photosynthetic eukaryotes (Fig. S1), and in *A. thaliana* forms a gene family involved in fatty acid phytyl ester synthesis that is highly expressed during senescence and nitrogen deprivation (24); that is, these proteins scavenge toxic free phytol and fatty acids after thylakoid degradation. Family 41 (Fig. 2) is similar to family 1, albeit with an additional bacterium-derived gamma-glutamylcyclotransferase N-terminal domain involved in glutathione

metabolism. The taxonomic distribution of family 41 is restricted to red algae, suggesting that lineage-specific fusion events may have given rise to convergent functions to protect plastid membranes from abiotic stress.

S Genes with Potential Novel Functions in Photosynthetic Eukaryotes. Another important aspect of our network analysis was to provide the foundation for experimental analysis of novel genes, because S genes could also have introduced novel biochemical functions that are exclusive to photosynthetic eukaryotes. An example of this is family 42, which is restricted to red algae (Fig. S1). This S gene is composed of an N-terminal, bacterium-derived chaperone DnaJ domain fused to a phycocyanobilin (PCB) lyase domain of cyanobacterial origin. PCB lyases attach bilin chromophores to light-harvesting phycobiliproteins through thioether bonds to cysteine residues. This modular protein appears to be plastid-targeted in rhodophytes. Absent functional data, the biological relevance of family 42 remains unknown but suggests the possibility of stress-dependent regulation of PCB maturation via lyase-dependent chromophore attachment.

Similarly, a central innovation in plastid evolution was the evolution of the plastid translocons (Toc/Tic) to allow the controlled entry of proteins translated in the cytosol into the organelle. We find here that domains present in translocon proteins can be recruited into S genes. This appears to be the case for family 28 (Fig. 2), which is absent from Archaeplastida but present in the red alga-derived plastid-containing stramenopiles and dinoflagellates. This modular protein is composed of an N-terminal calmodulin domain of prokaryotic (noncyanobacterial) origin fused with a cyanobacterium-derived Tic20-like domain. The Tic20 domain is widely distributed among photosynthetic eukaryotes (25, 26), where it plays an essential role in the creation of a preprotein-sensitive channel or contributes to retargeting proteins to the apicoplast in secondary plastid-containing organisms such as *Toxoplasma gondii* (27). The function of this novel S gene defies easy explanation; nonetheless, the combination of a calcium-sensing EF hand (two canonical domains exist in diatoms) with a plastid membrane channel protein suggests a role in calcium-dependent protein translocation in secondary photosynthetic eukaryotes. In pea, association between a calmodulin domain and the inner-envelope translocon component Tic32 protein has been reported, because a calmodulin binds to the C-terminal region of Tic32 in the inner chloroplast membrane, affecting channel activity (28). Interestingly, analysis of the N terminus of the S gene, uniting a calmodulin with Tic20, from the diatom *Phaeodactylum tricornutum* 219117465, provides evidence for a signal sequence cleavage site between residues 21 and 22 (SignalP 4.1) and a conserved ASAFAP motif typical for plastid-destined proteins in this species (29). RNA-seq analysis of *P. tricornutum* cultures under replete and nitrogen (N)-depleted conditions shows that the expression of this S gene is significantly down-regulated (ca. fivefold; $P = 2.57e\text{-}23$) under N stress (30) (*S-Gene Expression Analysis*).

Finally, gene family 64 (Fig. 2) might correspond to a new putative symbiogenetic bacteriorhodopsin (31–34). This protein unites a bacteriorhodopsin domain with a seven-transmembrane helical region in the N terminus, a PAS domain, and a transduction signal region of cyanobacterial origin in the C terminus. Interestingly, the transduction signal region is composed of two domains that are similar to the transduction signal region of ETR1 in *A. thaliana*: a signal transduction histidine kinase domain and a signal receiver domain (35, 36). Moreover, the N-terminal bacteriorhodopsin domain is preceded by 100 amino acids that may be involved in targeting. This S-gene family is present only in dinoflagellates.

Conclusions

In this study, we analyzed protein domain origins and identified at least 67 S genes (encompassing 2,153 coding regions) that had previously escaped detection using phylogenetic methods. S-gene functions include redox regulation, response to light, Fe–S cluster assembly, and, putatively, formation of protein channels. A total of 42% are present both in a primary photosynthetic lineage and in secondary plastid-bearing algae, suggesting their ancient emergence and their potential importance in the process of plastid establishment (Fig. 3). In contrast to these ancient S genes, 29 are lineage-specific families (43%) and were likely more recently formed, showing that cyanobacterial domain recycling is an ongoing process with a potential role in niche adaptation (Fig. 3). In addition, 55 of the S-gene products are demonstrated or predicted to be plastid-targeted (Fig. 2 and Table S1), suggesting their evolution offered an effective way to address the protein colocalization challenge in photosynthetic eukaryotes; that is, when fused with an N-terminal cyanobacterial domain that was already plastid-targeted, the novel protein did not need to "reinvent" or recruit the organelle-targeting sequence. Our results further underline the extent to which algae reuse genetic information to create not only complex structures such as the dinoflagellate "eye" (37) and metabolic pathways with chimeric gene origins (38–40) but now endosymbiont-derived composite genes with important roles in plastid maintenance. We suspect that because the number of proposed phylogenetically composite lineages continues to increase with the availability of novel genome data (41) [e.g., the photosynthetic sisters to Apicomplexa, *Chromera velia* and *Vitrella brassicaformis* (42)], our analysis provides a lower bound on S-gene numbers. Moreover, because our protocol excluded S-gene candidates present in nonphotosynthetic eukaryotes, composite genes retained in formerly photosynthetic lineages (e.g., relatives of apicomplexans) were not considered in our analysis. It is also likely that modular proteins with components derived from the mitochondrial endosymbiont will soon be discovered.

Materials and Methods

Dataset Construction. We assembled a protein sequence database by downloading every archaeal, viral, and plasmid genome that was annotated as "complete" according to the NCBI Genome database in November 2013 (152, 3,769, and 4,294 genomes, respectively). We also retrieved 230 eubacterial genomes, with 1 representative randomly chosen per eubacterial family, with the exception of cyanobacterial genomes, from which we selected 16 genomes. Finally, we sampled 38 unicellular eukaryotic genomes across the eukaryotic tree of life: 19 for photosynthetic organisms and 19 that are nonphotosynthetic, with a comparable total gene number and phylogenetic diversity in their ribosomal proteins. The resulting 2,192,940 protein sequences were compared pairwise using BLASTP (43) (version 2.2.26) (*E*-value cutoff 1e-5) (see Dataset S1 for the list of genomes used).

Detection of S-Gene Families. Composite genes and their associated component genes were detected with FusedTriplets (8) (*E* value <1e-5) by scanning the BLASTP output. Composite genes that were present in photosynthetic eukaryotes were compared with the entire nonredundant NCBI database (BLASTP; *E* value <10e-5 and ≥80% mutual sequence overlap) to confirm that these sequences had no full-length homologs outside photosynthetic eukaryotes. These composite genes were identified as candidate S genes. All sequences were also clustered into gene families according to a previous method (44, 45). Briefly, an undirected graph was constructed in which each node corresponds to a sequence and two nodes are linked if the corresponding sequences show a BLAST hit with an *E* value <1e-5, ≥30% sequence identity, and a mutual sequence overlap ≥80%. Connected components in this graph were considered to be gene families. For each candidate S gene, we retrieved the corresponding component sequences, as identified by FusedTriplets. Component sequences were clustered into component families according to the following rule: If two component sequences overlapped by more than 80% of their lengths on the protein composite, they belonged to the same component family. Component families were assigned a phylogenetic origin corresponding to their taxonomic composition. Component families were considered to be of eukaryotic origin if all their sequences belonged to eukaryotes. When one or more sequences from a component family contained prokaryotic sequences, we considered the component family to be of prokaryotic origin. If the three best prokaryotic component genes, according to their BLASTP bitscore against the composite gene, matched with the same prokaryotic phylum (e.g., Cyanobacteria), we considered the component to have more specifically originated from that prokaryotic phylum. All S-gene component origins were confirmed by BLAST analysis against an extensive prokaryotic dataset (2,982 prokaryotic genomes, 8,422,211 sequences). Only candidate S-gene families with at least one of their associated components assigned to a cyanobacterial origin (i.e., putative endosymbiotic origin) were retained.

Gene Expression and Gene Distribution Investigation. To gain insights into gene expression and distribution of S genes, composite sequences were compared with the predicted proteins of the combined assemblies of the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) (46) and additional rhodophyte samples from the MMETSP (data.imicrobe.us/project/view/104) (BLASTP, *E* value <1e-5, ≥80% mutual sequence overlap) (see Dataset S1 for the list of combined assemblies used).

Prediction of Plastid Localization. ChloroP (47) (version 1.1) and ASAFind (29) (version 1.1.7) were used to predict the putative cellular localization of the 67 S proteins listed in Fig. 2. Proteomic data were also used for four species: *A. thaliana* (48), *C. reinhardtii* (49), *Cyanophora paradoxa* (50), and *Ostreococcus tauri* (51).

Exon Analysis. A total of 13 genomes had GenBank files available. For these taxa, we retrieved each exon sequence for each S gene. Exon sequences were blasted against S genes; if one exon contained all domains from the S gene according to the Conserved Domain Database (52), the corresponding S-gene family was considered as not to be subject to exon misincorporation.

ACKNOWLEDGMENTS. We thank Nicole Wagner (Rutgers) for doing the PCR and sequence analysis of the *Picochlorum* species. E.Z. and D.B. are grateful to the Rutgers University School of Environmental and Biological Sciences and members of the Genome Cooperative at School of Environmental and Biological Sciences for supporting this research. E.B. is funded by the European Research Council (FP7/2007-2013 Grant Agreement 615274).

1. Martin W, et al. (1998) Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* 393(6681):162–165.
2. Reyes-Prieto A, Hackett JD, Soares MB, Bonaldo MF, Bhattacharya D (2006) Cyanobacterial contribution to algal nuclear genomes is primarily limited to plastid functions. *Curr Biol* 16(23):2320–2325.
3. Martin W, et al. (2002) Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc Natl Acad Sci USA* 99(19):12246–12251.
4. Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattacharya D (2004) A molecular timeline for the origin of photosynthetic eukaryotes. *Mol Biol Evol* 21(5):809–818.
5. McFadden GI (2014) Origin and evolution of plastids and photosynthesis in eukaryotes. *Cold Spring Harb Perspect Biol* 6(4):a016105.
6. Bapteste E, et al. (2012) Evolutionary analyses of non-genealogical bonds produced by introgressive descent. *Proc Natl Acad Sci USA* 109(45):18266–18272.
7. Haggerty LS, et al. (2014) A pluralistic account of homology: Adapting the models to the data. *Mol Biol Evol* 31(3):501–516.
8. Jachiet P-A, Pogorelcnik R, Berry A, Lopez P, Bapteste E (2013) MosaicFinder: Identification of fused gene families in sequence similarity networks. *Bioinformatics* 29(7):837–844.
9. Jachiet P-A, Colson P, Lopez P, Bapteste E (2014) Extensive gene remodeling in the viral world: New evidence for nongradual evolution in the mobilome network. *Genome Biol Evol* 6(9):2195–2205.
10. Leonard G, Richards TA (2012) Genome-scale comparative analysis of gene fusions, gene fissions, and the fungal tree of life. *Proc Natl Acad Sci USA* 109(52):21402–21407.
11. Rockwell NC, Lagarias JC, Bhattacharya D (2014) Primary endosymbiosis and the evolution of light and oxygen sensing in photosynthetic eukaryotes. *Front Ecol Evol* 2(66).
12. Halliwell B (2006) Reactive species and antioxidants. Redox biology is a fundamental theme of aerobic life. *Plant Physiol* 141(2):312–322.
13. Liu X, et al. (2013) Structural insights into the N-terminal GIY–YIG endonuclease activity of *Arabidopsis* glutaredoxin AtGRXS16 in chloroplasts. *Proc Natl Acad Sci USA* 110(23):9565–9570.
14. Jiao Y, Ma L, Strickland E, Deng XW (2005) Conservation and divergence of light-regulated genome expression patterns during seedling development in rice and *Arabidopsis*. *Plant Cell* 17(12):3239–3256.
15. Gao H, et al. (2013) *Arabidopsis thaliana* Nfu2 accommodates [2Fe-2S] or [4Fe-4S] clusters and is competent for in vitro maturation of chloroplast [2Fe-2S] and [4Fe-4S] cluster-containing proteins. *Biochemistry* 52(38):6633–6645.
16. Schippers JHM, et al. (2008) The *Arabidopsis* onset of leaf death5 mutation of quinolinate synthase affects nicotinamide adenine dinucleotide biosynthesis and causes early ageing. *Plant Cell* 20(10):2909–2925.
17. Katoh A, Uenohara K, Akita M, Hashimoto T (2006) Early steps in the biosynthesis of NAD in *Arabidopsis* start with aspartate and occur in the plastid. *Plant Physiol* 141(3):851–857.
18. Narayana Murthy UM, et al. (2007) Characterization of *Arabidopsis thaliana* SufE2 and SufE3: Functions in chloroplast iron-sulfur cluster assembly and NAD synthesis. *J Biol Chem* 282(25):18254–18264.
19. Tan B-C, et al. (2003) Molecular characterization of the *Arabidopsis* 9-*cis* epoxycarotenoid dioxygenase gene family. *Plant J* 35(1):44–56.
20. Lundquist PK, et al. (2012) The functional network of the *Arabidopsis* plastoglobule proteome based on quantitative proteomics and genome-wide coexpression analysis. *Plant Physiol* 158(3):1172–1192.
21. Schulze T, et al. (2013) The heme-binding protein SOUL3 of *Chlamydomonas reinhardtii* influences size and position of the eyespot. *Mol Plant* 6(3):931–944.
22. Overmann J (2010) The phototrophic consortium "*Chlorochromatium aggregatum*"—A model for bacterial heterologous multicellularity. *Adv Exp Med Biol* 675:15–29.
23. Friml J, Wiśniewska J, Benková E, Mendgen K, Palme K (2002) Lateral relocation of auxin efflux regulator PIN3 mediates tropism in *Arabidopsis*. *Nature* 415(6873):806–809.
24. Lippold F, et al. (2012) Fatty acid phytyl ester synthesis in chloroplasts of *Arabidopsis*. *Plant Cell* 24(5):2001–2014.
25. Töpel M, Jarvis P (2011) The Tic20 gene family: Phylogenetic analysis and evolutionary considerations. *Plant Signal Behav* 6(7):1046–1048.
26. Kasmati AR, Töpel M, Patel R, Murtaza G, Jarvis P (2011) Molecular and genetic analyses of Tic20 homologues in *Arabidopsis thaliana* chloroplasts. *Plant J* 66(5):877–889.
27. van Dooren GG, Tomova C, Agrawal S, Humbel BM, Striepen B (2008) *Toxoplasma gondii* Tic20 is essential for apicoplast protein import. *Proc Natl Acad Sci USA* 105(36):13574–13579.
28. Chigri F, et al. (2006) Calcium regulation of chloroplast protein translocation is mediated by calmodulin binding to Tic32. *Proc Natl Acad Sci USA* 103(43):16051–16056.
29. Gruber A, Rocap G, Kroth PG, Armbrust EV, Mock T (2015) Plastid proteome prediction for diatoms and other algae with secondary plastids of the red lineage. *Plant J* 81(3):519–528.
30. Levitan O, et al. (2015) Remodeling of intermediate metabolism in the diatom *Phaeodactylum tricornutum* under nitrogen stress. *Proc Natl Acad Sci USA* 112(2):412–417.
31. Béjà O, et al. (2000) Bacterial rhodopsin: Evidence for a new type of phototrophy in the sea. *Science* 289(5486):1902–1906.
32. Slamovits CH, Okamoto N, Burri L, James ER, Keeling PJ (2011) A bacterial proteorhodopsin proton pump in marine eukaryotes. *Nat Commun* 2:183.
33. Avelar GM, et al. (2014) A rhodopsin-guanylyl cyclase gene fusion functions in visual perception in a fungus. *Curr Biol* 24(11):1234–1240.
34. Scheib U, et al. (2015) The rhodopsin-guanylyl cyclase of the aquatic fungus *Blastocladiella emersonii* enables fast optical control of cGMP signaling. *Sci Signal* 8(389):rs8.
35. Müller-Dieckmann HJ, Grantz AA, Kim SH (1999) The structure of the signal receiver domain of the *Arabidopsis thaliana* ethylene receptor ETR1. *Structure* 7(12):1547–1556.
36. Chang C, Kwok SF, Bleecker AB, Meyerowitz EM (1993) *Arabidopsis* ethylene-response gene ETR1: Similarity of product to two-component regulators. *Science* 262(5133):539–544.
37. Gavelis GS, et al. (2015) Eye-like ocelloids are built from different endosymbiotically acquired components. *Nature* 523(7559):204–207.
38. Oborník M, Green BR (2005) Mosaic origin of the heme biosynthesis pathway in photosynthetic eukaryotes. *Mol Biol Evol* 22(12):2343–2353.
39. Frommolt R, et al. (2008) Ancient recruitment by chromists of green algal genes encoding enzymes for carotenoid biosynthesis. *Mol Biol Evol* 25(12):2653–2667.
40. Reyes-Prieto A, Bhattacharya D (2007) Phylogeny of Calvin cycle enzymes supports Plantae monophyly. *Mol Phylogenet Evol* 45(1):384–391.
41. Nelson-Sathi S, et al. (2015) Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* 517(7532):77–80.
42. Woo YH, et al. (2015) Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. *eLife* 4:e06974.
43. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402.
44. Alvarez-Ponce D, Lopez P, Bapteste E, McInerney JO (2013) Gene similarity networks provide tools for understanding eukaryote origins and evolution. *Proc Natl Acad Sci USA* 110(17):E1594–E1603.
45. Harel A, Karkar S, Cheng S, Falkowski PG, Bhattacharya D (2015) Deciphering primordial cyanobacterial genome functions from protein network analysis. *Curr Biol* 25(5):628–634.
46. Keeling PJ, 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(6):e1001889.
47. Emanuelsson O, Nielsen H, von Heijne G (1999) ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci* 8(5):978–984.
48. Sun Q, et al. (2009) PPDB, the Plant Proteomics Database at Cornell. *Nucleic Acids Res* 37(Database issue):D969–D974.
49. Terashima M, Specht M, Naumann B, Hippler M (2010) Characterizing the anaerobic response of *Chlamydomonas reinhardtii* by quantitative proteomics. *Mol Cell Proteomics* 9(7):1514–1532.
50. Facchinelli F, et al. (2013) Proteomic analysis of the *Cyanophora paradoxa* muroplast provides clues on early events in plastid endosymbiosis. *Planta* 237(2):637–651.
51. Le Bihan T, et al. (2011) Shotgun proteomic analysis of the unicellular alga *Ostreococcus tauri*. *J Proteomics* 74(10):2060–2070.
52. Marchler-Bauer A, et al. (2015) CDD: NCBI's conserved domain database. *Nucleic Acids Res* 43(Database issue):D222–D226.
53. Perrineau M-M, et al. (2014) Evolution of salt tolerance in a laboratory reared population of *Chlamydomonas reinhardtii*. *Environ Microbiol* 16(6):1755–1766.
54. Gorman DS, Levine RP (1965) Cytochrome *f* and plastocyanin: Their sequence in the photosynthetic electron transport chain of *Chlamydomonas reinhardti*. *Proc Natl Acad Sci USA* 54(6):1665–1669.
55. Foflonker F, et al. (2015) Genome of the halotolerant green alga *Picochlorum* sp. reveals strategies for thriving under fluctuating environmental conditions. *Environ Microbiol* 17(2):412–226.
56. Leliaert F, et al. (2012) Phylogeny and molecular evolution of the green algae. *CRC Crit Rev Plant Sci* 31(1):1–46.
57. Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.

Article

Formation of essential chimerical genes at the origin of eukaryotes

Submitted

Raphaël Méheust[1], Debashish Bhattacharya[2], James O McInerney[3], Philippe Lopez[1] and Eric Bapteste[1,*]

[1] Sorbonne Universités, UPMC Univ Paris 06, CNRS, Evolution Paris Seine - Institut de Biologie Paris Seine (EPS - IBPS), 75005 Paris, France

[2]Department of Ecology, Evolution and Natural Resources, Rutgers University, New Brunswick, NJ 08901, USA

[3]Michael Smith Building, The University of Manchester. Oxford Rd, Manchester M13 9PL, United Kingdom.

[*]Author to contact for correspondence.

Summary

Understanding how eukaryotes evolved from the symbiotic association of two prokaryotic partners is a major focus of microbiology, molecular evolution, population genetics, cell biology and genomics. Here we show that chimerical nuclear genes (S-genes) built upon prokaryotic domains, are critical for explaining the leap forward in cellular complexity achieved during eukaryogenesis. Over 300 S-gene families contributed solutions to many of the challenges faced by early eukaryotes: enhancing the informational machinery, handling introns, fighting genotoxicity within the cell, and ensuring functional protein interactions in a larger, more compartmentalized cell. Remarkably, Bacteria contributed 10-fold more S-genes than Archaea, including a twofold greater contribution to informational functions than Archaea. Therefore, there is a large hidden bacterial contribution to the evolution of eukaryotes, implying that fundamental eukaryotic properties do not strictly follow the traditional informational/operational divide for archaeal/bacterial contributions to eukaryogenesis.

Introduction

It has recently been demonstrated that endosymbiosis and the subsequent evolution of eukaryotic photosynthetic lineages was concomitant with the formation of a novel class of nuclear genes, referred to as symbiogenetic genes (S-genes)[1]. New genes can evolve in many ways[2]: by duplication[3], *de novo* formation[4], or by the fusion of gene fragments (i.e., domains) that give rise to novel chimeric proteins[5]. S-genes are in the latter class and emerged in photosynthetic eukaryotes from the union of domains acquired by endosymbiotic gene transfer (EGT) from the plastid to the host nucleus, with domains of other origins. S-genes identified in algae and plants are primarily involved in the integration of an oxygen-evolving, toxic endosymbiont in the eukaryotic cell. Specifically, recycled genetic domains from plastid DNA contributed to the enhancement of metabolic integration and reactive oxygen species (ROS) detoxification within the host eukaryote[1].

Plastids are not the first nor the only organelles potentially present in eukaryotes[6]. Mitochondrial acquisition occurred earlier, likely driving eukaryogenesis. This major evolutionary transition[7,8] took place about two billion years ago and involved two prokaryotic partners, one ancestral archaeon[9] and one ancestral alpha-proteobacterium[10,11]. Even though the details and the genetic, physiological, and structural extent of their merger remain to be established[12,13], there is a consensus that eukaryotes are a genetic chimera because they are comprised of at least two genomes: a nuclear genome and one or more endosymbiotic genomes (mitochondrion, mitosome, hydrogenosome, or plastid)[10].

During the evolution of eukaryotes, the mitochondrial genome has been significantly reduced in size, with many genes being lost, and others being transferred, either intact or in pieces, to the host eukaryotic nucleus through EGT[14–16]. These EGT-derived genes mainly encode operational functions, whereas genes of archaeal origin are usually involved in informational functions[17,18]. In addition to these genes ancestrally inherited from the symbiotic partners, eukaryotes also contain lineage-specific genes[18,19] created during and after eukaryogenesis. As a result, numerous eukaryotic features and processes (e.g., the nucleolus, the cytoskeleton, the DNA replication and transcription systems), while inherited from prokaryotes[20,21], were tinkered with and made more complex[22,23], *via* the addition of essential components that lack prokaryotic homologs[23,24]. Furthermore, eukaryotes have also evolved novel features (e.g., endoplasmic reticulum, golgi, peroxisomes, spliceosome) without prokaryotic counterparts[25]. These innovations occurred early during eukaryogenesis, because

the Last Eukaryotic Common Ancestor (LECA) was endowed with most of the structural traits present in extant lineages[23,26].

We predicted that aspects of this leap forward in organizational and compositional complexity from a consortium of prokaryotes may have been prompted by the evolution of S-genes during the first steps of eukaryogenesis. Phylogenetic methods that use simultaneous alignment of colinear proteins sharing significant sequence similarity over all, or most, of their lengths are useful to analyze the contribution of transferred intact genes to eukaryote evolution. However, the detection of reticulate sequence evolution, such as the fusion and recycling of domains derived from heterologous proteins, benefits from alternative network approaches. Here we used sequence similarity networks[27] that rely on reconstruction of both full and partial (i.e., protein domain) sequence relationships using pairwise protein similarity values to determine whether S-genes played a critical role in eukaryogenesis.

We report a massive burst of S-genes (303 gene families) early in eukaryotic evolution. These chimeric proteins contributed essential components to macromolecular eukaryotic complexes such as the ubiquitin system, the spliceosome, the SSU-processome, the transcription and translation systems, and were involved in membrane trafficking and lipid metabolism. Remarkably, twice as many S-genes of bacterial than of archaeal origin were detected in eukaryotic informational genes. Fundamental eukaryotic properties are thus derived from pieces of prokaryotic genes that have recombined with other domains. Early in their history, and thereafter, eukaryotes exploited domains from multiple co-interacting genomes to retool their own functional repertoire. This observation lies outside of the traditional informational versus operational divide of genetic contributions of archaeal and bacterial lineages, respectively, to the origin of eukaryote gene inventories.

Results and discussion

Massive early creation of S-genes

We searched for homology relationships between 614,589 nuclear-encoded proteins from 38 protists sampled across the tree of life and 1,151,256 proteins from 382 prokaryotes. Briefly, we compared all sequences by BLAST[28], using their sequence similarity to generate clusters (i.e., homologs that can be aligned over 80% of their length, see Methods) that were considered as gene families. This protocol led to 6,733 clusters containing sequences from at least 3 eukaryotic taxa. We considered that a family was multidomain and composite

(Extended Data Fig. 1) when >50% of sequences from the family encoded at least two domains (using CDD[29] or Pfam[30]) and fusedTriplets[27] that indicated chimerism (Extended Data Fig. 2). This conservative protocol returned 1,621 composite multidomain gene families. We classified these families into 3 groups, based on the homology (or lack of homology) of composite eukaryotic sequences with prokaryotic sequences from a reference dataset of 2,982 complete prokaryotic genomes (8,422,211 proteins) (Extended Data Fig. 1). Initially, we found that 633 gene families comprised composite eukaryotic genes with a prokaryotic origin; i.e. both the composite eukaryotic genes and at least one prokaryotic gene could be aligned over their full lengths. The origin of these composite genes likely predated LECA. Thereafter, composite eukaryotic genes in 383 gene families did not share detectable local similarity with prokaryotic sequences, and were thus likely to be eukaryotic innovations. Finally, 605 gene families corresponded to S-genes, because only partial homology was detected between composite eukaryotic and prokaryotic sequences. These 605 genes are noteworthy because they evolved from combining and recycling at least one genetic fragment of prokaryotic ancestry, either archaeal or bacterial, usually with eukaryotic genetic fragments, within a eukaryotic host lineage.

The distribution of S-genes across eukaryotic lineages reveals that 50% of these families (e.g., 303 gene families) are present both in Opimoda and Diphoda and therefore were likely present in LECA (Fig. 1). S-genes with a restricted taxonomic distribution are compatible with their formation at multiple phylogenetic depths, secondary loss in multiple lineages[31] and/or gene fission[32] of ancestral S-genes. The massive burst of S-genes in the earliest diverging eukaryotes may be an outcome of the extensive genome remodeling due to intron invasion[33] and gene duplication in LECA[3].

New essential eukaryotic components

Early S-genes contributed in many important ways to eukaryogenesis. Functional predictions suggest they were involved in cellular processes and signalling; primarily in the ‘O’ (Post-translational modification, protein turnover, chaperones) category, but also in the ‘U’ (Intracellular trafficking, secretion, and vesicular transport), ‘D’ (Cell cycle control and mitosis) and ‘Z’ (Cytoskeleton) categories, in information storage and processing (mainly the ‘A’ (RNA processing and modification), ‘K’ (Transcription), ‘L’ (DNA Replication and

repair) and ‘J’ (Translation) categories), as well as in metabolism (particularly the ‘I’ (Lipid metabolism) category) (Fig. 2).

A detailed gene-by-gene analysis (Fig. 3, Table S1) substantiates the relevance of S-genes to eukaryote evolution. These composite genes are key components of the replisome (families 41894 and 8452), the spliceosome (families 5353, 14116 and 7536), the transcriptional (families 15440, 8572 and 31114) and translational machineries (families 6980, 15594 and 4775), ribosome biogenesis and assembly (families 9105, 9136 and 4331), chromatin and chromosome structure (families 3752, 5196 and 60478), and DNA repair (families 19268, 39836 and 16839) (Fig. 3, Table S1). S-genes augmented the informational machinery during eukaryogenesis by adding new components to existing processes[22–24]. Defence against parasitic nucleic acids such as introns may explain why eukaryotic gene expression requires additional processing steps not observed in prokaryotes[34]. Indeed, dealing with introns was a major function of early arising S-genes, consistent with the notion that introns ‘plagued’ early eukaryotic genomes (Extended Data Fig. 3). Tinkering with the DNA repair system is supported by the following observations in microbiology. Prokaryotic endosymbionts within a free-living prokaryotic host have not been described thus far, indicating that this nested lifestyle is likely difficult to establish. Genotoxicity might be one of many barriers to the success of such endosymbioses[35,36]. During early eukaryogenesis, the DNA within the proto-mitochondrion was likely adversely impacted by the chemically harsh environment resulting from the inclusion of that organelle within its host. In addition, the organelle generated ROS rendering the cellular environment toxic for host DNA, if this genome was not protected by the nuclear membrane. Interestingly, two out of three components of the MRX complex, involved in repairing DNA double-strand breaks using homologous recombination[37], are S-genes (families 18347 and 18341) that provide protection from genotoxicity.

Moreover, S-gene evolution addressed another challenge faced by eukaryotes: early eukaryotic cells were larger and more compartmentalized than individual prokaryotic cells, which presumably limited protein-protein interactions because their interaction requires some form of coordinated intracellular targeting. We report 303 occurrences of physical associations of multiple domains on a single novel eukaryotic gene, whereas these domains are not so tightly connected in prokaryotes. This genetic remodeling guaranteed the direct interaction of these domains once translated into proteins in the eukaryotic cell. In contrast, domains encoded in separate genes are less likely to be able to interact in a larger

compartmentalized cell[38]. Consistent with this notion that S-genes stabilize functional interactions, and assuming that some operons were inherited from the bacterial and archaeal partners, we infer that 20 ancestral prokaryotic operons, encoding functions such as proton transport, transmembrane transport or DNA-templated transcription, fused into S-genes during early eukaryote evolution. The transformation of operons into S-genes facilitates the coordinated expression of interacting proteins and solves the problem of decoupled transcription and translation in eukaryotes. (Extended Data Table 1). The sparse taxonomic distribution of 14 other prokaryotic operons suggests they evolved into S-genes later during eukaryotic evolution or were secondarily lost from eukaryotic lineages.

Many early S-genes are involved in chaperone systems and protein folding that may also have contributed to dealing with an increase in cell complexity[3]. Six S-gene families containing a DnaJ domain and 11 S-genes with isomerase activities act as chaperones and folding catalysts (Table S1). S-genes are also involved in intracellular trafficking such as the Golgi-REG interface vis-à-vis the COPI and COPII coating machineries (families 3724, 3693, 63542 and 7977). Finally, early S-genes were frequently involved in post-translational modification and protein turnover, with at least 13 S-genes belonging to the ubiquitin system and the proteasome. These proteins, although of archaeal origin[39], are known to have diversified via architectural rearrangements in early eukaryotes with the evolution of further complexity in some lineages[40]. In a primitive eukaryotic cell already harbouring complex endomembrane compartments, early developments in post-translational and trafficking systems were likely to have been advantageous.

Early S-genes also contributed metabolic functionality with involvement in lipid transport and metabolism, with six represented in glycerophospholipid metabolism, which is important for membrane biogenesis (Extended Data Fig. 4). Of note, subsequent lineage specific tinkering of metabolic S-genes was an important process as illustrated by their sparse taxonomic distribution (Extended Data Fig. 5).

Overall, the 605 S-genes detected in this analysis (with 303 presumably present in LECA) contributed to important systems and processes in eukaryotes (Fig. 2 and Fig. 3). In the model organism *Saccharomyces cerevisiae*, 46 out 115 S-gene families are essential (Table S1) (a higher ratio when compared to the ratio of essential genes (103) in non symbiogenetic composite gene families (341)). S-genes also have a higher degree in the yeast PPI networks (median = 36.00; 1sr Qu. = 18.50;3rd Qu. = 56.00) than other composite genes

(median = 26.00; 1sr Qu. = 14.00; 3rd Qu. = 45.00), indicating that they associate with a higher number of protein partners (Table S1). This essentiality and high degree in PPI networks of S-genes makes sense because 52 of them encode proteins involved in macromolecular complexes, 34 of which contribute to key eukaryotic informational macromolecular machineries in yeast (Table S1).

Phylogenetic origins of S-genes

Clustering S-gene according to the phylogenetic assignation of their components (e.g., similar to sequences in Archaea, Bacteria, prokaryotes in general or eukaryotes) showed that components do not associate randomly. Very few S-genes (only 7, cluster 6 in Fig. 4) have combined fragments of archaeal and bacterial origins. This result might be surprising if one considers that genetic fragments from these two prokaryotic sources have occurred together in the same genome for about two billion years[10]. In fact, most S-genes (413) contained a component of bacterial origin that is either combined with another bacterial (cluster 2) or eukaryotic cluster (cluster 3), whereas only 46 S-genes with a clear component of archaeal origin were identified (clusters 5, 6 and 7). In order to understand this limited number of S-genes derived from Archaea, we looked in detail at clusters 1 (45 S-genes) and 4 (79 S-genes), which correspond to S-genes with components of prokaryotic origin (i.e., components similar to prokaryotes that we cannot assign to Archaea or Bacteria according to our parameters). We observed that 36 and 55 families in cluster 1 and 4 respectively, contain at least one archaeal sequence in the top three hits of their components (Table S2). These observations suggest that some families in clusters 1 and 4 may contain components of archaeal origin that are identified as prokaryotic because of the limited number of genomes available from Archaea. Nevertheless, S-genes are largely of bacterial origin, whereas S-genes with archaeal components are more rare, which is consistent with the analysis of full-length genes[18,41].

Interestingly, the phylogenetic origin of S-gene components also correlates with functions (Fig. 4). S-genes with archaeal components (clusters 5, 6, and 7 in Fig. 4) (46 S-genes) are mostly associated with informational functions (31 out of 46) (chi square test, adjusted P = 0.00163, Extended Data Fig. 6), whereas S-genes of bacterial origins (clusters 2, 3 and 6 in Fig. 4) are mostly involved in operational functions, typically metabolism (clusters 2, 82 out of 154 S-genes involved in metabolism) (chi square test, adjusted P = 0.02271, Extended Data Fig. 6). S-genes with bacterial and eukaryotic components are enriched in

cellular process and signalling (86 out of 208 S-genes are involved in cellular processes and signalling are present in cluster 3) (chi square test, adjusted P = 0.05529, Extended Data Fig. 6). At first glance, the evolution of S-genes thus seems consistent with the findings by Rivera and Lake[17] on the origin of eukaryotic genes: i.e., genes wholly inherited from an archaeal ancestor are involved in informational functions, whereas genes wholly of bacteria origin are involved in operational functions. However, although this correlation exists for S-genes in relative proportion, when the number of families is considered, S-gene families with bacterial origins encode twice as many informational processes (72) than S-gene families with archaeal origins (31). Thus, there is a large hidden bacterial contribution to the evolution of eukaryotes, beyond operational functions consistent with[18]. Specifically, S-genes with bacterial origins are more involved in RNA processing ('A') and to a lesser extent in transcription ('K') than archaeal components containing S-genes (Extended Data Fig. 7). Of note, 9 families involved in the spliceosome encode a bacterial component, whereas only two show a likely archaeal origin (Table S1). The complementary observation, a possible subgenic hidden contribution of Archaea to eukaryogenesis, is not true: Archaea did not contribute many genetic fragments to S-genes associated with operational genes in eukaryotes. Thus, not only at the gene level[41] but also at the subgenic level, the evolvablity of DNA derived from Archaea appears more limited than the evolvability of DNA derived from Bacteria in nuclear genomes. Whereas S-genes with bacterial components are found in all functional categories, this is not the case for S-genes with archaeal components.

Finally, S-gene formation continued after eukaryotic diversification into major phyla. For example, 32% of S-genes are present in a single eukaryotic lineage (191 families), and could serve as synapomorphies (i.e., adaptive functions) for these groups[32]. In particular among the SAR group, ciliates contain a high proportion of exclusive S-genes (38 families, Extended Data Fig. 5). Ciliates are known for their complex mechanisms of programmed genome rearrangements[42] which facilitate new chimeric gene creation[43]. S-genes in ciliates do not seem to fulfil random functions: i.e., they are mostly involved in cellular processes and signalling (21 S-genes) with 13 playing a role in signal transduction mechanisms (Table S1).

In summary, given the complex nature of eukaryogenesis it is not surprising that valuable genetic information was exploited in many different ways to remodel host cell biology. Our results demonstrate that S-genes were a key part of this process, with over 300 composite sequences formed during the early phases of eukaryogenesis. We propose that these S-gene families helped address many of the challenges faced by early eukaryotes:

enhancing the informational machinery, handling introns, countering genotoxicity within the cell, and ensuring functional protein interactions in a larger, more compartimentalized cell. Moreover, it is surprising that only 46 S-genes contain an archaeal domain, which, on a per-gene basis, is about 10-fold less than donated by Bacteria. Furthermore, in terms of the absolute number of gene families, Bacteria made a twofold greater contribution to informational functions than Archaea. Therefore, fundamental eukaryotic properties do not follow strictly the traditional informational/operational divide for archaeal/bacterial contributions to eukaryogenesis.

Methods

Dataset construction.

We assembled a protein sequence database by downloading every archaeal, viral, and plasmid genome that was annotated as "complete" according to the NCBI Genome database on November 2013 (152, 3769 and 4294 genomes, respectively). We also retrieved 230 eubacterial genomes, with one representative randomly chosen per eubacterial family. Finally, we sampled 38 unicellular eukaryotic genomes across the eukaryotic tree of life: 19 for photosynthetic organisms and 19 that are non-photosynthetic, with a comparable total gene number and phylogenetic diversity in their ribosomal proteins. The resulting 2,192,940 protein sequences were compared pairwise using BLASTP[28] (version 2.2.26) (E-value cutoff 1e-5) (see Table S3 for the list of genomes used).

Domain and functional annotations

Domains were predicted using the Conserved Domain Database[29] (CDD) (version 3.13) (default parameters) and Pfam[30] (version 29.0) (default parameters). Sequences were functionally annotated by the category of their best HmmScan[44] match (version 3.1) (E-value cutoff 1e-5) against eukaryotic EggNog database[45] (version 4.5). *S.cerevisiae* genes were annotated with the DEG database[46] (version 13.3) and protein protein interactions with the BioGRID database[43] (version 3.4.136).

Detection of S-gene families.

Composite genes were detected with FusedTriplets[27] (E-value < 1e-5) by scanning the BLASTP output. All sequences were also clustered into gene families according to the method used in[41]. Briefly, an undirected graph was constructed in which each node corresponds to a sequence and two nodes are linked if the corresponding sequences show a BLAST hit with an e-value < 1e-5, <= 30% sequence identity and a mutual sequence overlap >= 80%. Connected components in this graph were considered gene families. Families with only eukaryotic sequences, at least three different eukaryote species, more than 50% of genes detected as composite by FusedTriplets and with at least two domains were kept for further analysis. In order to be as comprehensive as possible, each gene was blasted against an

extensive prokaryotic dataset. (2,982 prokaryotic genomes, 8,422,211 sequences). If all sequences of a family had no full-length homologs (i.e, no mutual alignment cover > 80%) but show partial similarity with prokaryote sequences, the composite family was considered as S-gene family.

For each S-gene, prokaryotic component sequences were clustered into component families according to the following rule: if two component sequences overlapped by more than 70% of their lengths on the protein composite, they belonged to the same component family. A refining procedure has been performed in order to merge overlapping and/or nested components families. Two component families were merged if one family was included by more than 70% of its length into the other one.

Component families were assigned a broad phylogenetic origin corresponding to their taxonomic composition. If the ten best prokaryotic component sequences, according to their BLASTP bitscore against the composite gene, matched with the same prokaryotic domain (e.g., Archaea or Bacteria), we considered the component to have more specifically originated from that prokaryotic domain. If component family contained less than ten sequences or if archaeal and bacterial sequences were both present among the ten best sequences, we considered the component to originate from prokaryotes.

Operon-like composite detection

Operon-like composites were detected using the ProOpDB database[47] where 191 out of 382 genomes used in this study are referenced in the database. Briefly, if two components of a composite were found in an operon in the same prokaryote, the composite is considered as an operon-like composite.

References

1. Méheust, R., Zelzion, E., Bhattacharya, D., Lopez, P. & Bapteste, E. Protein networks identify novel symbiogenetic genes resulting from plastid endosymbiosis. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 3579–84 (2016).

2. Kaessmann, H. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* **20,** 1313–26 (2010).

3. Makarova, K. S., Wolf, Y. I., Mekhedov, S. L., Mirkin, B. G. & Koonin, E. V. Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. *Nucleic Acids Res.* **33,** 4626–38 (2005).

4. McLysaght, A. & Guerzoni, D. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **370,** 20140332 (2015).

5. Kawai, H. *et al.* Responses of ferns to red light are mediated by an unconventional photoreceptor. *Nature* **421,** 287–90 (2003).

6. Karnkowska, A. *et al.* A Eukaryote without a Mitochondrial Organelle. *Curr. Biol.* (2016). doi:10.1016/j.cub.2016.03.053

7. Szathmáry, E. & Smith, J. M. The major evolutionary transitions. *Nature* **374,** 227–32 (1995).

8. Szathmáry, E. Toward major evolutionary transitions theory 2.0. *Proc. Natl. Acad. Sci. U. S. A.* 1421398112– (2015). doi:10.1073/pnas.1421398112

9. Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521,** 173–179 (2015).

10. McInerney, J. O., O'Connell, M. J. & Pisani, D. The hybrid nature of the Eukaryota and a consilient view of life on Earth. *Nat. Rev. Microbiol.* **12,** 449–55 (2014).

11. Williams, T. A., Foster, P. G., Cox, C. J. & Embley, T. M. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* **504,** 231–6 (2013).

12. O'Malley, M. A. The first eukaryote cell: an unfinished history of contestation. *Stud. Hist. Philos. Biol. Biomed. Sci.* **41,** 212–24 (2010).

13. Lane, N. & Martin, W. The energetics of genome complexity. *Nature* **467,** 929–34 (2010).

14. Timmis, J. N., Ayliffe, M. A., Huang, C. Y. & Martin, W. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat. Rev. Genet.* **5,** 123–35 (2004).

15. O'Malley, M. A. Endosymbiosis and its implications for evolutionary theory. *Proc. Natl. Acad. Sci. U. S. A.* (2015). doi:10.1073/pnas.1421389112

16. Archibald, J. M. Endosymbiosis and Eukaryotic Cell Evolution. *Curr. Biol.* **25,** R911–R921 (2015).

17. Rivera, M. C., Jain, R., Moore, J. E. & Lake, J. A. Genomic evidence for two functionally distinct gene classes. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 6239–44 (1998).

18. Cotton, J. A. & McInerney, J. O. Eukaryotic genes of archaebacterial origin are more important than the more numerous eubacterial genes, irrespective of function. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 17252–5 (2010).

19. Esser, C. *et al.* A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Mol. Biol. Evol.* **21,** 1643–60 (2004).

20. Koonin, E. V & Yutin, N. The dispersed archaeal eukaryome and the complex archaeal ancestor of eukaryotes. *Cold Spring Harb. Perspect. Biol.* **6,** a016188 (2014).

21. Koonin, E. V. Origin of eukaryotes from within archaea, archaeal eukaryome and bursts of gene gain: eukaryogenesis just made easier? *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **370,** (2015).

22. McInerney, J., Pisani, D. & O'Connell, M. J. The ring of life hypothesis for eukaryote origins is supported by multiple kinds of data. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **370,** 20140323 (2015).

23. Koumandou, V. L. *et al.* Molecular paleontology and complexity in the last eukaryotic common ancestor. *Crit. Rev. Biochem. Mol. Biol.* **48,** (2013).

24. Mast, F. D., Barlow, L. D., Rachubinski, R. A. & Dacks, J. B. Evolutionary

mechanisms for establishing eukaryotic cellular complexity. *Trends in Cell Biology* **24,** 435–442 (2014).

25. Gabaldón, T. & Pittis, A. A. Origin and evolution of metabolic sub-cellular compartmentalization in eukaryotes. *Biochimie* **119,** 262–8 (2015).

26. Koonin, E. V. The origin and early evolution of eukaryotes in the light of phylogenomics. *Genome Biol.* **11,** 209 (2010).

27. Jachiet, P.-A., Pogorelcnik, R., Berry, A., Lopez, P. & Bapteste, E. MosaicFinder: identification of fused gene families in sequence similarity networks. *Bioinformatics* **29,** 837–844 (2013).

28. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–402 (1997).

29. Marchler-Bauer, a. *et al.* CDD: NCBI's conserved domain database. *Nucleic Acids Res.* **43,** D222–D226 (2014).

30. Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **40,** D290–301 (2012).

31. Ku, C. *et al.* Endosymbiotic origin and differential loss of eukaryotic genes. *Nature* **524,** 427–437 (2015).

32. Leonard, G. & Richards, T. A. Genome-scale comparative analysis of gene fusions, gene fissions, and the fungal tree of life. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 21402–7 (2012).

33. Irimia, M. & Roy, S. W. Origin of spliceosomal introns and alternative splicing. *Cold Spring Harb. Perspect. Biol.* **6,** (2014).

34. Madhani, H. D. The frustrated gene: origins of eukaryotic gene expression. *Cell* **155,** 744–9 (2013).

35. Gross, J. & Bhattacharya, D. Uniting sex and eukaryote origins in an emerging oxygenic world. *Biol. Direct* **5,** 53 (2010).

36. Bernstein, H., Byerly, H. C., Hopf, F. A. & Michod, R. E. Genetic damage, mutation, and the evolution of sex. *Science* **229,** 1277–81 (1985).

37. Symington, L. S. DNA repair: Making the cut. *Nature* **514,** 39–40 (2014).

38. Shieh, Y.-W. *et al.* Operon structure and cotranslational subunit association direct protein assembly in bacteria. *Science* **350,** 678–680 (2015).

39. Humbard, M. A. *et al.* Ubiquitin-like small archaeal modifier proteins (SAMPs) in Haloferax volcanii. *Nature* **463,** 54–60 (2010).

40. Grau-Bové, X., Sebé-Pedrós, A. & Ruiz-Trillo, I. The eukaryotic ancestor had a complex ubiquitin signalling system of archaeal origin. *Mol. Biol. Evol.* **32,** 726–39 (2014).

41. Alvarez-Ponce, D., Lopez, P., Bapteste, E. & McInerney, J. O. Gene similarity networks provide tools for understanding eukaryote origins and evolution. *Proc. Natl. Acad. Sci. U. S. A.* **110,** E1594–603 (2013).

42. Chen, X. *et al.* The Architecture of a Scrambled Genome Reveals Massive Levels of Genomic Rearrangement during Development. *Cell* **158,** 1187–1198 (2014).

43. Chatr-Aryamontri, A. *et al.* The BioGRID interaction database: 2015 update. *Nucleic Acids Res.* **43,** D470–8 (2015).

44. Eddy, S. Profile hidden Markov models. *Bioinformatics* **14,** 755–763 (1998).

45. Huerta-Cepas, J. *et al.* eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44,** D286–93 (2015).

46. Luo, H., Lin, Y., Gao, F., Zhang, C.-T. & Zhang, R. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Res.* **42,** D574–D580 (2013).

47. Taboada, B., Ciria, R., Martinez-Guerrero, C. E. & Merino, E. ProOpDB: Prokaryotic Operon DataBase. *Nucleic Acids Res.* **40,** D627–31 (2012).

48. Derelle, R. *et al.* Bacterial proteins pinpoint a single eukaryotic root. *Proc. Natl. Acad. Sci.* 201420657 (2015). doi:10.1073/pnas.1420657112

49. de Duve, C. The origin of eukaryotes: a reappraisal. *Nat. Rev. Genet.* **8,** 395–403 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgments

E.B. is funded by European Research Council (FP7/2007-2013 Grant Agreement #615274). D.B is grateful to the Rutgers University School of Environmental and Biological Sciences and members of the Genome Cooperative at SEBS for supporting this research.

Author Contributions

E.B, P.L and R.M designed the study. R.M performed analyses. All authors analyzed the data and wrote the paper.

Author Information

Fasta sequences of S-genes can be found at http://www.evol-net.fr/downloads/SgenesEukaryogenesis.zip. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.M at raphael.meheust@gmail.com.

Legends:

Figure 1. Putative phylogeny of eukaryotes, based on[48], that shows the distribution of 605 S-gene families. Family evolution reconstruction was performed using Dollo parsimony. The four boxes correspond to the number of families involved in metabolism (red), information storage and processing (blue), cellular processes and signalling (green), and poorly characterized processes (white).

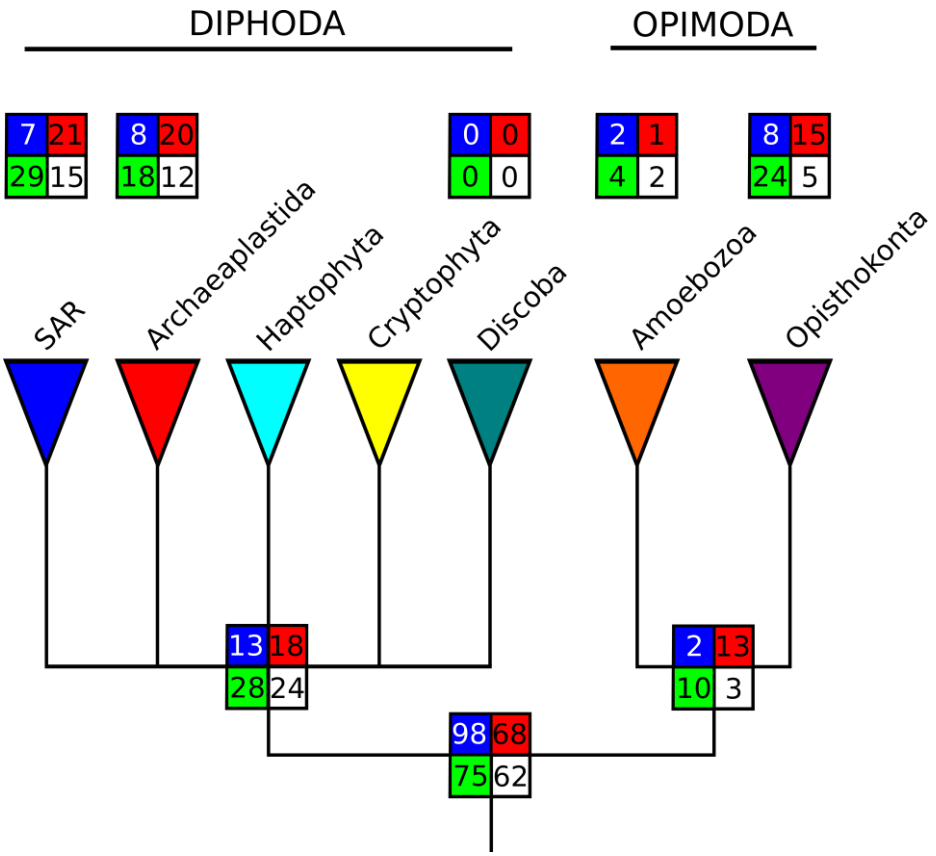


Figure 2. Functional annotation of the 605 S-genes based on COG categories. S-gene families were divided into early (S-genes found in both Opimoda and Diphoda, 303 gene families, in pink) intermediate (S-genes found either in Opimoda or Diphoda, 111 gene families, in green) and lineage-specific (S-genes found in one eukaryotic supergroups, 191 gene families, in blue) (COG category definitions can be found here: http://eggnogdb.embl.de/download/eggnog_4.5/COG_functional_categories.txt).

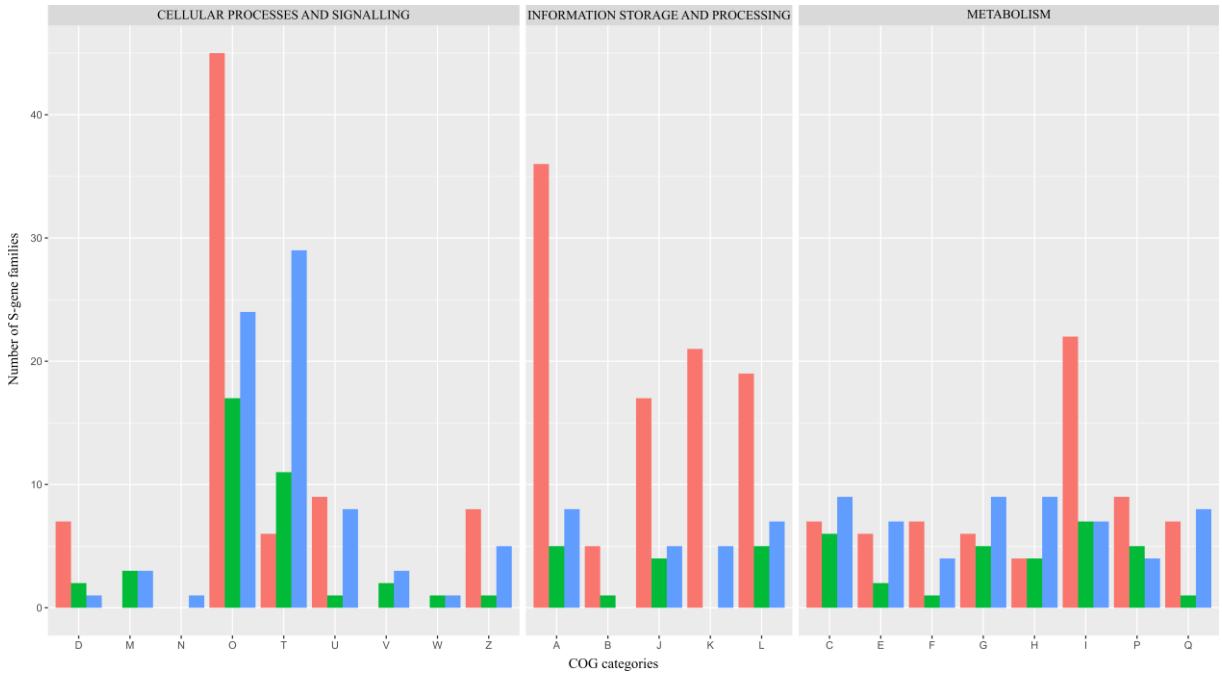


Figure 3. Mapping of the functions of 605 S-genes in a eukaryotic cell (figure adapted from[49]). Numbers in red correspond to functions containing essential S-genes in yeast.

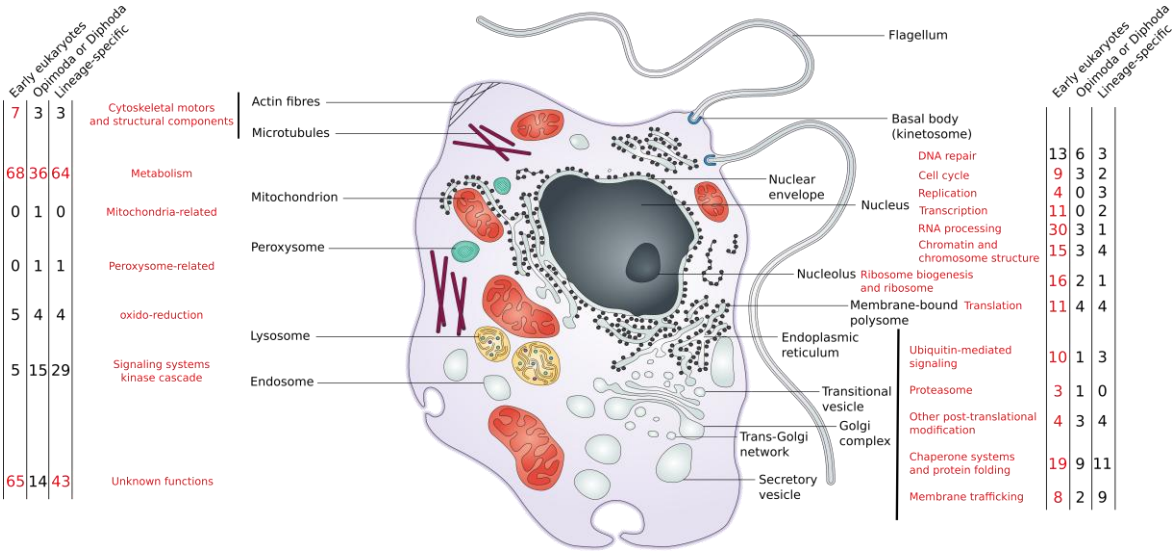
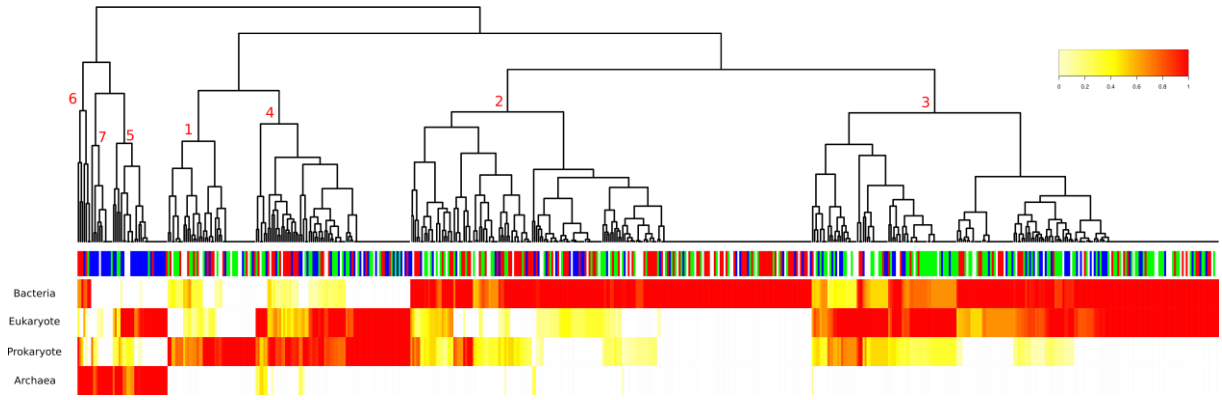
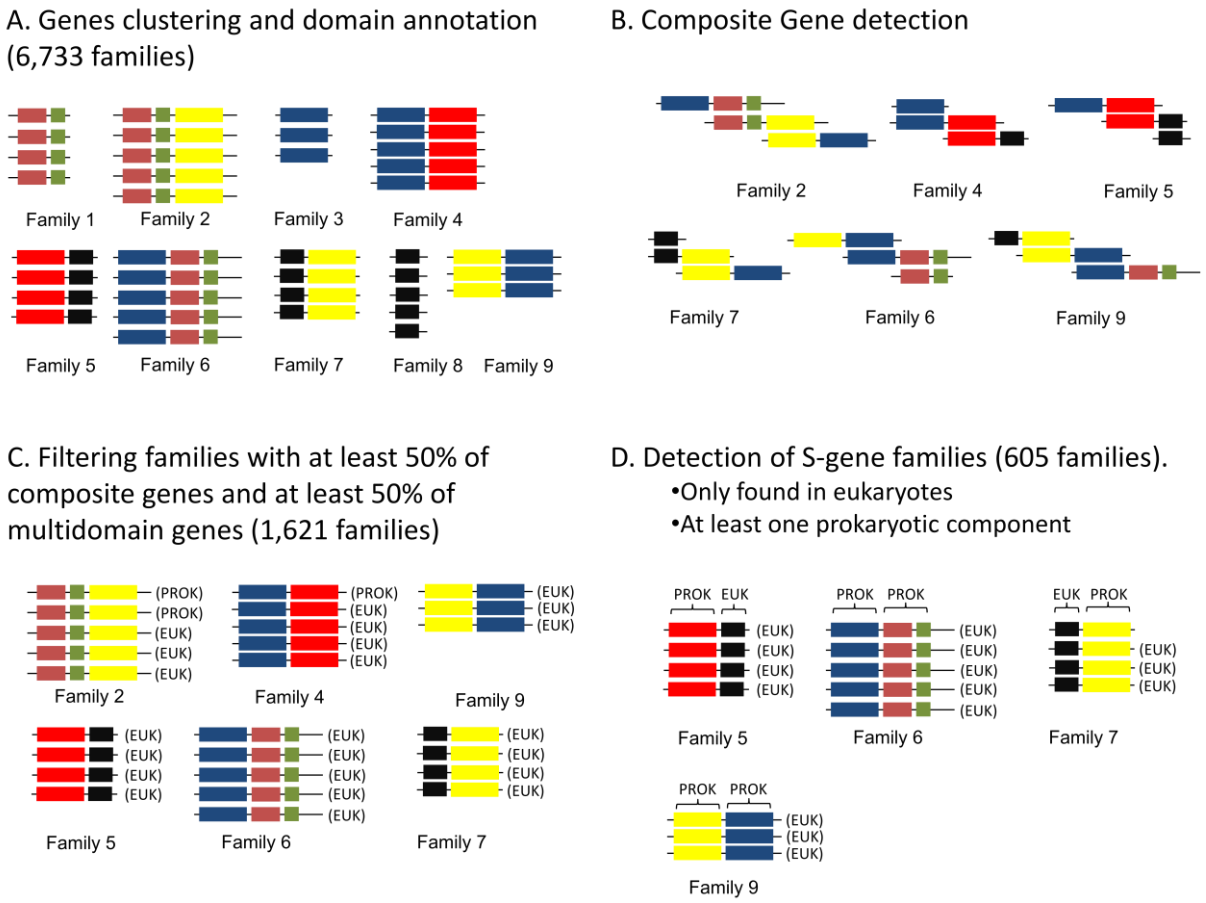


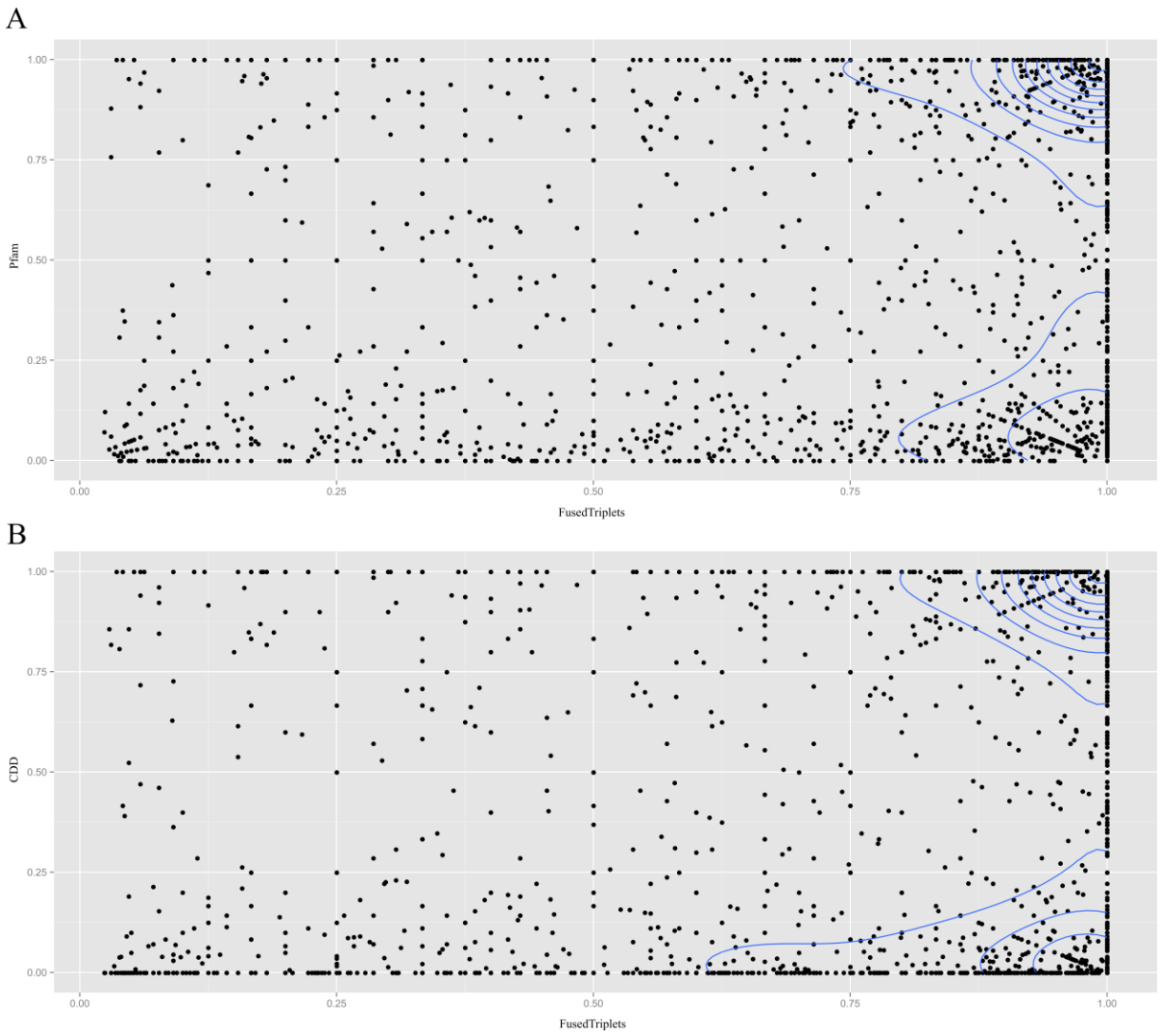
Figure 4. Hierarchical clustering of S-gene families according to their component origins. The heatmap represents the ratio of genes in a given family (columns) that have at least one component of a given origin (eukaryotic, archaeal, bacterial or prokaryotic; the rows). White lines correspond to the absence of a component from a given origin in every gene in the given S-gene family. The colored lines correspond to the presence of at least one component of the given origin in a given percentage of genes in the given S-gene family (red lines denote that all [100%] genes contains a given origin component). The colored top bar indicates the functional annotation.



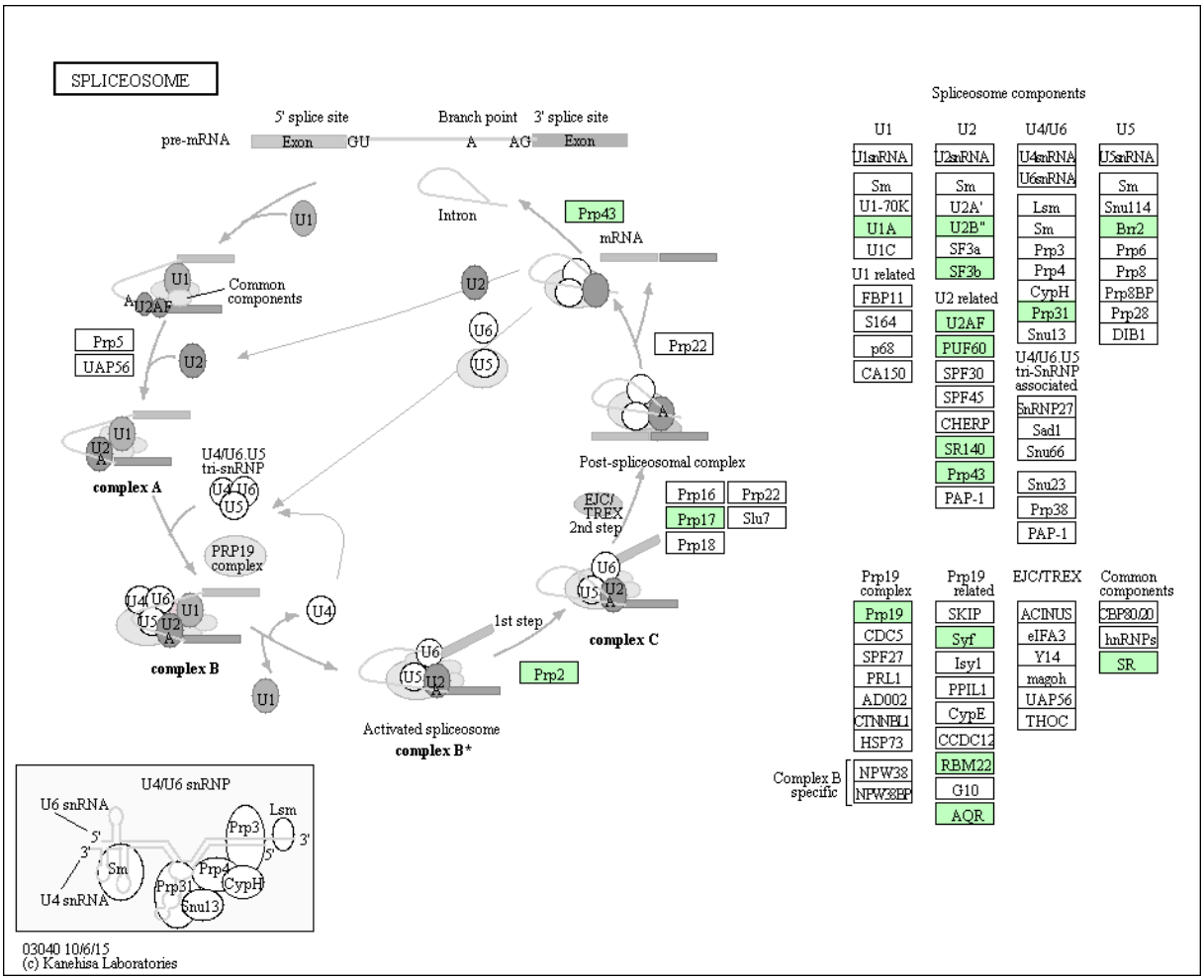
Extended Data Figure 1. Protocol used for the detection of S-gene families. A. Sequences have been clustered in gene families. B. Composite genes have been detected using FusedTriplets. C. Gene families detected as composite and having at least two domains have been kept for further analysis. D. Composite gene families only found in eukaryotes and having at least one component of prokaryotic origin were considered as S-gene families.



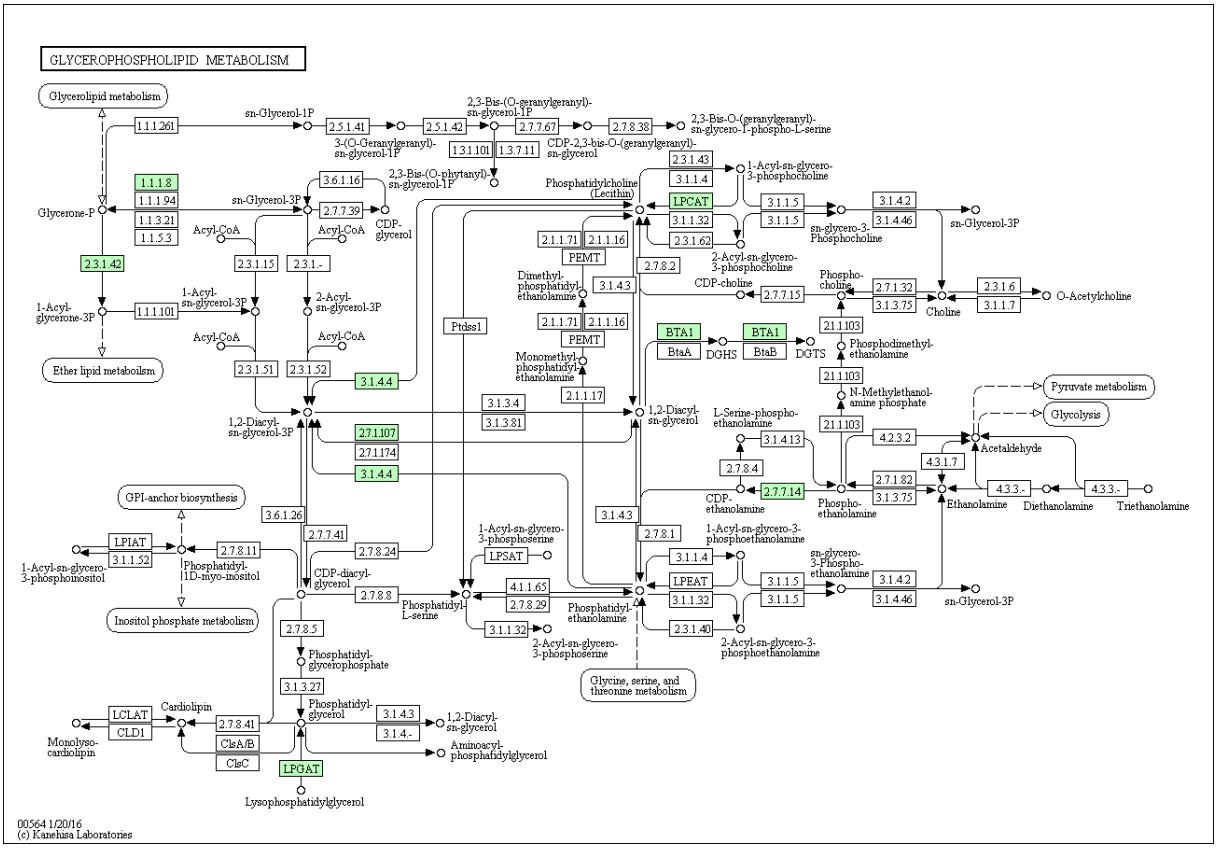
Extended Data Figure 2. Two-dimensional density graph of percentage of families detected as composite according to fusedTriplets (x-axis) and with at least two known domains according to Pfam (A) and CDD (B) (y-axis). Each point corresponds to a family. Since these points can stack, isodensity lines in blue delimit regions having constant density.



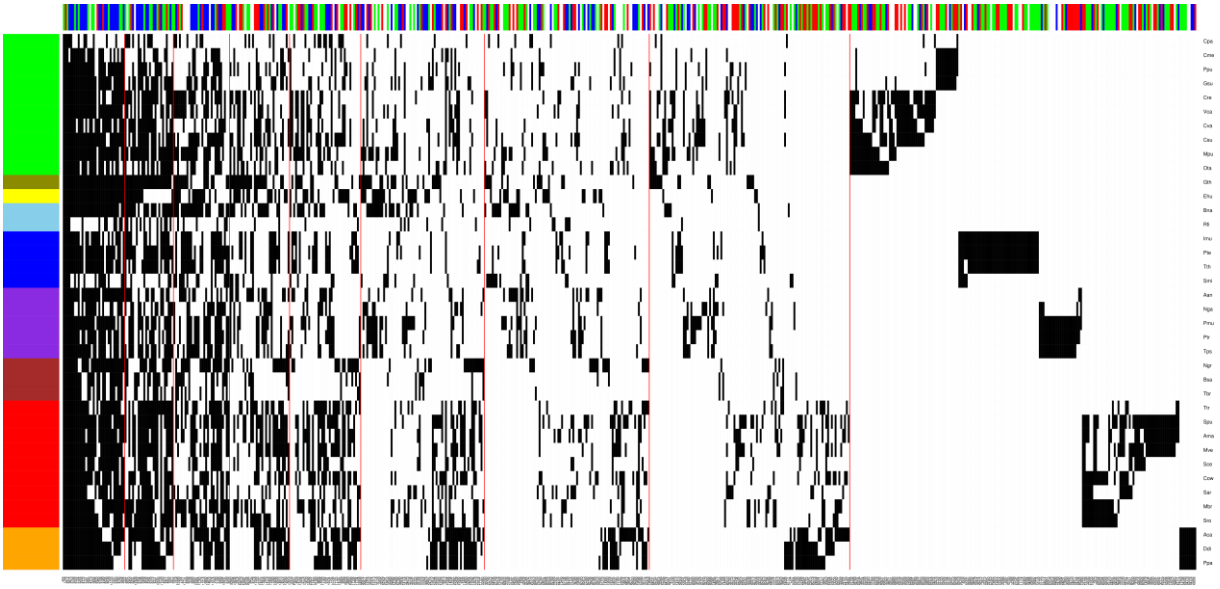
Extended Data Figure 3. Kegg map of the spliceosome showing the 19 S-genes.



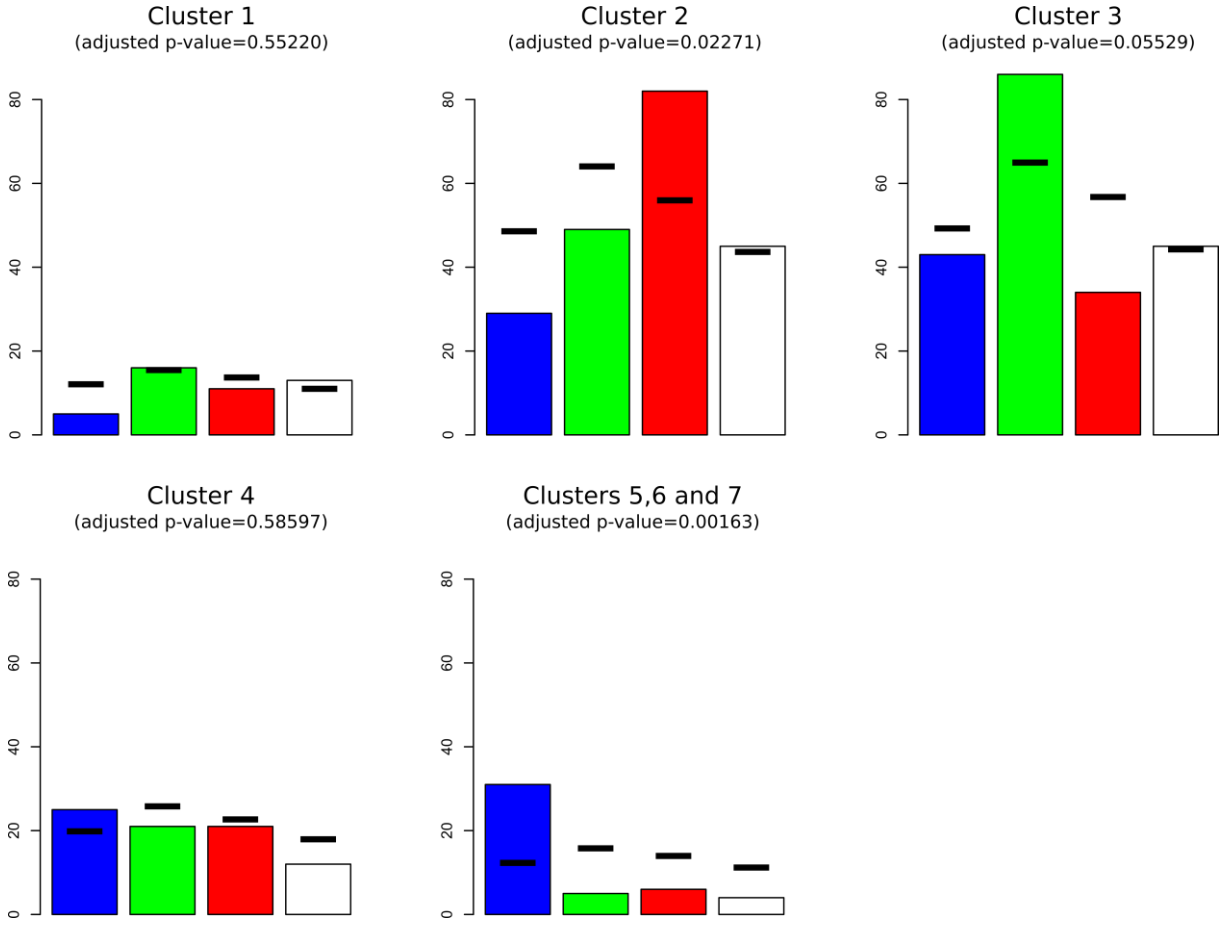
Extended Data Figure 4. Kegg map of the glycerophospholipid pathway showing the 6 S-genes.



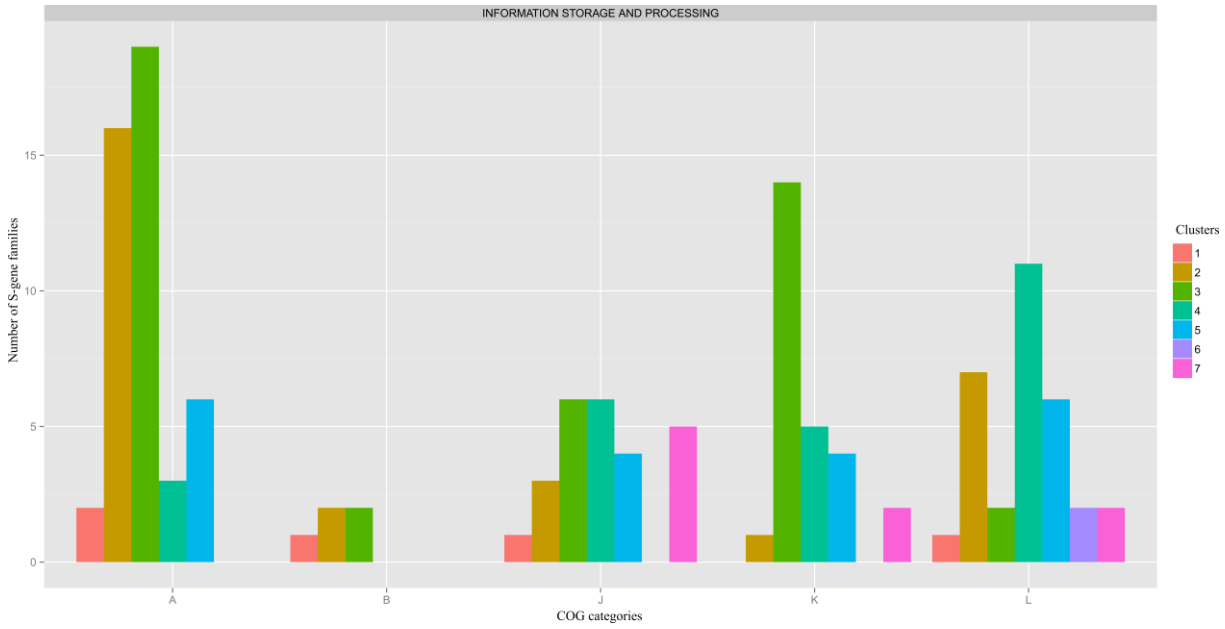
Extended Data Figure 5. Distribution of 605 S-gene families across eukaryotic species. The heatmap represents the presence (black line) or absence (white line) of a given S-gene family in a eukaryotic species (each line represents a given species, each column represents a given family). Eukaryotic species are colored with respect to their classification into major supergroups (light green: Archaeaplastida, dark yellow: Cryptophytes, yellow: Haptophytes, light blue: Rhizaria, blue: Alveolates, purple: Stramenopiles, brown: excavates, red: Opisthokonts, orange: Amoebozoa). The colored top bar indicates the functional annotation of the S-gene families according to COG (red: metabolism, blue: information storage and processing, green: cellular processes and signalling, white: poorly characterized). The heatmap is structured along its x-axis, based on the number of eukaryotic supergroups containing the S-gene family, binned in decreasing order (from the left: S-gene families distributed in all 9 supergroups, to the right: S-gene families present in a single supergroup but in at least three species; each bin is separated by a thin red line).



Extended Data Figure 6. Chi square test on the distribution of COG categories. Color code is the same as in Fig. 4. Barplots correspond to observed proportions while black lines correspond to expected proportions.



Extended Data Figure 7. Functional annotation of the S-genes involved in information storage and processing according to the different clusters in Fig. 4.



Extended Data Table 1. The 34 operon-like composite families, along with the prokaryotic phyla where these operons were detected.

Family	Cluster	Distribution	Crenarchaeota	Euryarchaeota	Korarchaeota	Thaumarchaeota	Actinobacteria	Aquificae	Bacteroidetes	Chlamydiae	Chlorobi	Chloroflexi	Chrysiogenetes	Cyanobacteria	Deferribacteres	Deinococcus-Thermus	Dictyoglomi	Elusimicrobia	Fibrobacteres	Firmicutes	Fusobacteria	Gemmatimonadetes	Planctomycetes	Proteobacteria	Spirochaetes	Synergistetes	Thermotogae	Verrucomicrobia
9304	1	Early						1			1													1				
4453	2	Early	1	2			19			1		2		3		1								22				2
7614	2	Early					11																	5				
8629	2	Early		1			27		2		1		1						1	1	1		1	20				2
8949	2	Early		8			1																	12				
12311	2	Early	12	6			17			1			1							2				13				
12806	2	Early		5			1		1					1				1				1						2
12884	2	Early																						1				1
12885	2	Opisthokonta										1					1			1				1		1	1	
14803	2	Early												1														
15326	2	Early					1																					
16965	2	Diphoda					1									1												
21942	2	Opimoda					7					2								2				9				1
21948	2	Diphoda												1										12				
21962	2	Early	10				3	1				1	1		1	1	1	1		8	1			21	1	1		
24670	2	Early					1																	1				
25132	2	Opisthokonta	10																									
26810	2	Early																					1					
26839	2	Diphoda																						1				
26893	2	Diphoda					16													1				2				
39861	2	Opimoda	2		1		2						1							4				1	1	1	1	
43676	2	Archaeaplastida		1			9			1			1			2				5		1	1	13				
43725	2	Archaeaplastida																		1				3				
44677	2	SAR																						2				
45806	2	Amoebozoa					1																	5				
50538	2	Early																					1					
10810	3	Early	18	23		1	30					2			1	1	1	1	1	7	1	1		21	1			4
4588	4	Early		2																								
7080	4	Early		2																								
7988	4	SAR		3																								
15589	4	Opisthokonta		2																								
16841	4	Early		2																								
41766	5	Early	23	22																								
9066	7	Early	23	22																								

Table S1. Annotation of the 605 S-gene families that were detected in our study.

Table S2. Detailed origin of prokaryotic S-gene components. Fam: S-gene family, Cpt: component, Cluster: cluster number according to Fig. 4, Bacteria: number of hits from Bacteria, Archaea: number of hits from Archaea. For each S-gene component, the rank in the BLAST search and the taxonomic assignation of the 25 sequences with the best hits to that component were reported (Aci: *Acidobacteria*, Act: *Actinobacteria*, Aqu: *Aquificae*, Arm: *Armatimonadetes*, Bac: *Bacteroidetes*, Chl: *Chloroflexi*, Cre: *Crenarchaeota*, Cya: *Cyanobacteria*, Def: *Deferribacteres*, Dei: *Deinococcus-Thermus*, Eur: *Euryarchaeota*, Fir: *Firmicutes*, Fus: *Fusobacteria*, Gem: *Gemmatimonadetes*, Ign: *Ignavibacteriae*, Nit: *Nitrospirae*, Pla: *Planctomycetes*, Pro: *Proteobacteria*, Spi: *Spirochaetes*, Syn: *Synergistetes*, Ten: *Tenericutes*, Tha: *Thaumarchaeota*, The: *Thermotogae*, Ver: *Verrucomicrobia*, roo: Unknown). Red cells correspond to bacterial phyla while blue cells correspond to archaeal phyla. When only one S-gene component is described, the unrepresented S-gene components from the S-gene family are either exclusively found in photosynthetic eukaryotes, or have diverged too much to be confidently assigned to a prokaryotic group.

Table S3. List of 38 eukaryote genomes and the 382 prokaryotic genomes used in our comparative analysis

Hundreds of novel chimeric genes and of symbiogenetic genes with bacterial origins contributed to Halobacterial evolution

In prep.

Raphaël Méheust[1], Robertson Thane Papke[2], François-Joseph Lapointe[3], Philippe Lopez[1] and Eric Bapteste[1]

[1]Sorbonne Universités, UPMC Univ Paris 06, Institut de Biologie Paris Seine, Centre National de la Recherche Scientifique, Unité Mixte de Recherche 7138 Evolution Paris Seine, 75005, Paris, France

[2]Department of Molecular and Cell Biology, University of Connecticut, 06269 Storrs, CT, USA

[3]Département de sciences biologiques, Université de Montréal, Montréal, Québec, Canada

Introduction

Halobacteria are Archaea which, unlike their ancestors thriving in anaerobic environments, live in oxygenic, highly salted niches. This major transition in lifestyle required that this lineage faced at least two challenges. It involved numerous changes in the genomes and in the physiology of the first Halobacteria (a process termed 'halogenesis'), as well as subsequent genomic optimizations (e.g. in particular, proteins encoded within Halobacteria show lower isoelectric points than their homologs outside this group). While these latter changes can result from point mutation, abundant lateral gene transfers from bacteria have repeatedly been invoked to explain the evolution and adaptation to oxygenic lifestyle of this fascinating archaeal lineage (*1*). Bacterial contribution was observed in phylogenetic studies, largely focused on the acquisition by Halobacteria of full-sized genes from bacterial donors, either via a sudden and massive introgressive process (*2, 3*), or in a more piecemeal fashion (*4, 5*). A thousand gene families with bacterial origins were thus detected in the halobacterial group (*2, 3*). However, additional events of gene remodeling leading to the creation of novel genes, possibly contributing to halogenesis and to the subsequent evolution of halobacteria, were not as systematically explored thus far. Yet, gene remodeling has been described in prokaryotes, mainly as a result of fusion and fission of genes (*6*). Moreover, the transfer of genetic fragments (e.g. domains), i.e. subgenic regions shorter than entire genes, have also been reported for prokaryotes. This latter process of genetic acquisition could in principle be followed by genomic rearrangements when the laterally acquired domains combine with genetic material already present in their new host genomes. In eukaryotes, this process led to the evolution of symbiogenetic genes, when subgenic regions from endosymbionts merged together or with the host DNA in the nucleus (*7*). In the case of photosynthetic eukaryotes, 67 such S-genes with possible adaptive functions, were recently reported (*7*).

Importantly, the detection of reticulate sequence evolution, such as the fusion and recycling of domains derived from heterologous proteins, is best studied using network approaches than conventional phylogenetic approaches. Here, we used sequence similarity networks (*8*) that rely on reconstruction of both full and partial (i.e., protein domain) sequence relationships using pairwise protein similarity values to test whether gene remodeling could have been involved in the emergence of Halobacteria and in their subsequent evolution. We report the creation of hundreds of novel chimeric genes, both early in the evolution of Halobacteria (during Halogenesis) and later in diverged Halobacterial groups. We distinguish 3 classes of such chimeric genes : genes derived from material from archaeal genomes, genes derived from genetic material from prokaryotes that one cannot confidently assigned to bacteria or archae, and genes derived (at least in part) from material from bacterial genomes (Halobacterial S-genes). These latter genes constitute a different gene pool from the laterally acquired bacterial genes detected by Nelson-Sathi, and hence reveal an additional substantial bacterial contribution to the evolution of Halobacteria. All these novel chimeric genes, S-genes and other composite genes derived from Archaea (or more generally prokaryotes), showed significant resident time in halophiles genomes, since their taxonomic distribution is not random and since their isolectectric point were optimized to allow their encoded proteins to operate in salty environments. Importantly, while these novel composite genes are largely involved in metabolic functions, many of them also played a role in adapting Halobacteria to an oxygenic lifestyle.

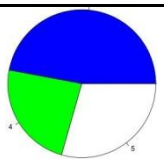
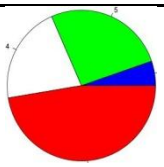
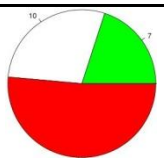
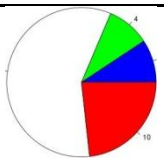
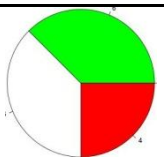
Results

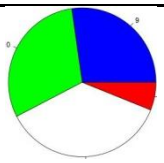
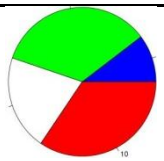
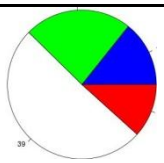
Detection of composite genes and S-genes in Halobacteria (Table 1, Fig. S1, Fig. S2). We clustered 1,816,486 archaeal proteins from 802 genomes into 49,269 families. 6,417 of them (132,458 proteins) contain only proteins from at least 3 different genomes of Halobacteria. These 132,458 proteins have been further aligned over an extended bacterial database of 7,239,663 sequences from 2,078 bacterial genomes in order to removed families with full-length similarities with bacterial proteins. From the 6,417 families, 5,558 have been kept and are therefore good candidates for novel, clade specific genes because they likely originated during or after the emergence of Halobacteria. We tested whether these exclusively halobacterial genes were composite, i.e. whether some of their constitutive regions matched with unrelated gene families (in particular in 7,239,663 bacterial sequences). We combined this detection of component and composite genes) with an additional step of domain annotation. This protocol returned 320 composite gene families, exclusive to Halobacteria.

We classified these families into 3 major groups, based on the phylogenetic assignation of their components (Table 1 and Fig. S1). First, there were 68 families of composite genes (clusters 1 and 10 in Table 1 and Fig. S1) that exclusively combined components found only in archaeal genomes. Second, there were 126 composite gene families, which presented at least one component of bacterial origin (clusters 3, 4, 6 and 9 in Table 1 and Fig. S1). Therefore, we labeled these genes, S-genes. Only 7 of these S-genes correspond to a gene family amongst the 1089 laterally acquired bacterial genes described by Nelson-Sathi (*2*). This very limited overlap means that our S-genes are *bona fide* genetic innovations (Table S1), and point to an additional significant bacterial contribution to the evolution of Halobacteria. Taxonomic assignation of these bacterial components by BLAST comparisons suggests that many independent bacterial sources might have been involved as donors of these recycled fragments (Fig. S2). Finally, clusters 2, 5, 7 and 8 correspond to 136

families with components of prokaryotic origins (i.e., components similar to prokaryotes that we cannot assign to Archaea or Bacteria according to our parameters). These clusters may contain additional S-genes with components of bacterial origins.

Table 1. Classification of the 320 chimeric families found in Halobacteria according to their component (domain) origins. Pie charts correspond to the distribution of functional annotations of the chimeric families for each class (blue: information storage and processing, red: metabolism, white: poorly characterized, green: cellular processes and signaling).

Clusters	Domain A origin	Domain B origin	Number of families	Pie chart of cog categories
1	**Halobacteria**	**Archaea**	51	
10	**Archaea**	**Archaea**	17	
9	**Archaea**	**Bacteria**	19	
4	**Bacteria**	**Bacteria**	35	
3	**Bacteria**	**Halobacteria**	43	
7	**Prokaryote**	**Prokaryote**	16	

5 and 8	**Prokaryote**	**Archaea**	33	
6	**Prokaryote**	**Bacteria**	29	
2	**Prokaryote**	**Halobacteria**	77	

S-genes are significantly involved in metabolism (Fig. 1, Table S1). Functional analysis showed that the 126 S-genes are enriched in metabolic functions (47 out of 126, one-sided Fisher test, P-value=1.681e-09). This result adds further evidence that bacteria contributed to metabolic functions of Halobacteria (*2*, *1*) and that metabolic bacterial genes can be generally recycled in genetic mergers (*9*). Looking in details (Fig. 1), all metabolic categories are more represented in S-genes than in the 2 other major classes of chimeric genes except for Q (« Secondary metabolites biosynthesis, transport and catabolism ») and E (" Amino acid transport and metabolism ") categories (Fig. 1). S-gene families are particularly involved in carbohydrate transport and metabolism (G category in Fig. 1).

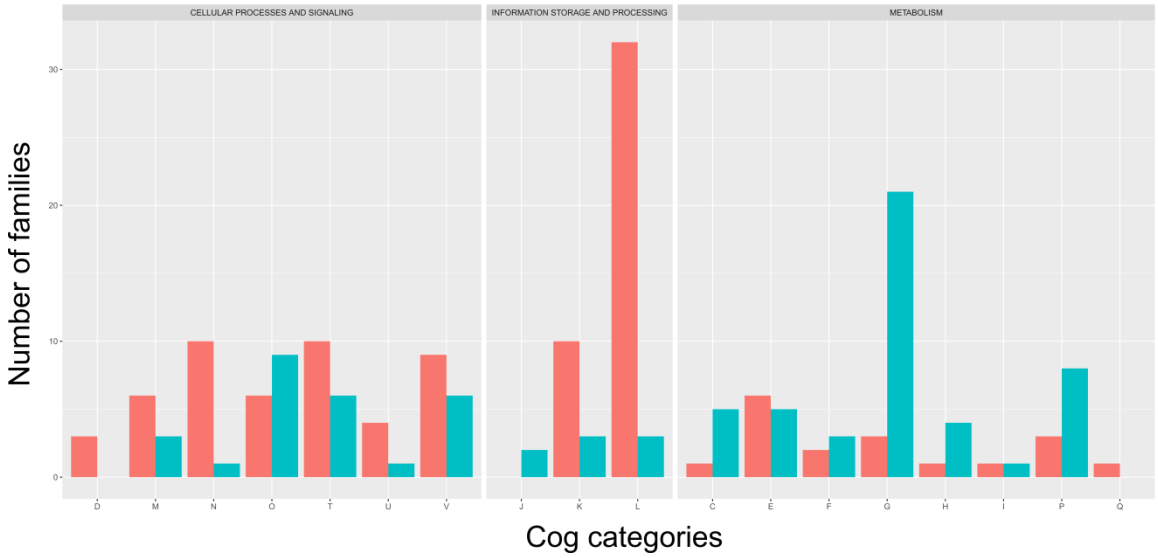


Figure 1: Barplot of functional annotation of the 126 S-gene families (blue) and other chimeric families (red). (COG category definitions can be found here: http://eggnogdb.embl.de/download/eggnog_4.5/COG_functional_categories.txt).

Most of chimeric genes are sparsely distributed but the few conserved families contain genes involved in salt and aerobic lifestyle (Fig 2: family distribution). The distribution of the 320 chimeric families across halobacterial genomes shows that most novel composite gene families are sparsely distributed (Fig. 2). Interestingly, this sparse distribution is not random with respect to currently recognized groups of Halobacteria. We used a Mantel test (P-value=0.001) to confirm that chimeric genes were mainly shared by multiple genomes from the same Halobacterial groups. This type of sharing means that chimeric genes have persisted in these groups for a certain period of time, and therefore have likely some adaptive value. Moreover, the distribution of these chimeric genes is not strictly limited group specific: while chimeric genes are mostly shared by related genomes, they are also laterally transferred between halobacterial groups. (i.e. the heatmap show that composite families, which are well conserved in some lineages are also present in several genomes of other Halobacterial groups).

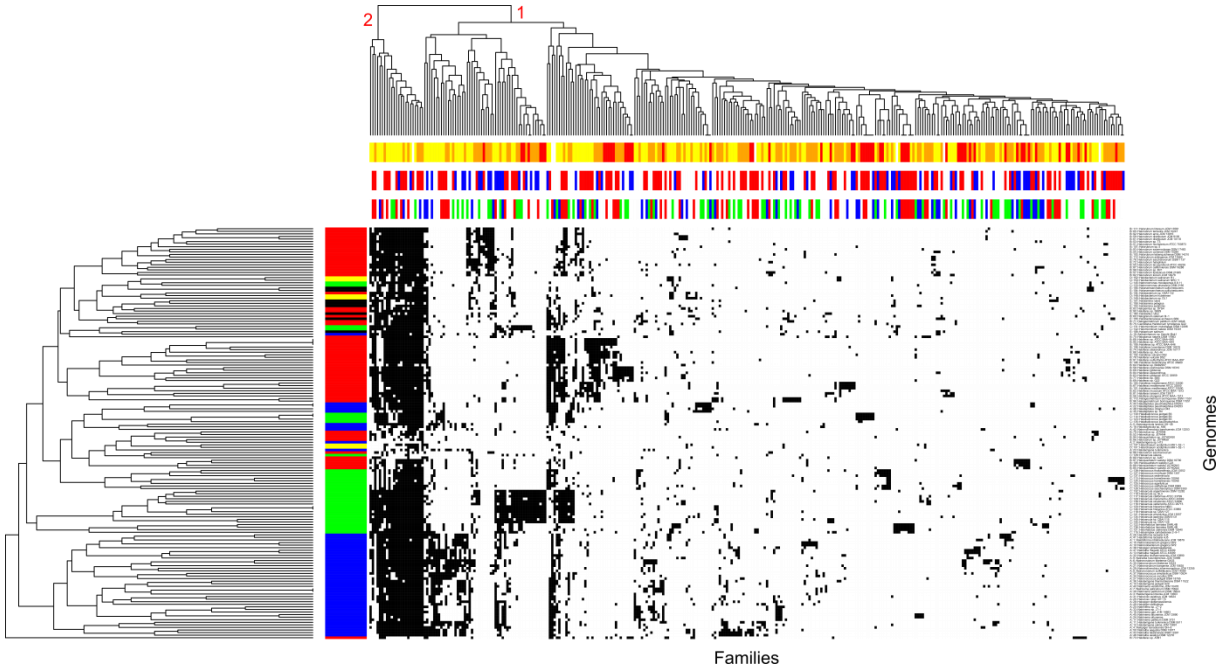


Figure 2. Distribution of the 320 chimeric gene families in Halobacteria. The heatmap represents the presence (black line) or absence (white line) of a given family in Halobacteria genomes (each line represents a given genome, each column represents a given family). Halobacteria genomes are colored with respect to their classification into major clades (red: clade B, blue: clade A, green: clade C, yellow: clade D and black: unknown). The colored top bar on families indicates the functional annotation of the families according to COG categories (red: metabolism, blue: information storage and processing, green: cellular processes and signalling, white: poorly characterized). A hierarchical clustering has been performed both on columns and rows.

Remarkably, a minority of novel chimeric gene families are broadly distributed across Halobacteria (cluster 2 in Fig. 2, 23 families). Genes within these families also show a larger divergence in primary sequences (measured in %identity between homologous sequences). Taken together, their broad distribution and the accumulation of substitution in their sequences suggest that these genes are ancient, and were possibly invented during Halogenesis. Focusing on their functions may help us to understand how the first halobacteria tackled the challenges of adaptation to an aerobic and salty environment. Many of these families seem involved in adhesion and/or are localized in the periplasm or the membrane. While many of these genes were involved in sugar metabolism (cog category G in Table S1) and many contain carbohydrate binding module (families 17758, 22083, 25805, 3843, 8114 and 8367), there is also a clear signal for more specific contribution to a change in lifestyle.

Chimeric genes invented during halogenesis are involved in redox activities (XXX number of genes), consistent with the need for early Halobacteria to thrive in an oxigenic environment (Table S1). XX chimeric gene families invented during halogenesis were also (XX number) involved in protein folding, which is likely important to deal with high salt concentration environment (Table S1).

Long term presence of chimeric and S-genes in Halobacteria. As indicated above, adaptation to a high salt level environment requires decreasing isolectric point of proteins. In order to assess the long term presence of chimeric genes in genomes, we calculated their isoelectric points. Results showed that isolectric points of chimeric genes and of S-genes do not differ from that of the rest of the halobacterial proteins, and are significantly lower than other archaeal and bacterial proteins (Fig. S3). These lower isoelectric points in chimeric genes are likely the result of a process of genetic optimization of their genetic fragments. Consistently, there is a significant difference in isoeletric points between the top five bacterial sequences matching with the bacterial components of S-genes and these bacterial components (Fig. S4). These calculations suggest that these chimeric proteins have experienced amino acid changes in order to adapt in their new environment, confirming their significant time of residency in halophiles.

Conclusion

Our network analyses show that over 320 novel chimeric genes evolved in Halobacteria. At least 23 such gene families appeared early in the evolution of Halobacteria, possibly during Halogenesis and were largely conserved since that time. 297 additional chimeric gene families appeared later, in already diverged Halobacterial groups. Importantly, 126 of these novel chimeric genes derived from genetic material from bacterial genomes. These halobacterial S-genes unravel a massive additional bacterial contribution to the evolution of Halobacteria, in

addition to the many reported cases of LGT. All these genes were more than transient visitors of a few halobacterial genomes: they were optimized to code for proteins with low isolectric points and are distributed in multiple related genomes, suggesting that, while not necessarily essential, composite and S-genes certainly play a role in the biology of Halobacteria. While these novel chimeric genes are largely involved in metabolic functions, many of them also played a role in adapting Halobacteria to an oxygenic lifestyle.

Materials and methods

Dataset creation

We assembled a protein sequence database by downloading every archaeal genome from to the NCBI Genome database in April 2016 (803 genomes, 1,816,486 proteins). 2,078 eubacteria genomes annotated as complete according to the NCBI Genome database (7,239,663 proteins) were downloaded with one representative per species according to the NCBI taxonomy.

Construction of gene families

The 1,816,486 archaeal protein sequences were compared pairwise using BLASTP (*10*) (version 2.2.26). Proteins were clustered into families using their sequence similarity (i.e., proteins that can be aligned over 80% of their length and showing protein identities over 30% are considered as homologs and so belong to the same family). The archaeal protein sequences were compared against the 7,239,663 eubacterial protein sequences. Families with only halobacteria proteins from at least 3 distinct genomes and no global similarities with eubacterial sequences (mutual cover > 80%, pid >25%) were kept for S-gene detection.

Detection and origin assignment of component families

For each retained sequences, component sequences were clustered into component families according to the following rule: if two component sequences overlapped by more than 70% of their lengths on the protein composite, they belonged to the same component family. A refining procedure has been done in order to merge overlapping and/or nested components

families, two component families were merged if one family is included by more than 70% of its length into the other one.

Component families were assigned a broad phylogenetic origin corresponding to their taxonomic composition. If the five best prokaryotic component sequences, according to their BLASTP bitscore against the composite gene, matched with the same prokaryotic domain (e.g., Archaea or Bacteria), we considered the component to have more specifically originated from that prokaryotic domain. If component family contained less than five sequences or if archaeal and bacterial sequences were both present among the five best sequences, we considered the component to originate from prokaryotes.

Domain and functional annotations

Domains were predicted using the conserved domain database (CDD) (version 3.13) (*11*) (default parameters). Sequences were functionally annotated with the halobacteria profiles dataset from the EggNog database (version 4.5) (*12*) (default parameters). For each family, if more than 60% of gene members share the same EggNog annotation, this EggNog annotation has been assigned to the family unless an unknown annotation has been assigned. Cellular localization has been detected using the PSORTdb (version 3.0) (*13*) (default parameters). For each family, the more abundant localization annotation has been used as family localization.

References

1. P. López-García, Y. Zivanovic, P. Deschamps, D. Moreira, Bacterial gene import and mesophilic adaptation in archaea. *Nat. Rev. Microbiol.* **13**, 447–56 (2015).

2. S. Nelson-Sathi *et al.*, Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 20537–42 (2012).

3. S. Nelson-Sathi *et al.*, Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature*. **517**, 77–80 (2015).

4. M. Groussin *et al.*, Gene Acquisitions from Bacteria at the Origins of Major Archaeal Clades Are Vastly Overestimated. *Mol. Biol. Evol.* **33**, 305–10 (2016).

5. E. A. Becker *et al.*, Phylogenetically driven sequencing of extremely halophilic archaea reveals strategies for static and dynamic osmo-response. *PLoS Genet.* **10**, e1004784 (2014).

6. S. Pasek, J.-L. Risler, P. Brézellec, Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics*. **22**, 1418–23 (2006).

7. R. Méheust, E. Zelzion, D. Bhattacharya, P. Lopez, E. Bapteste, Protein networks identify novel symbiogenetic genes resulting from plastid endosymbiosis. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 3579–84 (2016).

8. P.-A. Jachiet, R. Pogorelcnik, A. Berry, P. Lopez, E. Bapteste, MosaicFinder: identification of fused gene families in sequence similarity networks. *Bioinformatics*. **29**, 837–844 (2013).

9. R. Méheust, P. Lopez, E. Bapteste, Metabolic bacterial genes and the construction of

high-level composite lineages of life. *Trends Ecol. Evol.* **30**, 127–9 (2015).

10. S. F. Altschul *et al.*, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–402 (1997).

11. a. Marchler-Bauer *et al.*, CDD: NCBI's conserved domain database. *Nucleic Acids Res.* **43**, D222–D226 (2014).

12. J. Huerta-Cepas *et al.*, eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44**, D286-93 (2015).

13. M. A. Peabody, M. R. Laird, C. Vlasschaert, R. Lo, F. S. L. Brinkman, PSORTdb: expanding the bacteria and archaea protein subcellular localization database to better reflect diversity in cell envelope structures. *Nucleic Acids Res.* **44**, D663-8 (2016).

Figure legends

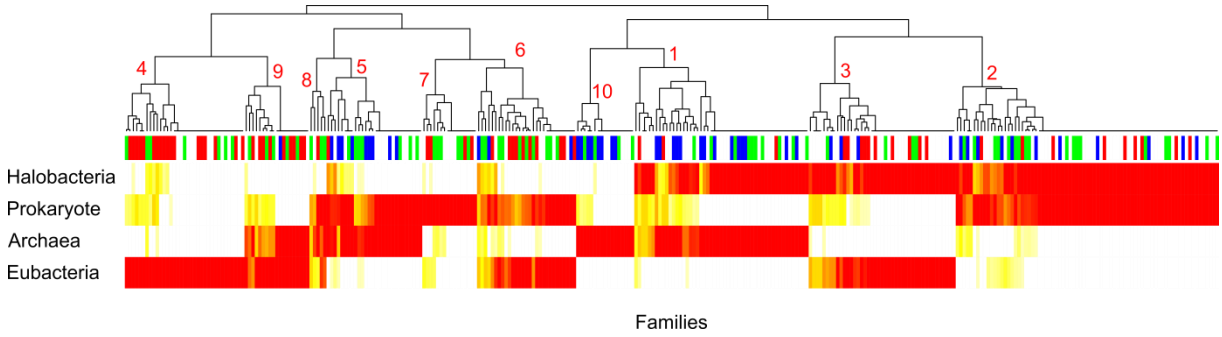


Figure S1. Hierarchical clustering of S-genes families according to their component origins. The heatmap represents the ratio of genes in a given family (columns) which have at least one component of a given origin (halobacteria, archaeal, bacterial or prokaryotic, rows). White lines correspond to the absence of component from a given origin in every gene in the given S-gene family. Color lines correspond to presence of at least one component of the given origin in a given percentage of genes in the given S-gene family (red lines mean that all (100%) genes contains a given origin component). A hierarchical clustering has been performed on families. The colored top bar indicates the functional annotation of the S-gene families according to COG (red: metabolism, blue: information storage and processing, green: cellular processes and signaling, white: poorly characterized).

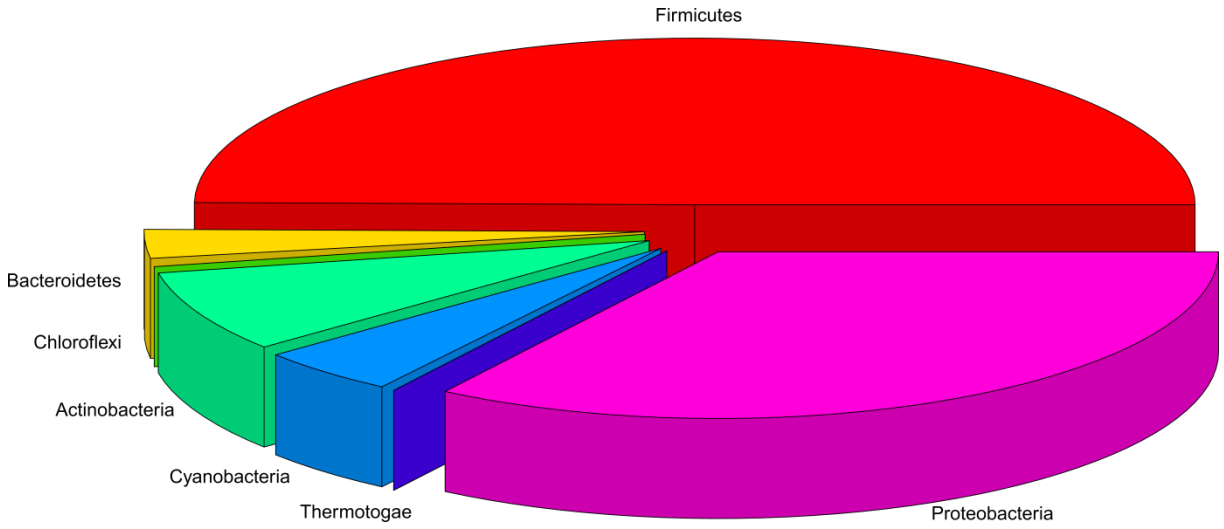


Figure S2. Pie chart of bacterial sources of S-gene families. For each bacterial component of S-genes, we looked at the phylum in which belong the top five hit sequences. We only considered bacterial components where the top five hit sequences belong to the same bacterial phylum.

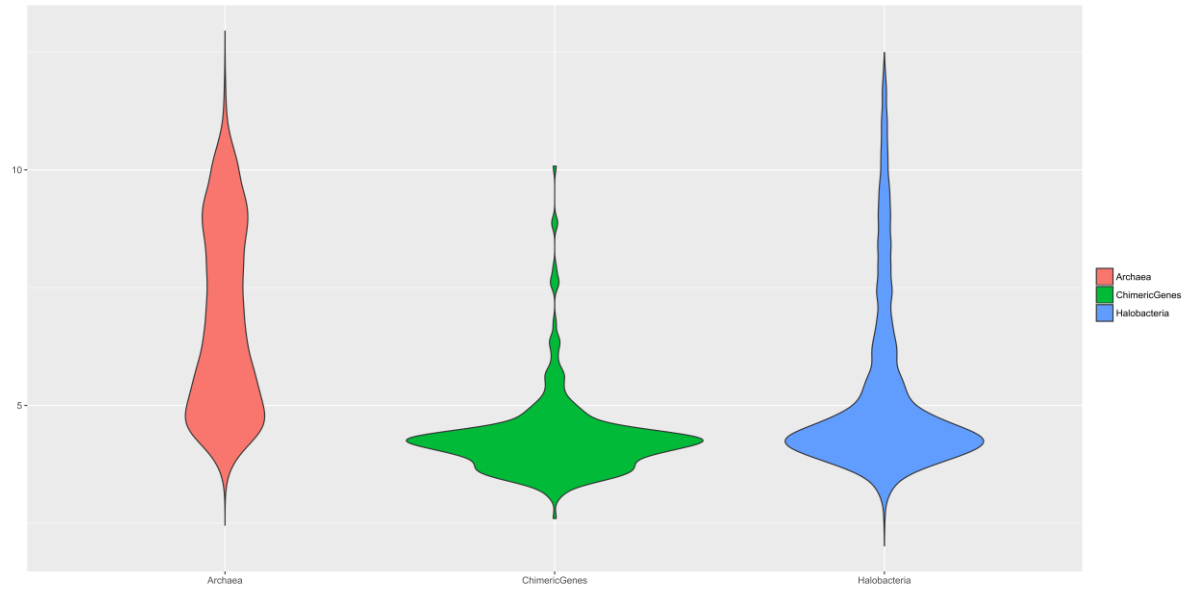


Figure S3. Violin plots of the distribution of isoelectric points (y-axis) of full-length proteins.

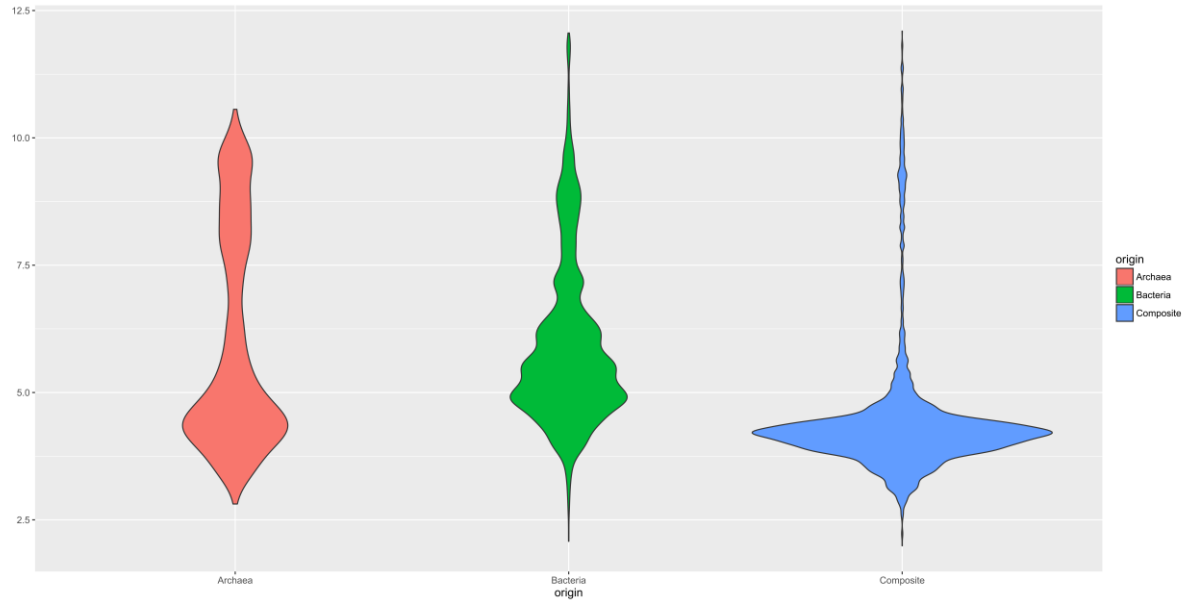


Figure S4. Violin plots of the distribution of isoelectric points (y-axis) of components (x-axis).

Table S1. Annotation of the 320 S-gene families detected.

Table S2. Detailed origin of prokaryotic S-gene components. Fam: S-gene family, Cpt: component, Cluster: cluster number according to Fig. 2, Bacteria: number of hits from Bacteria, Archaea: number of hits from Archaea. For each S-gene component, the rank in the BLAST search and the taxonomic assignation of the 25 sequences with the best hits to that component were reported (Aci: *Acidobacteria*, Act: *Actinobacteria*, Aqu: *Aquificae*, Arm: *Armatimonadetes*, Bac: *Bacteroidetes*, Chl: *Chloroflexi*, Cre: *Crenarchaeota*, Cya: *Cyanobacteria*, Def: *Deferribacteres*, Dei: *Deinococcus-Thermus*, Eur: *Euryarchaeota*, Fir: *Firmicutes*, Fus: *Fusobacteria*, Gem: *Gemmatimonadetes*, Ign: *Ignavibacteriae*, Nit: *Nitrospirae*, Pla: *Planctomycetes*, Pro: *Proteobacteria*, Spi: *Spirochaetes*, Syn: *Synergistetes*, Ten: *Tenericutes*, Tha: *Thaumarchaeota*, The: *Thermotogae*, Ver: *Verrucomicrobia*, roo: Unknown). Red cells correspond to bacterial phyla while blue cells correspond to archaeal phyla. When only one S-gene component is described, the unrepresented S-gene components from the S-gene family are either exclusively found in halobacteria, or have diverged too much to be confidently assigned to a prokaryotic group.

V. Conclusion et perspectives

Au cours de cette thèse, je me suis principalement intéressé à l'impact des endosymbioses dans la création de nouveaux gènes chimériques constitués de fragments d'au moins un partenaire symbiotique. Ces nouveaux gènes ont été nommés gènes symbiogénétiques afin de les différencier des autres gènes chimériques.

De tels gènes ont été identifiés dans les génomes eucaryotes, les génomes d'eucaryotes photosynthétiques ainsi que dans les génomes d'halobactéries. Leur existence étend le concept de mosaïcisme au niveau infra-génique et réaffirme que l'évolution des gènes et des génomes n'est pas qu'un processus graduel et arborescent; au sein des génomes cohabitent des gènes d'origines distinctes acquis ou crées *via* différents mécanismes. Les S-gènes représentent une telle classe de gènes. Bien évidemment tout arrive en biologie, il s'agit ensuite de savoir si ces mécanismes ou processus sont significativement importants quantitativement et/ou qualitativement. Au regard des premiers résultats, la contribution des S-gènes apparait modeste en terme de nombre de gènes créés, du moins clairement plus modeste que celle des gènes issus de duplications mais sans doute plus importante que celle des gènes créés *de novo* chez les eucaryotes. Une des hypothèses énoncées au début de ma thèse était que les S-gènes conféraient des propriétés émergentes à la cellule qui les exprimait et pouvaient donc être à l'origine de nouveaux traits. Comme expliqué précédemment (voir chapitre III.B.2.d), il est compliqué d'identifier le pouvoir adaptatif d'un gène et les résultats obtenus ne nous permettent pas de répondre. Cependant, la distribution et les fonctions des S-gènes laisse penser qu'une partie peut être associée avec l'eucaryogenèse, l'acquisition de la photosynthèse chez les eucaryotes ainsi que l'émergence des halobactéries comme cela a pu être suggéré pour les duplications et l'eucaryogénèse [127].

L'étude des origines phylogénétiques des composantes des S-gènes a permis de distinguer plusieurs règles d'associations. Ces règles ne sont pas aléatoires. Par exemple dans le cas des eucaryotes, la part de S-gènes possédant des composantes provenant à la fois d'une archaea et d'une bactérie est minime malgré la coexistence au sein du même génome de gènes archaea et bactériens depuis plusieurs centaines de millions d'années. Les composantes ayant une même origine ont plus de chance d'être associées au sein d'un S-gène. De même, la proportion de S-gènes issus de l'association d'une composante d'un symbionte et d'une composante de l'hôte (par exemple un morceau bactérien et un morceau eucaryote) est importante. Ces derniers demandent une étude approfondie car la composante de l'hôte peut

correspondre à la composante d'un symbionte ayant tellement divergé qu'il n'est plus possible de l'assigner phylogénétiquement au symbionte. L'utilisation de profils HMM (Hidden Markov Model), plus sensibles que la recherche de similarité par BLAST, pourrait être une solution. De la même manière que pour les gènes complets, les fonctions des S-gènes sont corrélées avec les origines de leurs composantes.

Les trois études sur les S-gènes se sont particulièrement intéressées aux S-gènes anciens, c'est à dire formés lors des processus d'émergence des eucaryotes, des eucaryotes photosynthétiques et des halobactéries, afin de trouver une association entre ces processus et l'apparition de S-gènes. Or les trois études ont aussi suggéré que la création de S-gènes est continue et qu'une grande partie de ces gènes est spécifique des lignées plus récentes. Une suite naturelle de mes travaux serait donc d'étudier les familles spécifiques de lignées. Deux autres perspectives seraient d'étendre mes travaux à d'autres organismes issus de transitions égalitaires comme les autres groupes d'Archaea ou encore *Paulinella chromatophora* [201] et d'aller rechercher des S-gènes chez les organismes multicellulaires.

Ces trois années de travail nous ont permis de trouver des règles d'associations entre l'origine et les fonctions des S-gènes. Par contre, nous n'avons pas réussi à dégager de liens clairs pour comprendre ce qui conduit les gènes/les fragments de gènes à s'associer entre eux. Il existe de nombreuses études suggérant ou démontrant le pourquoi de ces associations mais aucune ne semble avoir examiné cette question d'une manière globale. Deux gènes codés à différents locus peuvent être transcrits puis traduits séparément en protéines pour ensuite interagir ensemble dans la cellule. Pourquoi ces gènes/fragments de gènes se retrouvent associés ? Au-delà des gènes composites, l'ordre des gènes n'est pas aléatoire dans les génomes eucaryotes [202]. Les gènes ayant les mêmes patrons d'expression ont plus tendance à co-localiser [202], de même que les gènes codant pour des protéines impliquées dans les mêmes voies métaboliques [203–207]. Cette question est d'autant plus intéressante chez les procaryotes où la transcription et la traduction sont couplées. Une étude récente a montré que l'organisation en opéron des gènes permet de guider plus efficacement l'assemblage d'un complexe protéique par rapport à des gènes non organisés en opéron [208]. Les gènes composites présentent le cas le plus extrême d'association puisque les composantes sont directement fusionnées entre elles. Sur ce dernier point, plusieurs études ont mis en évidence des règles d'association. Par exemple les gènes codant des protéines interagissant ensemble ont tendance à fusionner ensemble [183,184] et une étude de l'équipe de Sarah Teichmann a montré que l'assemblage des complexes protéiques contraint l'ordre des fusions [186].

D'autres études ont pointé du doigt la toxicité ou la fragilité de certains produits enzymatiques pour expliquer pourquoi certains gènes codant pour des enzymes se trouvent fusionnés entre eux [209–211]. Une idée qui a émergé des travaux sur les eucaryotes et les eucaryotes photosynthétiques est que la création de gènes chimériques pourrait être un bon moyen d'adresser dans le bon tempo plusieurs protéines dans le bon compartiment cellulaire. Dans les cellules eucaryotes qui possèdent de nombreux compartiments, les protéines travaillant dans un compartiment cellulaire doivent être importé du cytoplasme où elles sont traduites vers le bon compartiment cellulaire. Les fusionner peut être un bon moyen pour les adresser au bon endroit et dans le bon tempo. Approfondir cette hypothèse de co-localisation me semble être la perspective la plus intéressante de mon travail. Elle pourrait peut être permettre de comprendre pourquoi les génomes eucaryotes possèdent en moyenne plus de gènes composites que les génomes procaryotes.

VI. Références

1 Darwin, C. (1859) *On the Origin of the Species*,

2 Nutman, A.P. *et al.* (2016) Rapid emergence of life shown by discovery of 3,700-million-year-old microbial structures. *Nature* DOI: 10.1038/nature19355

3 Bapteste, E. *et al.* (2012) Evolutionary analyses of non-genealogical bonds produced by introgressive descent. *Proc. Natl. Acad. Sci. U. S. A.* 109, 18266–18272

4 Halary, S. *et al.* (2010) Network analyses structure genetic diversity in independent genetic worlds. *Proc. Natl. Acad. Sci.* 107, 127–132

5 Corel, E. *et al.* (2016) Network-Thinking: Graphs to Analyze Microbial Complexity and Evolution. *Trends Microbiol.* 24, 224–37

6 Dagan, T. *et al.* (2008) Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc. Natl. Acad. Sci. U. S. A.* 105, 10039–44

7 McFall-Ngai, M. *et al.* (2013) Animals in a bacterial world, a new imperative for the life sciences. *Proc. Natl. Acad. Sci. U. S. A.* 110, 3229–36

8 Brito, I.L. *et al.* (2016) Mobile genes in the human microbiome are structured from global to individual scales. *Nature* 535, 435–439

9 Pedersen, H.K. *et al.* (2016) Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature* 535, 376–381

10 Guidi, L. *et al.* (2016) Plankton networks driving carbon export in the oligotrophic ocean. *Nature* 532, 465–70

11 Moran, N.A. and Sloan, D.B. (2015) The Hologenome Concept: Helpful or Hollow? *PLOS Biol.* 13, e1002311

12 Nelson-Sathi, S. *et al.* (2015) Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* 517, 77–80

13 Ereshefsky, M. and Pedroso, M. (2015) Rethinking evolutionary individuality. *Proc. Natl. Acad. Sci. U. S. A.* 112, 10126–32

14 Lima-Mendez, G. *et al.* (2015) Ocean plankton. Determinants of community structure in the global plankton interactome. *Science* 348, 1262073

15 Gilbert, S.F. *et al.* (2015) Eco-Evo-Devo: developmental symbiosis and developmental plasticity as evolutionary agents. *Nat. Rev. Genet.* 16, 611–622

16 Liu, J. *et al.* (2015) Metabolic co-dependence gives rise to collective oscillations within biofilms. *Nature* DOI: 10.1038/nature14660

17 Bailleul, B. *et al.* (2015) Energetic coupling between plastids and mitochondria drives CO2 assimilation in diatoms. *Nature* DOI: 10.1038/nature14599

18 Overmann, J. (2010) The phototrophic consortium “Chlorochromatium aggregatum” - a model for bacterial heterologous multicellularity. *Adv. Exp. Med. Biol.* 675, 15–29

19 Amin, S.A. *et al.* (2015) Interaction and signalling between a cosmopolitan phytoplankton and associated bacteria. *Nature* 522, 98–101

20 Kiers, E.T. and West, S.A. (2015) Evolving new organisms via symbiosis. *Science (80-.).* 348, 392–394

21 Blouin, N.A. and Lane, C.E. (2012) Red algal parasites: models for a life history evolution that leaves photosynthesis behind again and again. *Bioessays* 34, 226–35

22 Cordero, O.X. *et al.* (2012) Public good dynamics drive evolution of iron acquisition strategies in natural bacterioplankton populations. *Proc. Natl. Acad. Sci. U. S. A.* 109, 20059–64

23 Bobay, L.-M. *et al.* (2013) The adaptation of temperate bacteriophages to their host genomes. *Mol. Biol. Evol.* 30, 737–51

24 Nakayama, T. *et al.* (2014) Complete genome of a nonphotosynthetic cyanobacterium in a diatom reveals recent adaptations to an intracellular lifestyle. *Proc. Natl. Acad. Sci.*

111, 11407–12
25 Levin, S.A. (2014) Public goods in relation to competition, cooperation, and spite. *Proc. Natl. Acad. Sci.* 111 Suppl, 10838–45
26 Smillie, C.S. *et al.* (2011) Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480, 241–4
27 Yoon, H.S. *et al.* (2011) Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science* 332, 714–7
28 Werner, G.D.A. *et al.* (2014) Evolution of microbial markets. *Proc. Natl. Acad. Sci. U. S. A.* 111, 1237–44
29 Husnik, F. *et al.* (2013) Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell* 153, 1567–78
30 Molloy, S. (2013) Symbiosis: a symbiotic mosaic. *Nat. Rev. Microbiol.* 11, 510
31 Douglas, A.E. (2014) Symbiosis as a general principle in eukaryotic evolution. *Cold Spring Harb. Perspect. Biol.* 6,
32 Decelle, J. *et al.* (2012) An original mode of symbiosis in open ocean plankton. *Proc. Natl. Acad. Sci. U. S. A.* 109, 18000–5
33 Brucker, R.M. and Bordenstein, S.R. (2013) The hologenomic basis of speciation: gut bacteria cause hybrid lethality in the genus Nasonia. *Science* 341, 667–9
34 Modi, S.R. *et al.* (2013) Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature* 499, 219–22
35 Kåhrström, C.T. (2013) Metagenomics: With a little help from my phage friends. *Nat. Rev. Genet.* 14, 517
36 Moissl-Eichinger, C. and Huber, H. (2011) Archaeal symbionts and parasites. *Curr. Opin. Microbiol.* 14, 364–70
37 Wrede, C. *et al.* (2012) Archaea in symbioses. *Archaea* 2012, 596846
38 Thompson, A.W. *et al.* (2012) Unicellular cyanobacterium symbiotic with a single-celled eukaryotic alga. *Science* 337, 1546–50
39 Keeling, P.J. *et al.* (2015) Symbiosis becoming permanent: Survival of the luckiest. *Proc. Natl. Acad. Sci.* 112, 10101–10103
40 Szathmáry, E. and Smith, J.M. (1995) The major evolutionary transitions. *Nature* 374, 227–32
41 Szathmáry, E. (2015) Toward major evolutionary transitions theory 2.0. *Proc. Natl. Acad. Sci. U. S. A.* DOI: 10.1073/pnas.1421398112
42 Gray, M.W. and Doolittle, W.F. (1982) Has the endosymbiont hypothesis been proven? *Microbiol. Rev.* 46, 1–42
43 Sapp, J. (1998) *Evolution by association: A history of symbiosis*, 29
44 Archibald, J.M. (2015) Endosymbiosis and Eukaryotic Cell Evolution. *Curr. Biol.* 25, R911–R921
45 Koonin, E. V. (2015) Archaeal ancestors of eukaryotes: not so elusive any more. *BMC Biol.* 13, 84
46 Koonin, E. V (2015) Origin of eukaryotes from within archaea, archaeal eukaryome and bursts of gene gain: eukaryogenesis just made easier? *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 370,
47 Martin, W.F. *et al.* (2015) Endosymbiotic theories for eukaryote origin. *Philos. Trans. R. Soc. B Biol. Sci.* 370, 20140330
48 Lake, J.A. (2015) Eukaryotic origins. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 370, 20140321
49 Koonin, E. V (2010) The origin and early evolution of eukaryotes in the light of phylogenomics. *Genome Biol.* 11, 209
50 Ettema, T.J.G. (2016) Evolution: Mitochondria in the second act. *Nature* 531, 9–10

51 Embley, T.M. and Martin, W. (2006) Eukaryotic evolution, changes and challenges. *Nature* 440, 623–30
52 Karnkowska, A. *et al.* (2016) A Eukaryote without a Mitochondrial Organelle. *Curr. Biol.* DOI: 10.1016/j.cub.2016.03.053
53 Williams, T.A. *et al.* (2013) An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* 504, 231–236
54 Woese, C.R. and Fox, G.E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. U. S. A.* 74, 5088–5090
55 Lake, J.A. *et al.* (1984) Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 81, 3786–3790
56 Guy, L. and Ettema, T.J.G. (2011) The archaeal "TACK" superphylum and the origin of eukaryotes. *Trends Microbiol.* 19, 580–7
57 Spang, A. *et al.* (2015) Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* 521, 173–179
58 Archibald, J.M. (2015) Evolution: Gene transfer in complex cells. *Nature* advance on,
59 McInerney, J.O. *et al.* (2014) The hybrid nature of the Eukaryota and a consilient view of life on Earth. *Nat. Rev. Microbiol.* 12, 449–55
60 Lane, N. and Martin, W. (2010) The energetics of genome complexity. *Nature* 467, 929–34
61 Lynch, M. and Marinov, G.K. (2015) The bioenergetic costs of a gene. *Proc. Natl. Acad. Sci.* DOI: 10.1073/pnas.1514974112
62 Lynch, M. and Conery, J.S. (2003) The origins of genome complexity. *Science* 302, 1401–4
63 Lynch, M. and Marinov, G.K. (2016) Reply to Lane and Martin: Mitochondria do not boost the bioenergetic capacity of eukaryotic cells. *Proc. Natl. Acad. Sci. U. S. A.* 113, E667-8
64 Lane, N. and Martin, W.F. (2016) Mitochondria, complexity, and evolutionary deficit spending. *Proc. Natl. Acad. Sci. U. S. A.* 113, E666
65 Koonin, E. V (2015) Energetics and population genetics at the root of eukaryotic cellular and genomic complexity. *Proc. Natl. Acad. Sci. U. S. A.* 112, 15777–8
66 Pittis, A.A. and Gabaldón, T. (2016) Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. *Nature* 531, 101–4
67 O'Malley, M.A. (2010) The first eukaryote cell: an unfinished history of contestation. *Stud. Hist. Philos. Biol. Biomed. Sci.* 41, 212–24
68 Timmis, J.N. *et al.* (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat. Rev. Genet.* 5, 123–35
69 Rivera, M.C. *et al.* (1998) Genomic evidence for two functionally distinct gene classes. *Proc. Natl. Acad. Sci. U. S. A.* 95, 6239–44
70 Ku, C. *et al.* (2015) Endosymbiotic origin and differential loss of eukaryotic genes. *Nature* 524, 427–437
71 Lane, N. and Martin, W.F. (2015) Eukaryotes really are special, and mitochondria are why. *Proc. Natl. Acad. Sci. U. S. A.* 112, E4823
72 Dagan, T. and Martin, W. (2006) The tree of one percent. *Genome Biol.* 7, 118
73 Archibald, J.M. (2015) Genomic perspectives on the birth and spread of plastids. *Proc. Natl. Acad. Sci. U. S. A.* DOI: 10.1073/pnas.1421374112
74 Keeling, P.J. (2013) The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. *Annu. Rev. Plant Biol.* 64, 583–607
75 Rodríguez-Ezpeleta, N. *et al.* (2005) Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. *Curr. Biol.* 15, 1325–30
76 McFadden, G.I. and van Dooren, G.G. (2004) Evolution: red algal genome affirms a

common origin of all plastids. *Curr. Biol.* 14, R514-6
77 Martin, W. *et al.* (2002) Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc. Natl. Acad. Sci. U. S. A.* 99, 12246–51
78 Price, D. *et al.* (2012) Cyanophora paradoxa genome elucidates origin of photosynthesis in algae and plants. *Science (80-.).* 335, 843–847
79 Matsuzaki, M. *et al.* (2004) Genome sequence of the ultrasmall unicellular red alga Cyanidioschyzon merolae 10D. *Nature* 428, 653–7
80 Reyes-Prieto, A. *et al.* (2006) Cyanobacterial contribution to algal nuclear genomes is primarily limited to plastid functions. *Curr. Biol.* 16, 2320–5
81 Curtis, B.A. *et al.* (2012) Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature* 492, 59–65
82 Rogers, M.B. *et al.* (2007) The complete chloroplast genome of the chlorarachniophyte Bigelowiella natans: evidence for independent origins of chlorarachniophyte and euglenid secondary endosymbionts. *Mol. Biol. Evol.* 24, 54–62
83 Hirakawa, Y. *et al.* (2009) Protein targeting into secondary plastids of chlorarachniophytes. *Proc. Natl. Acad. Sci. U. S. A.* 106, 12820–5
84 Zimorski, V. *et al.* (2014) Endosymbiotic theory for organelle origins. *Curr. Opin. Microbiol.* 22, 38–48
85 Janouskovec, J. *et al.* (2010) A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. *Proc. Natl. Acad. Sci. U. S. A.* 107, 10949–54
86 Yoon, H.S. *et al.* (2002) The single, ancient origin of chromist plastids. *Proc. Natl. Acad. Sci. U. S. A.* 99, 15507–12
87 Burki, F. *et al.* (2012) The evolutionary history of haptophytes and cryptophytes: phylogenomic evidence for separate origins. *Proc. Biol. Sci.* 279, 2246–2254
88 Baurain, D. *et al.* (2010) Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles. *Mol. Biol. Evol.* 27, 1698–709
89 Shi, L.-X. and Theg, S.M. (2013) The chloroplast protein import system: from algae to trees. *Biochim. Biophys. Acta* 1833, 314–31
90 Kessler, F. and Schnell, D. (2009) Chloroplast biogenesis: diversity and regulation of the protein import apparatus. *Curr. Opin. Cell Biol.* 21, 494–500
91 Gould, S.B. *et al.* (2015) Protein Import and the Origin of Red Complex Plastids. *Curr. Biol.* 25, R515–R521
92 Dolezal, P. *et al.* (2006) Evolution of the molecular machines for protein import into mitochondria. *Science* 313, 314–8
93 Sommer, M.S. *et al.* (2007) Der1-mediated preprotein import into the periplastid compartment of chromalveolates? *Mol. Biol. Evol.* 24, 918–28
94 Stiller, J.W. *et al.* (2014) The evolution of photosynthesis in chromist algae through serial endosymbioses. *Nat. Commun.* 5, 5764
95 O'Malley, M.A. (2015) Endosymbiosis and its implications for evolutionary theory. *Proc. Natl. Acad. Sci. U. S. A.* DOI: 10.1073/pnas.1421389112
96 Becker, E.A. *et al.* (2014) Phylogenetically driven sequencing of extremely halophilic archaea reveals strategies for static and dynamic osmo-response. *PLoS Genet.* 10, e1004784
97 Groussin, M. *et al.* (2016) Gene Acquisitions from Bacteria at the Origins of Major Archaeal Clades Are Vastly Overestimated. *Mol. Biol. Evol.* 33, 305–10
98 Nelson-Sathi, S. *et al.* (2012) Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proc. Natl. Acad. Sci. U. S. A.* 109, 20537–42
99 Nowack, E.C.M. *et al.* (2008) Chromatophore genome sequence of Paulinella sheds

light on acquisition of photosynthesis by eukaryotes. *Curr. Biol.* 18, 410–8
100 Szamecz, B. *et al.* (2014) The genomic landscape of compensatory evolution. *PLoS Biol.* 12, e1001935
101 Wolf, Y.I. and Koonin, E. V (2013) Genome reduction as the dominant mode of evolution. *Bioessays* 35, 829–37
102 O'Malley, M.A. *et al.* (2016) Losing Complexity: The Role of Simplification in Macroevolution. *Trends Ecol. Evol.* 31, 608–621
103 Kaessmann, H. (2010) Origins, evolution, and phenotypic impact of new genes. *Genome Res.* 20, 1313–26
104 Lord, E. *et al.* (2016) BRIDES: A New Fast Algorithm and Software for Characterizing Evolving Similarity Networks Using Breakthroughs, Roadblocks, Impasses, Detours, Equals and Shortcuts. *PLoS One* 11, e0161474
105 Soucy, S.M. *et al.* (2015) Horizontal gene transfer: building the web of life. *Nat. Rev. Genet.* 16, 472–482
106 Jain, R. *et al.* (2003) Horizontal gene transfer accelerates genome innovation and evolution. *Mol. Biol. Evol.* 20, 1598–602
107 Ochman, H. *et al.* (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* 405, 299–304
108 Treangen, T.J. and Rocha, E.P.C. (2011) Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet.* 7, e1001284
109 Bapteste, E. *et al.* (2009) Prokaryotic evolution and the tree of life are two different things. *Biol. Direct* 4, 34
110 Doolittle, W.F. and Bapteste, E. (2007) Pattern pluralism and the Tree of Life hypothesis. *Proc. Natl. Acad. Sci. U. S. A.* 104, 2043–2049
111 Lukjancenko, O. *et al.* Comparison of 61 Sequenced Escherichia coli Genomes. , *Microbial Ecology*, 60. (2010) , 708–720
112 Tettelin, H. *et al.* (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *PNAS* 102, 13950–13955
113 Andam, C.P. and Gogarten, J.P. (2011) Biased gene transfer in microbial evolution. *Nat. Rev. Microbiol.* 9, 543–55
114 Frost, L.S. *et al.* (2005) Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.* 3, 722–32
115 Beiko, R.G. *et al.* (2005) Highways of gene sharing in prokaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 102, 14332–7
116 Ragan, M.A. (2001) Detection of lateral gene transfer among microbial genomes. *Curr. Opin. Genet. Dev.* 11, 620–6
117 Jain, R. *et al.* (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl. Acad. Sci. U. S. A.* 96, 3801–6
118 Cohen, O. *et al.* (2011) The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. *Mol. Biol. Evol.* 28, 1481–9
119 Dimitriu, T. *et al.* (2014) Genetic information transfer promotes cooperation in bacteria. *Proc. Natl. Acad. Sci.* DOI: 10.1073/pnas.1406840111
120 Sullivan, M.B. *et al.* (2006) Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biol.* 4, e234
121 McInerney, J.O. *et al.* (2011) The Public Goods Hypothesis for the evolution of life on Earth. *Biol. Direct* 6, 41
122 Keeling, P.J. and Palmer, J.D. (2008) Horizontal gene transfer in eukaryotic evolution. *Nat. Rev. Genet.* 9, 605–18
123 Rosen, M.J. *et al.* (2015) Microbial diversity. Fine-scale diversity and extensive

recombination in a quasisexual bacterial population occupying a broad niche. *Science* 348, 1019–23
124 Worden, A.Z. *et al.* (2009) Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas. *Science* 324, 268–72
125 Read, B.A. *et al.* (2013) Pan genome of the phytoplankton Emiliania underpins its global distribution. *Nature* advance on,
126 Mira, A. *et al.* (2010) The bacterial pan-genome:a new paradigm in microbiology. *Int. Microbiol.* 13, 45–57
127 Makarova, K.S. *et al.* (2005) Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. *Nucleic Acids Res.* 33, 4626–38
128 Alvarez-Ponce, D. *et al.* (2013) Gene similarity networks provide tools for understanding eukaryote origins and evolution. *Proc. Natl. Acad. Sci. U. S. A.* 110, E1594-603
129 Ku, C. *et al.* (2015) Endosymbiotic gene transfer from prokaryotic pangenomes: Inherited chimerism in eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* DOI: 10.1073/pnas.1421385112
130 Kondo, N. *et al.* (2002) Genome fragment of Wolbachia endosymbiont transferred to X chromosome of host insect. *Proc. Natl. Acad. Sci. U. S. A.* 99, 14280–5
131 Mi, S. *et al.* (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* 403, 785–9
132 Alsmark, C. *et al.* (2013) Patterns of prokaryotic lateral gene transfers affecting parasitic microbial eukaryotes. *Genome Biol.* 14, R19
133 Hirt, R.P. *et al.* Lateral gene transfers and the origins of the eukaryote proteome: A view from microbial parasites. , *Current Opinion in Microbiology*, 23. (2015) , Elsevier Ltd, 155–162
134 Katz, L.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. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 370, 20140324
135 Szöllősi, G.J. *et al.* (2015) Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 370, 20140335
136 Savory, F. *et al.* The Role of Horizontal Gene Transfer in the Evolution of the Oomycetes. , *PLoS Pathogens*, 11. May-(2015) , e1004805
137 Morris, P.F. *et al.* (2009) Multiple horizontal gene transfer events and domain fusions have created novel regulatory and metabolic networks in the oomycete genome. *PLoS One* 4, e6133
138 Richards, T.A. and Talbot, N.J. (2013) Horizontal gene transfer in osmotrophs: playing with public goods. *Nat. Rev. Microbiol.* 11, 720–7
139 Richards, T.A. *et al.* (2011) Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes. *Proc. Natl. Acad. Sci. U. S. A.* 108, 15258–63
140 Yue, J. *et al.* (2012) Widespread impact of horizontal gene transfer on plant colonization of land. *Nat. Commun.* 3, 1152
141 Ball, S.G. *et al.* (2013) Metabolic effectors secreted by bacterial pathogens: essential facilitators of plastid endosymbiosis? *Plant Cell* 25, 7–21
142 Moustafa, A. *et al.* (2008) Chlamydiae has contributed at least 55 genes to Plantae with predominantly plastid functions. *PLoS One* 3, e2205
143 Huang, J. and Gogarten, J.P. (2007) Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? *Genome Biol.* 8, R99
144 Brinkman, F.S.L. *et al.* (2002) Evidence that plant-like genes in Chlamydia species reflect an ancestral relationship between Chlamydiaceae, cyanobacteria, and the

chloroplast. *Genome Res.* 12, 1159–67
145 Greub, G. and Raoult, D. (2003) History of the ADP/ATP-translocase-encoding gene, a parasitism gene transferred from a Chlamydiales ancestor to plants 1 billion years ago. *Appl. Environ. Microbiol.* 69, 5530–5
146 Domman, D. *et al.* (2015) Plastid establishment did not require a chlamydial partner. *Nat. Commun.* 6, 6421
147 Ball, S.G. *et al.* (2016) Commentary: Plastid establishment did not require a chlamydial partner. *Front. Cell. Infect. Microbiol.* 6, 43
148 Esser, C. *et al.* (2004) A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Mol. Biol. Evol.* 21, 1643–60
149 Cotton, J.A. and McInerney, J.O. (2010) Eukaryotic genes of archaebacterial origin are more important than the more numerous eubacterial genes, irrespective of function. *Proc. Natl. Acad. Sci. U. S. A.* 107, 17252–5
150 Ohno, S. (1970) Evolution by Gene Duplication. *(1970)* at <papers2://publication/uuid/B88A8170-7082-4E75-83AE-AEBEAEE93729>
151 Zhao, L. *et al.* (2014) Origin and spread of de novo genes in Drosophila melanogaster populations. *Science* 343, 769–72
152 Gilbert, W. (1978) Why genes in pieces? *Nature* 271, 501
153 Chen, S. *et al.* (2013) New genes as drivers of phenotypic evolution. *Nat. Rev. Genet.* 14, 645–60
154 Courseaux, A. and Nahon, J.L. (2001) Birth of two chimeric genes in the Hominidae lineage. *Science* 291, 1293–7
155 Salim, H.M.W. *et al.* (2011) Detection of fused genes in eukaryotic genomes using gene deFuser: analysis of the Tetrahymena thermophila genome. *BMC Bioinformatics* 12, 279
156 Vogel, C. *et al.* (2004) Structure, function and evolution of multidomain proteins. *Curr. Opin. Struct. Biol.* 14, 208–16
157 Jachiet, P.-A. *et al.* (2014) Extensive gene remodeling in the viral world: new evidence for nongradual evolution in the mobilome network. *Genome Biol. Evol.* 6, 2195–205
158 Jachiet, P.-A. *et al.* (2013) MosaicFinder: identification of fused gene families in sequence similarity networks. *Bioinformatics* 29, 837–844
159 Liu, M. and Grigoriev, A. (2004) Protein domains correlate strongly with exons in multiple eukaryotic genomes--evidence of exon shuffling? *Trends Genet.* 20, 399–403
160 Long, M. (2000) A new function evolved from gene fusion. *Genome Res.* 10, 1655–7
161 Andreatta, M.E. *et al.* (2015) The Recent De Novo Origin of Protein C-Termini. *Genome Biol. Evol.* 7, 1686–701
162 Nekrutenko, A. and Li, W.H. Transposable elements are found in a large number of human protein-coding genes. , *Trends in Genetics*, 17. (2001) , 619–621
163 Wang, W. *et al.* (2006) High rate of chimeric gene origination by retroposition in plant genomes. *Plant Cell* 18, 1791–802
164 Zhang, P.G. *et al.* (2010) Extensive divergence in alternative splicing patterns after gene and genome duplication during the evolutionary history of Arabidopsis. *Mol. Biol. Evol.* 27, 1686–97
165 Jacob, F. (1977) Evolution and tinkering. *Science (80-.).* 196, 1161–1166
166 Long, M. and Langley, C.H. (1993) Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. *Science* 260, 91–5
167 Long, M. *et al.* (2003) The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.* 4, 865–75
168 Wang, W. *et al.* (2000) The origin of the Jingwei gene and the complex modular

structure of its parental gene, yellow emperor, in Drosophila melanogaster. *Mol. Biol. Evol.* 17, 1294–301
169 Zhang, J. *et al.* (2004) Evolving protein functional diversity in new genes of Drosophila. *Proc. Natl. Acad. Sci. U. S. A.* 101, 16246–50
170 Jones, C.D. and Begun, D.J. (2005) Parallel evolution of chimeric fusion genes. *Proc. Natl. Acad. Sci. U. S. A.* 102, 11373–8
171 Wilson, S.J. *et al.* (2008) Independent evolution of an antiviral TRIMCyp in rhesus macaques. *Proc. Natl. Acad. Sci. U. S. A.* 105, 3557–62
172 Rogers, R.L. *et al.* (2010) Adaptive impact of the chimeric gene Quetzalcoatl in Drosophila melanogaster. *Proc. Natl. Acad. Sci. U. S. A.* 107, 10943–8
173 Rogers, R.L. and Hartl, D.L. (2012) Chimeric genes as a source of rapid evolution in Drosophila melanogaster. *Mol. Biol. Evol.* 29, 517–29
174 Wang, W. *et al.* (2002) Origin of sphinx, a young chimeric RNA gene in Drosophila melanogaster. *Proc. Natl. Acad. Sci. U. S. A.* 99, 4448–53
175 Kawai, H. *et al.* (2003) Responses of ferns to red light are mediated by an unconventional photoreceptor. *Nature* 421, 287–90
176 Li, F.-W. *et al.* (2014) Horizontal transfer of an adaptive chimeric photoreceptor from bryophytes to ferns. *Proc. Natl. Acad. Sci. U. S. A.* 111, 6672–7
177 Suetsugu, N. *et al.* (2005) A chimeric photoreceptor gene, NEOCHROME, has arisen twice during plant evolution. *Proc. Natl. Acad. Sci. U. S. A.* 102, 13705–9
178 Thomson, T.M. *et al.* (2000) Fusion of the human gene for the polyubiquitination coeffector UEV1 with Kua, a newly identified gene. *Genome Res.* 10, 1743–56
179 Brennan, G. *et al.* (2008) TRIMCyp expression in Old World primates Macaca nemestrina and Macaca fascicularis. *Proc. Natl. Acad. Sci. U. S. A.* 105, 3569–74
180 Salim, H.M.W. *et al.* (2009) 1+1 = 3: a fusion of 2 enzymes in the methionine salvage pathway of Tetrahymena thermophila creates a trifunctional enzyme that catalyzes 3 steps in the pathway. *PLoS Genet.* 5, e1000701
181 Avelar, G.M. *et al.* (2014) A Rhodopsin-Guanylyl cyclase gene fusion functions in visual perception in a fungus. *Curr. Biol.* 24, 1234–1240
182 Yanai, I. *et al.* (2001) Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. *Proc. Natl. Acad. Sci. U. S. A.* 98, 7940–5
183 Enright, A.J. *et al.* (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402, 86–90
184 Marcotte, E.M. *et al.* (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science* 285, 751–3
185 Enright, A.J. and Ouzounis, C.A. (2001) Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. *Genome Biol.* 2, RESEARCH0034
186 Marsh, J.A. *et al.* (2013) Protein complexes are under evolutionary selection to assemble via ordered pathways. *Cell* 153, 461–70
187 Tsoka, S. and Ouzounis, C.A. (2000) Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. *Nat. Genet.* 26, 141–2
188 von Mering, C. *et al.* (2003) Genome evolution reveals biochemical networks and functional modules. *Proc. Natl. Acad. Sci. U. S. A.* 100, 15428–33
189 Richards, T.A. and Gomes, S.L. (2015) Protistology: How to build a microbial eye. *Nature* advance on,
190 Gavelis, G.S. *et al.* (2015) Eye-like ocelloids are built from different endosymbiotically acquired components. *Nature* advance on,
191 Frommolt, R. *et al.* (2008) Ancient recruitment by chromists of green algal genes

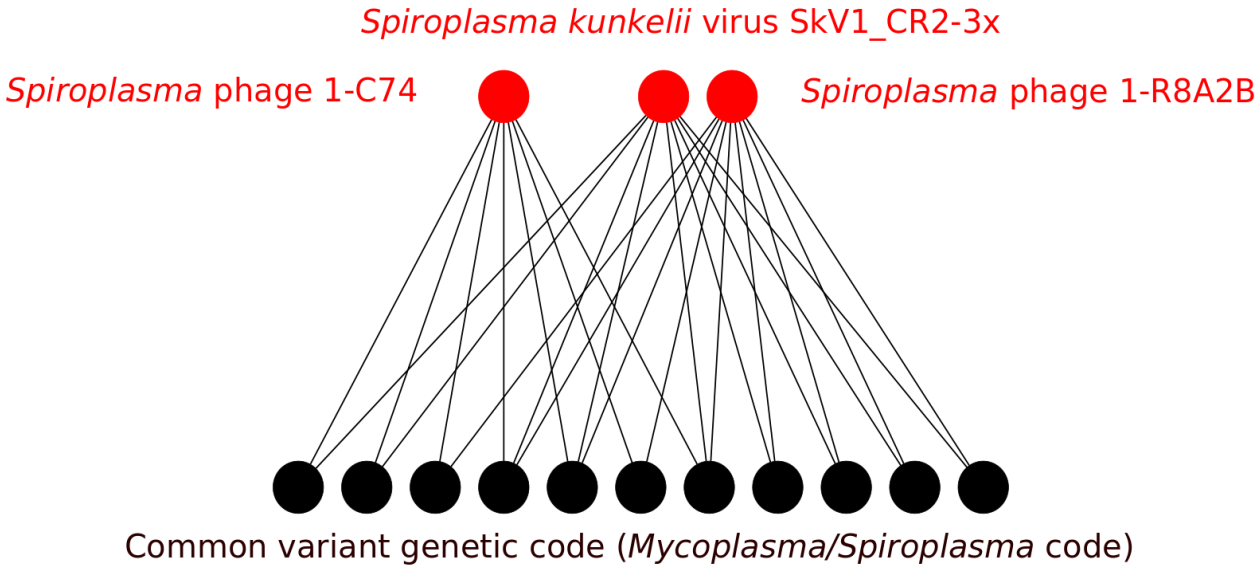
encoding enzymes for carotenoid biosynthesis. *Mol. Biol. Evol.* 25, 2653–67
192 Oborník, M. and Green, B.R. (2005) Mosaic origin of the heme biosynthesis pathway in photosynthetic eukaryotes. *Mol. Biol. Evol.* 22, 2343–53
193 Reyes-Prieto, A. and Bhattacharya, D. (2007) Phylogeny of Calvin cycle enzymes supports Plantae monophyly. *Mol. Phylogenet. Evol.* 45, 384–91
194 Grau-Bové, X. *et al.* (2014) The eukaryotic ancestor had a complex ubiquitin signalling system of archaeal origin. *Mol. Biol. Evol.* 32, 726–39
195 Bogumil, D. *et al.* (2014) Integration of two ancestral chaperone systems into one: the evolution of eukaryotic molecular chaperones in light of eukaryogenesis. *Mol. Biol. Evol.* 31, 410–8
196 Shabalina, S.A. and Koonin, E. V (2008) Origins and evolution of eukaryotic RNA interference. *Trends Ecol. Evol.* 23, 578–87
197 Stroud, D.A. *et al.* (2016) Accessory subunits are integral for assembly and function of human mitochondrial complex I. *Nature* DOI: 10.1038/nature19754
198 Madhani, H.D. (2013) The frustrated gene: origins of eukaryotic gene expression. *Cell* 155, 744–9
199 Mast, F.D. *et al.* Evolutionary mechanisms for establishing eukaryotic cellular complexity. , *Trends in Cell Biology*, 24. (2014) , Elsevier Ltd, 435–442
200 Méheust, R. *et al.* (2016) Protein networks identify novel symbiogenetic genes resulting from plastid endosymbiosis. *Proc. Natl. Acad. Sci. U. S. A.* 113, 3579–84
201 Méheust, R. *et al.* (2015) Metabolic bacterial genes and the construction of high-level composite lineages of life. *Trends Ecol. Evol.* 30, 127–9
202 Hurst, L.D. *et al.* (2004) The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.* 5, 299–310
203 Gori, K. *et al.* (2016) Clustering genes of common evolutionary history. *Mol. Biol. Evol.* DOI: 10.1093/molbev/msw038
204 Boycheva, S. *et al.* (2014) The rise of operon-like gene clusters in plants. *Trends Plant Sci.* 19, 447–59
205 Wong, S. and Wolfe, K.H. (2005) Birth of a metabolic gene cluster in yeast by adaptive gene relocation. *Nat. Genet.* 37, 777–82
206 Slot, J.C. and Rokas, A. (2010) Multiple GAL pathway gene clusters evolved independently and by different mechanisms in fungi. *Proc. Natl. Acad. Sci. U. S. A.* 107, 10136–41
207 Field, B. and Osbourn, A.E. (2008) Metabolic diversification--independent assembly of operon-like gene clusters in different plants. *Science* 320, 543–7
208 Shieh, Y.-W. *et al.* (2015) Operon structure and cotranslational subunit association direct protein assembly in bacteria. *Science* 350, 678–680
209 Hasnain, G. *et al.* (2013) Identification and characterization of the missing pyrimidine reductase in the plant riboflavin biosynthesis pathway. *Plant Physiol.* 161, 48–56
210 Frelin, O. *et al.* (2014) A directed-overflow and damage-control N-glycosidase in riboflavin biosynthesis. *Biochem. J.* DOI: 10.1042/BJ20141237
211 Hanson, A.D. *et al.* (2016) Metabolite Damage and Metabolite Damage Control in Plants. *Annu. Rev. Plant Biol.* 67, 131–52

VII. Annexes

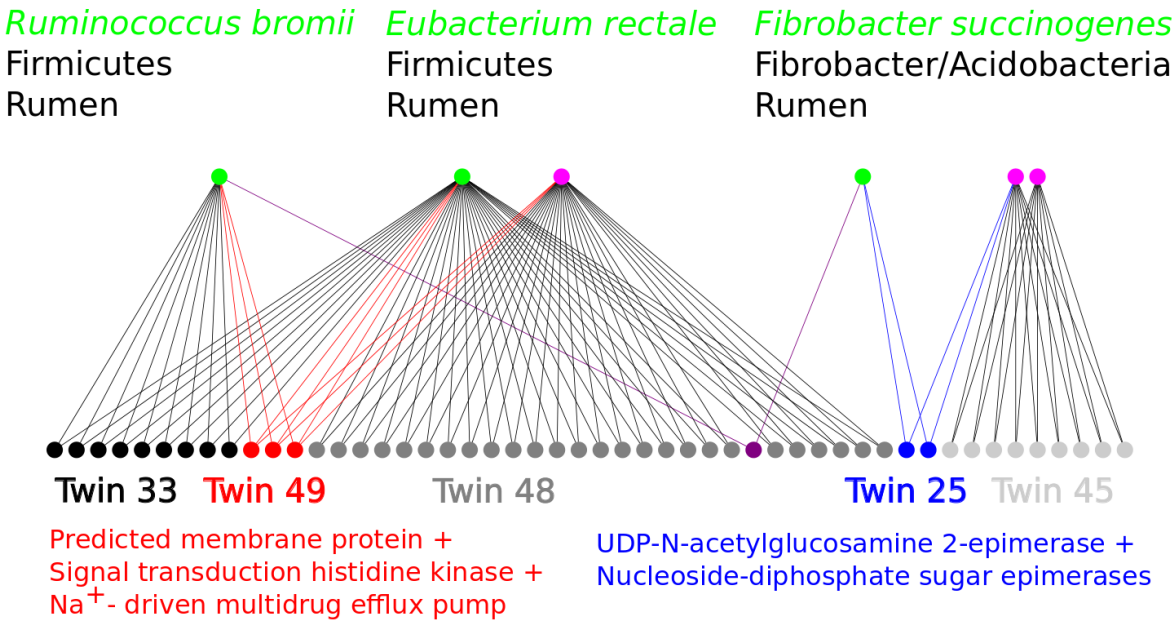
Les pages suivantes correspondent au matériel supplémentaire de l'article sur les graphes bipartis et l'externalisation de gènes (Article IV, chapitre III.A.1).

Supplementary Material

Suppl. Fig. 1. Connected component of the bipartite graph isolating a common variant of the genetic code shared by the three viral genomes on top, and the shared gene family content on the bottom.



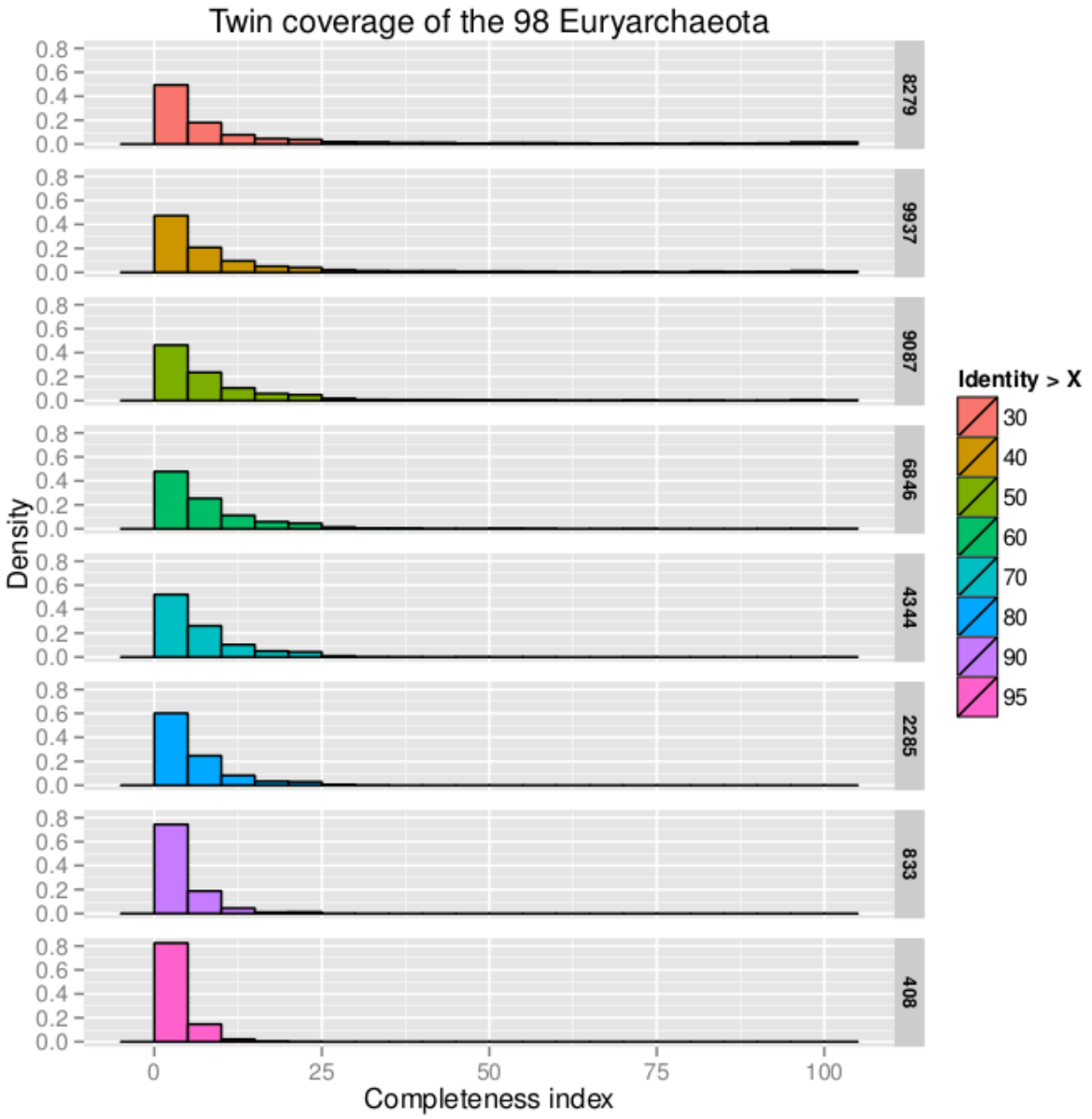
Suppl. Fig. 2. Connected component isolating ruminal bacteria and their associated plasmids.

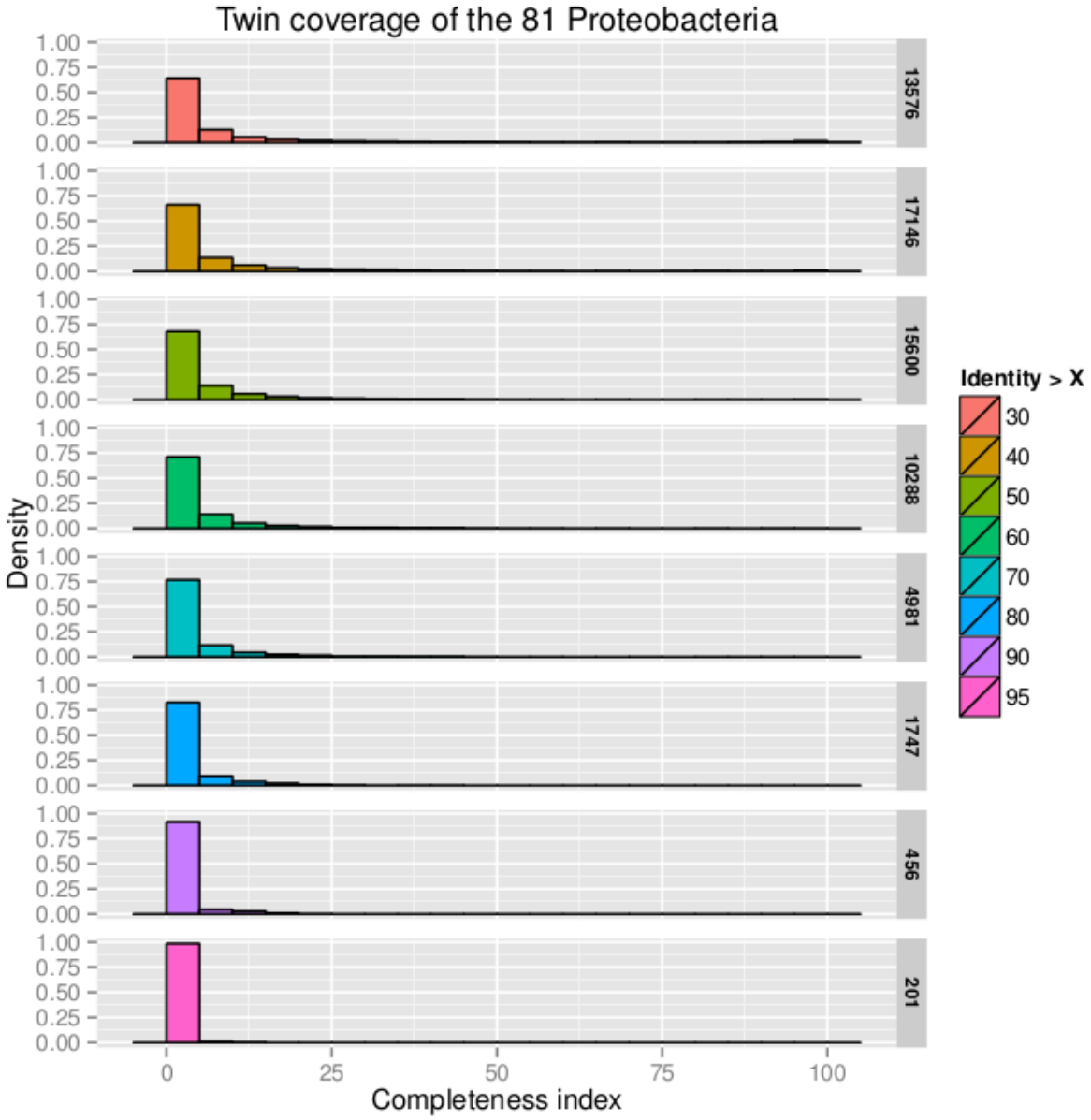


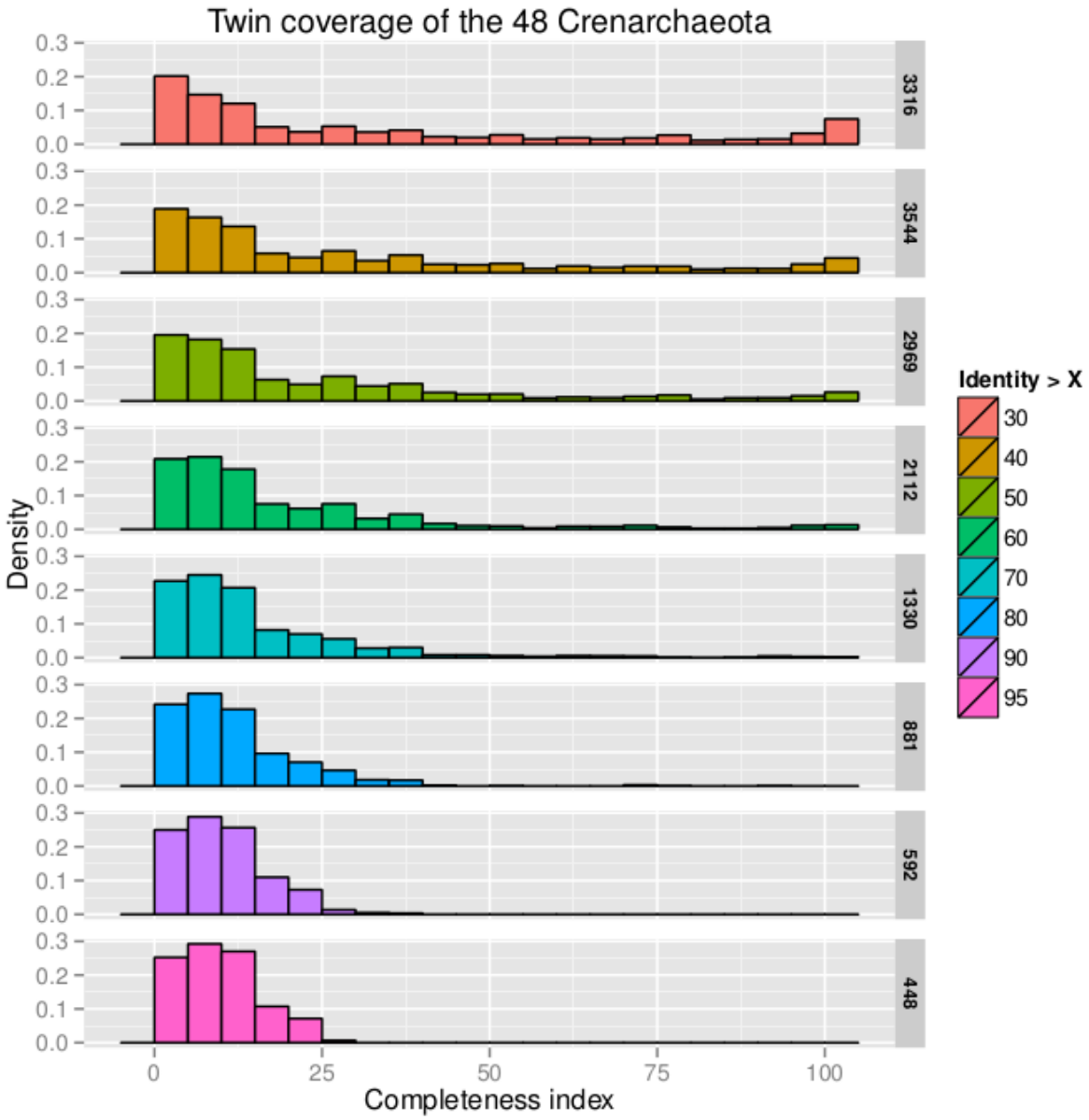
Twins 25 (blue) and 49 (red) exhibit functional properties related to environmental adaptivity. The purple bottom node is an articulation point.

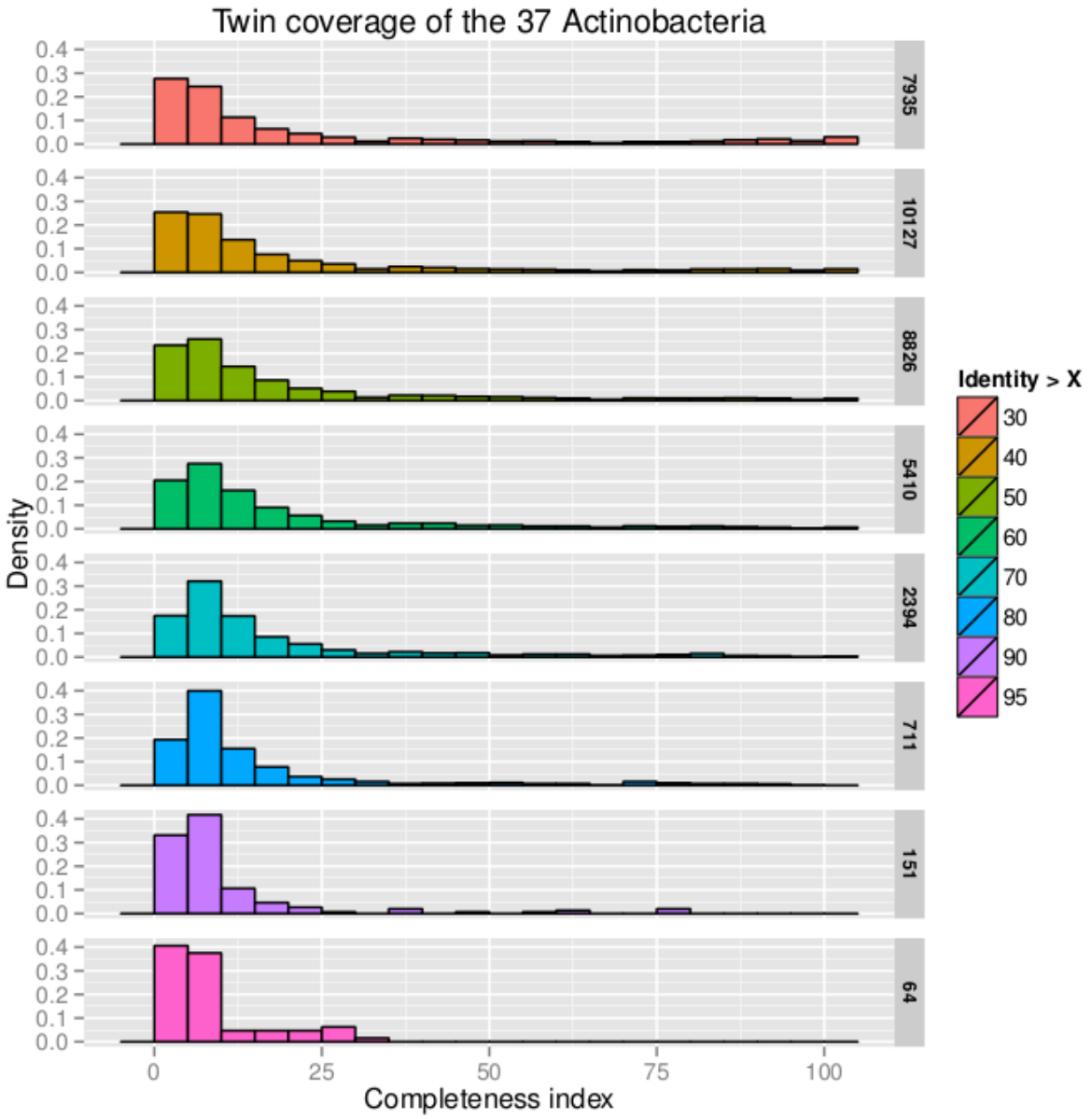
In Suppl. Fig. 3 and 4, we report the completeness of twin support relative to phylum and class, respectively. For each twin, we define the completeness index for a taxonomic category as the proportion of the members of this category that are included in the support of this twin. Core gene families correspond to twins having completeness index equal to 1 (or 100%). In the attached PDFs, we present for each threshold the distribution of the completeness indices over a given taxonomic category. On the right sidebar are reported the number of twins that the distribution is based upon. As a general trend, we confirm the patchiness of the gene family distribution, both at the phylum and the class levels. Notably, categories featuring substantial relative amounts of core gene families are environmentally (Thermococci, Thermoprotei, Halobacteria) or metabolically (Methanococci, Methanomicrobia, Cyanobacteria) specific lineages.

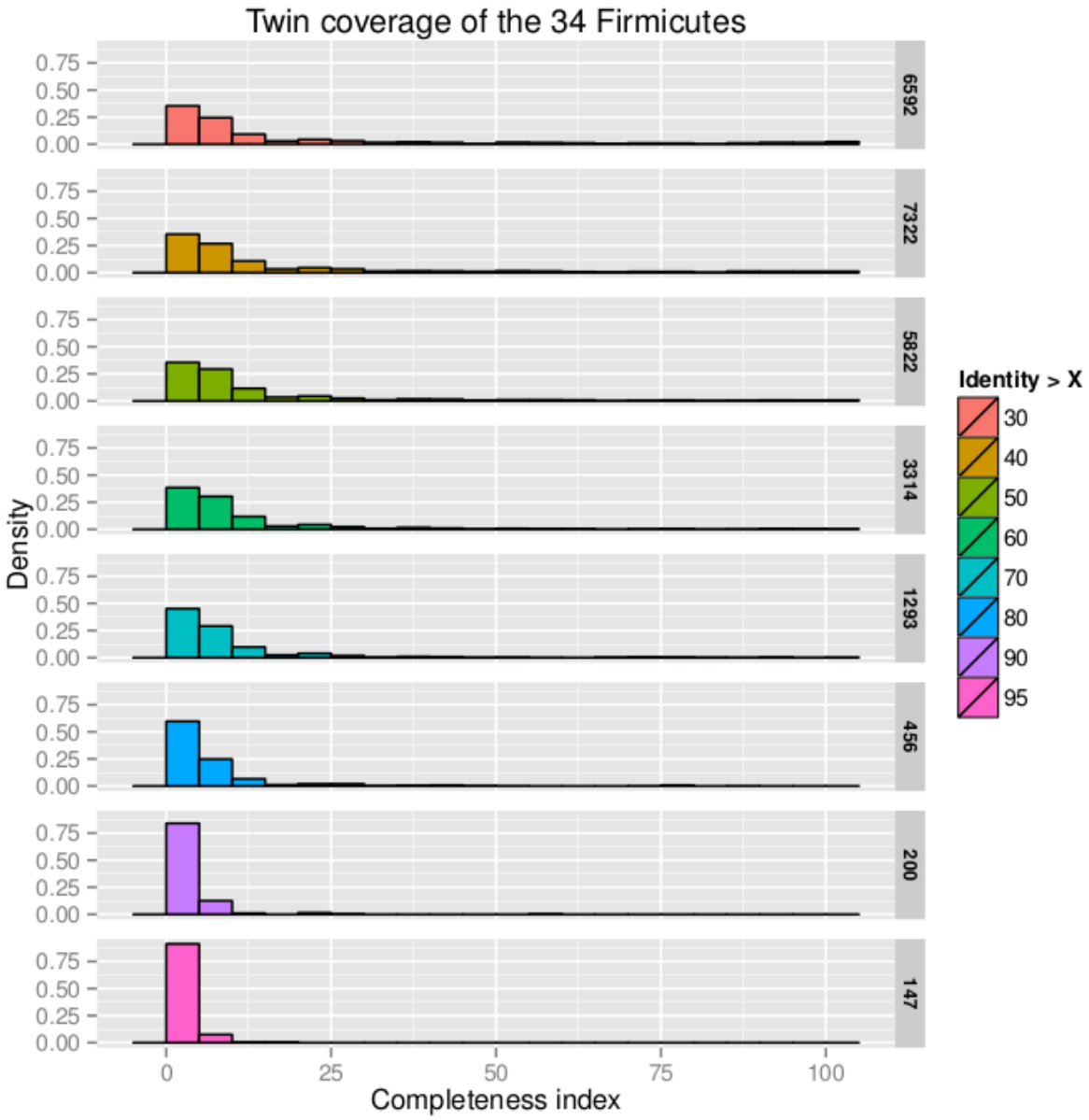
Suppl. Fig. 3-1 to 3-9. Completeness distributions for twin supports with respect to phylum. For each of the 9 phyla represented by at least 7 genomes, the graphs display for each identity threshold the distribution of the completeness indices of twins (the total number of twins being reported on the side bar).

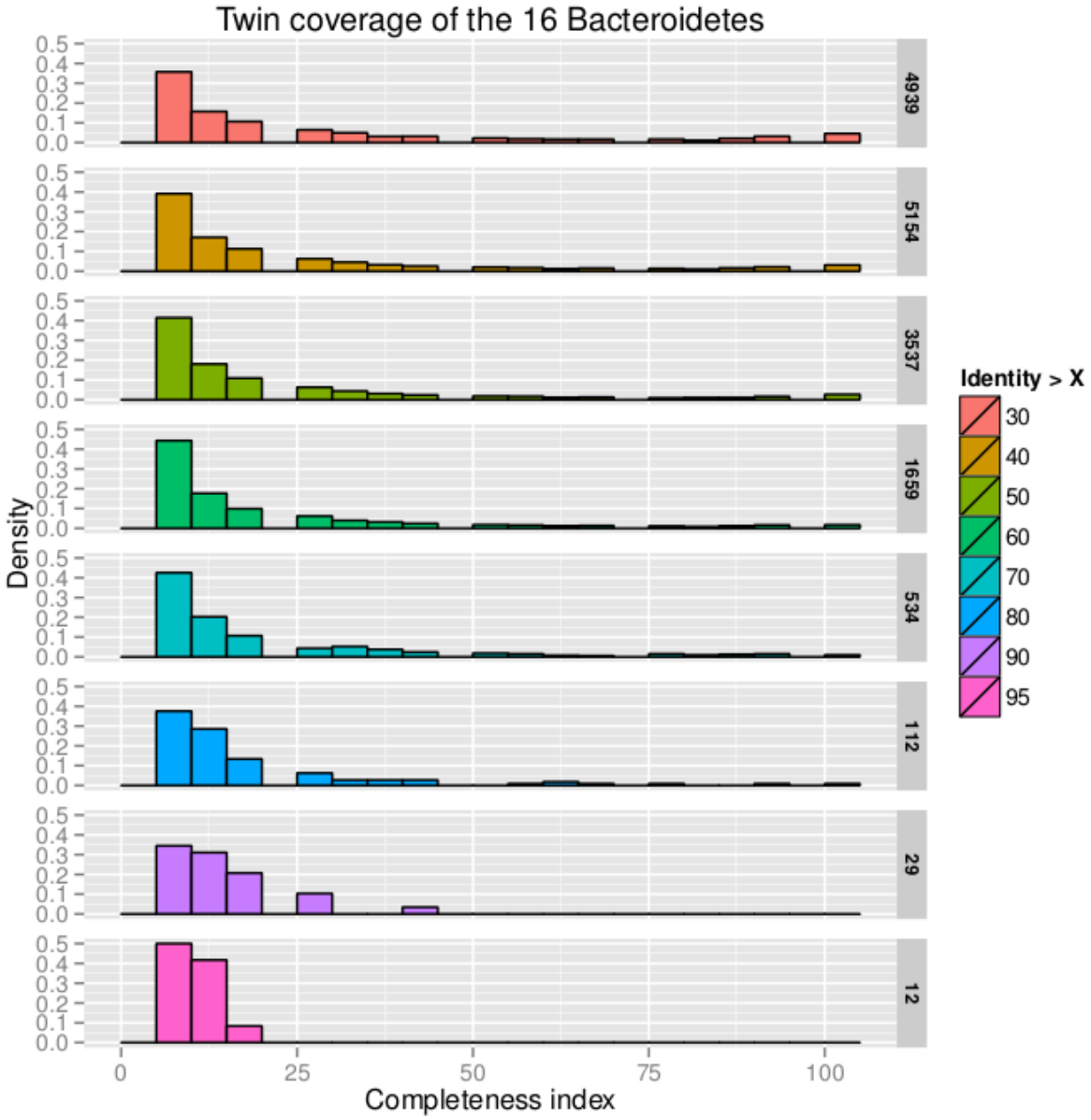

Twin coverage of the 98 Euryarchaeota
Density
0.8
0.6
0.4
0.2
0.0
8279
9937
9087
6846
4344
2285
833
408
0
25
50
75
100
Completeness index
Identity > X
30
40
50
60
70
80
90
95

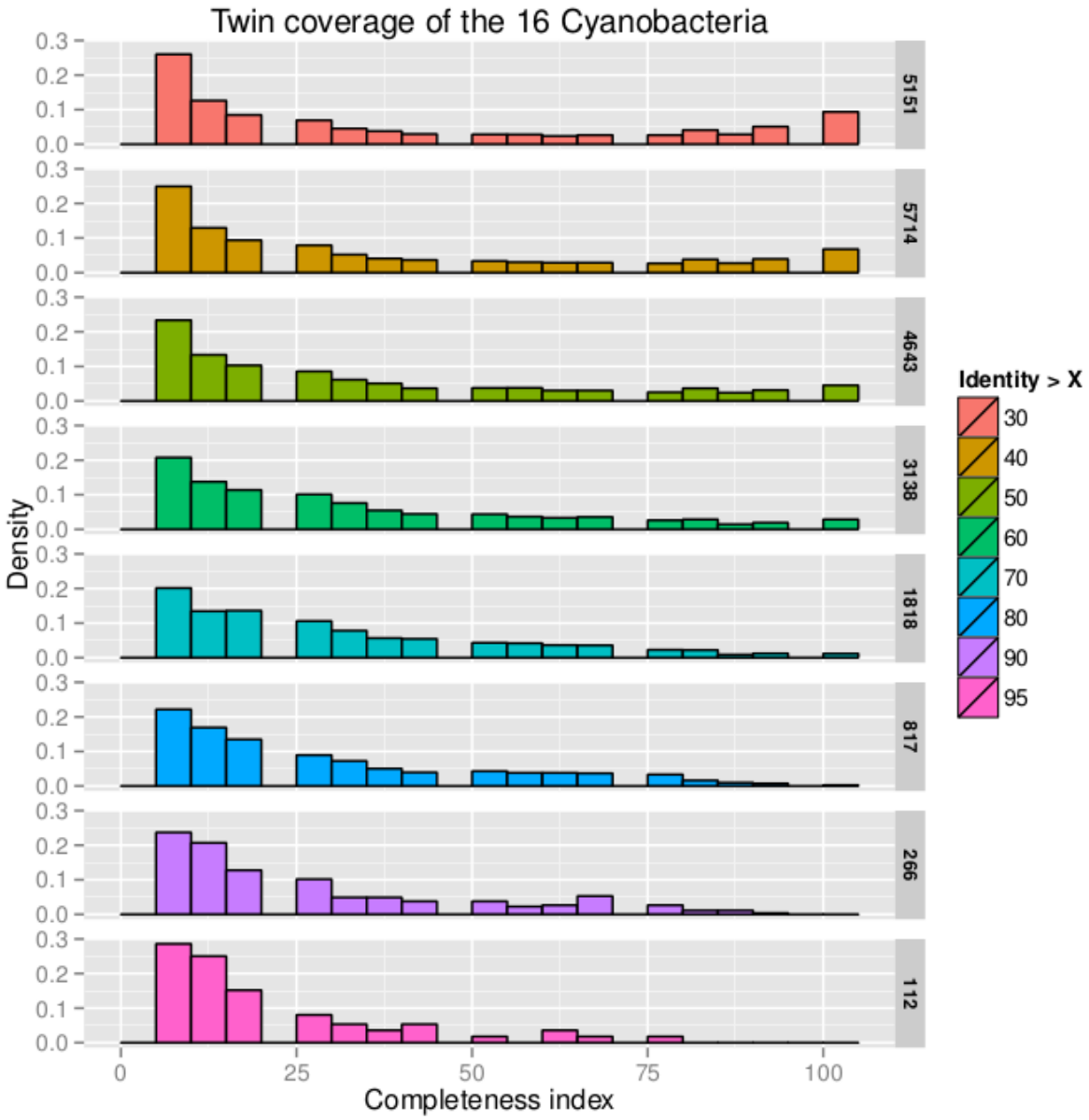

Twin coverage of the 81 Proteobacteria
Density
1.00
0.75
0.50
0.25
0.00
13576
17146
15600
10288
4981
1747
456
201
0
25
50
75
100
Completeness index
Identity > X
30
40
50
60
70
80
90
95

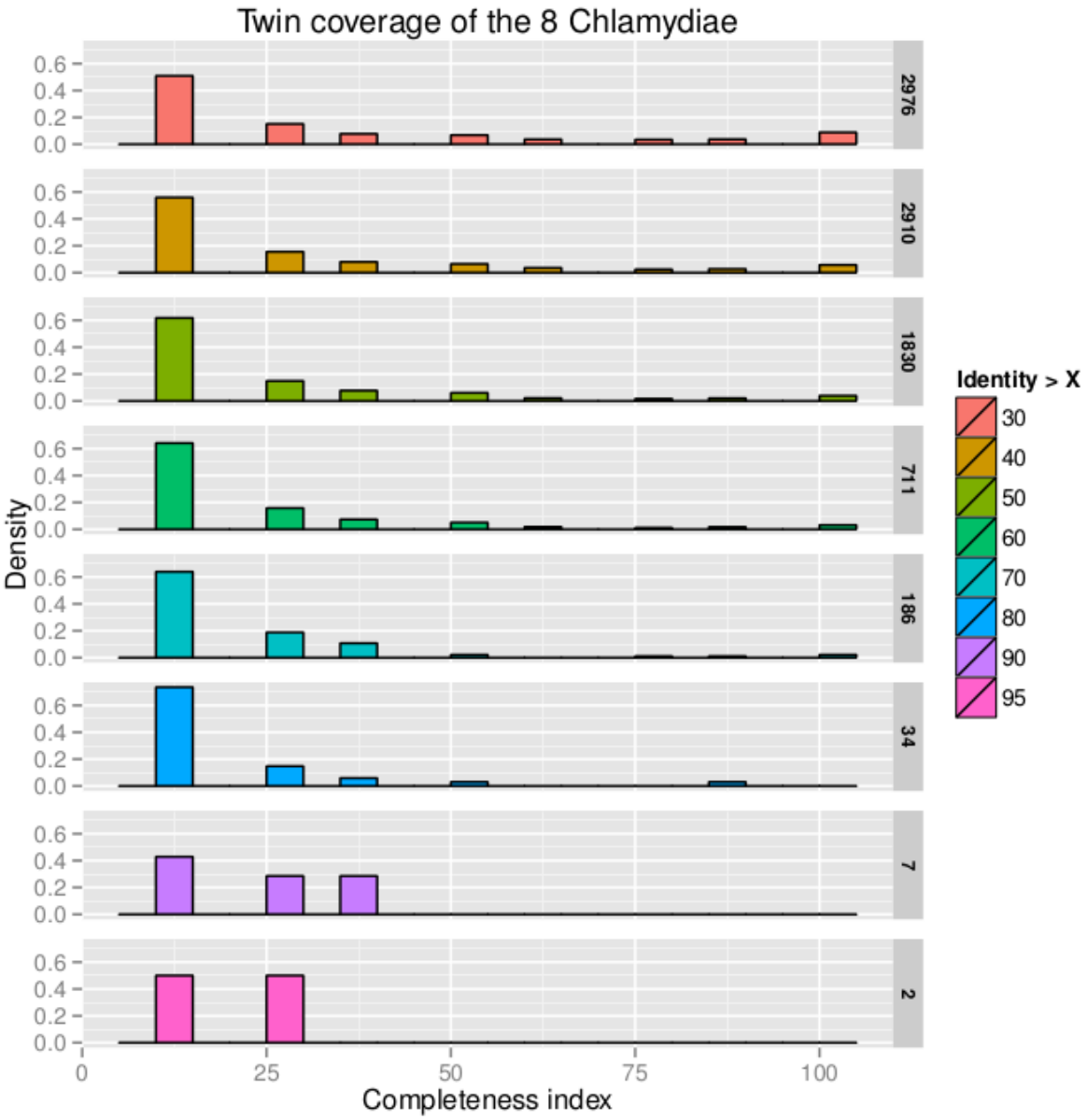

Twin coverage of the 48 Crenarchaeota
Density
0.3
0.2
0.1
0.0
3316
3544
2969
2112
1330
881
592
448
0
25
50
75
100
Completeness index
Identity > X
30
40
50
60
70
80
90
95

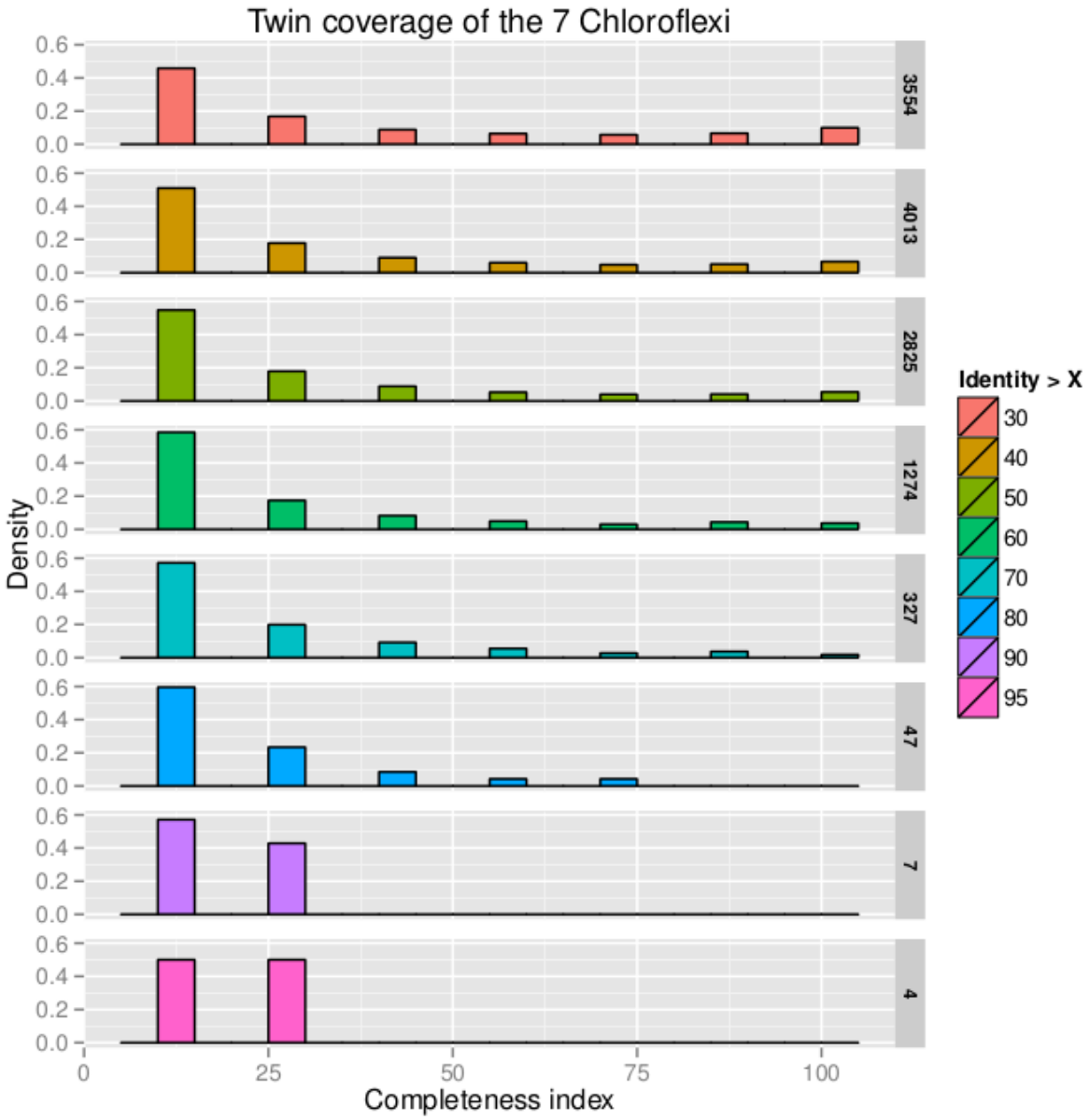

Twin coverage of the 37 Actinobacteria
Density
0.4
0.3
0.2
0.1
0.0
7935
10127
8826
5410
2394
711
151
64
0
25
50
75
100
Completeness index
Identity > X
30
40
50
60
70
80
90
95


Twin coverage of the 34 Firmicutes
Density
0.75
0.50
0.25
0.00
6592
7322
5822
3314
1293
456
200
147
0
25
50
75
100
Completeness index
Identity > X
30
40
50
60
70
80
90
95

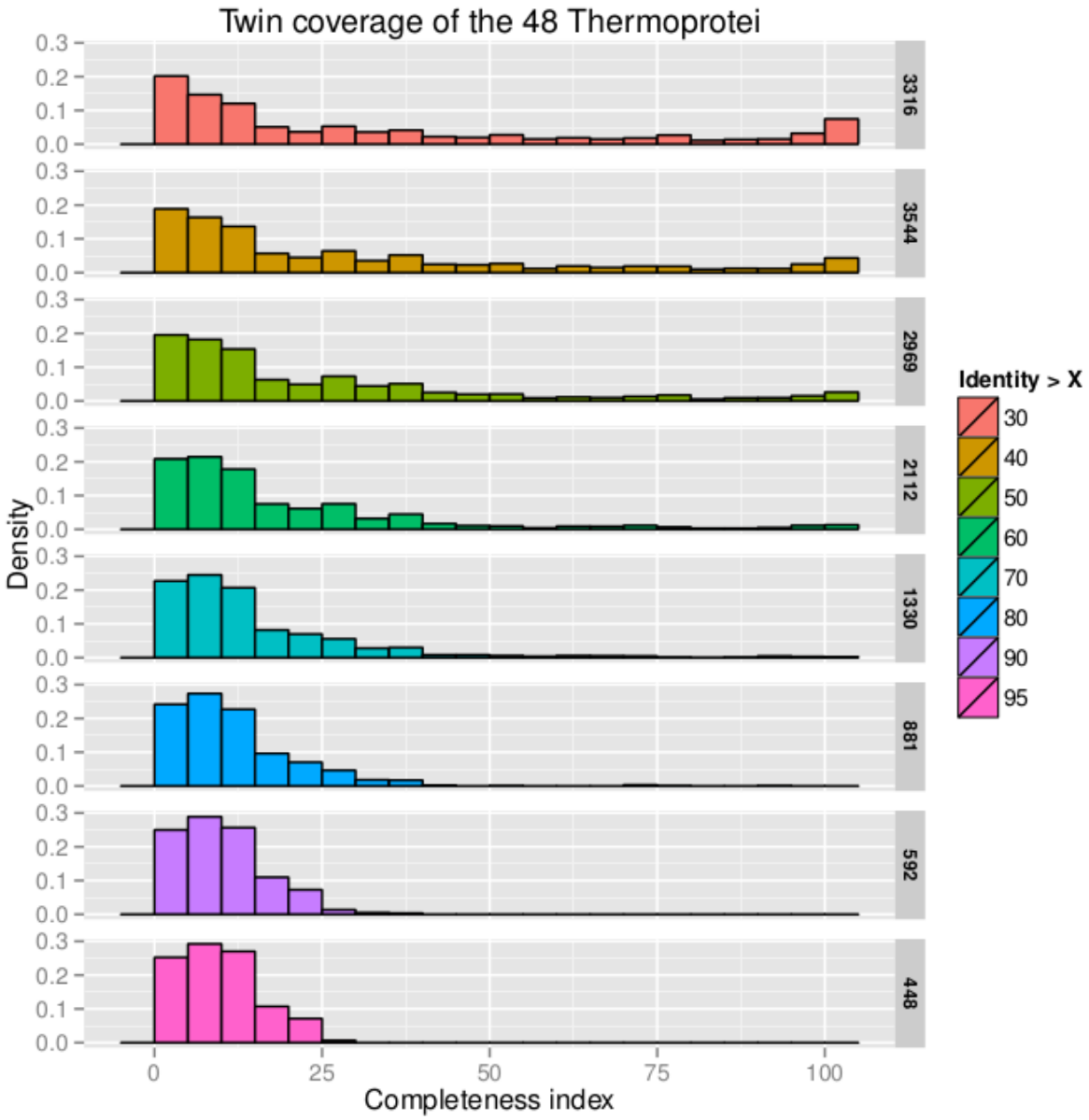

Twin coverage of the 16 Bacteroidetes
Density
0.5
0.4
0.3
0.2
0.1
0.0
4939
5154
3537
1659
534
112
29
12
0
25
50
75
100
Completeness index
Identity > X
30
40
50
60
70
80
90
95

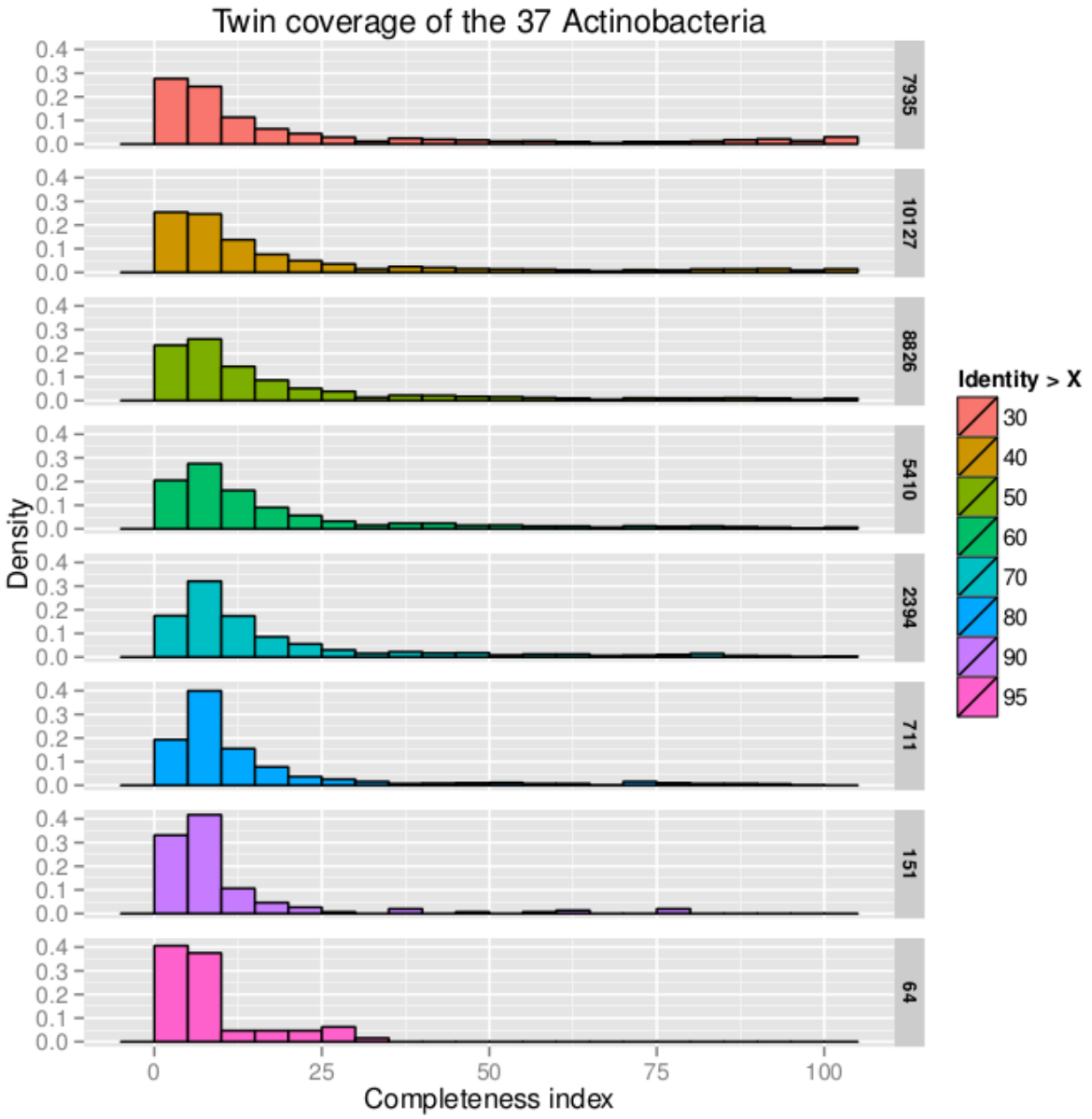

Twin coverage of the 16 Cyanobacteria
Density
0.3
0.2
0.1
0.0
5151
5714
4643
3138
1818
817
266
112
0
25
50
75
100
Completeness index
Identity > X
30
40
50
60
70
80
90
95

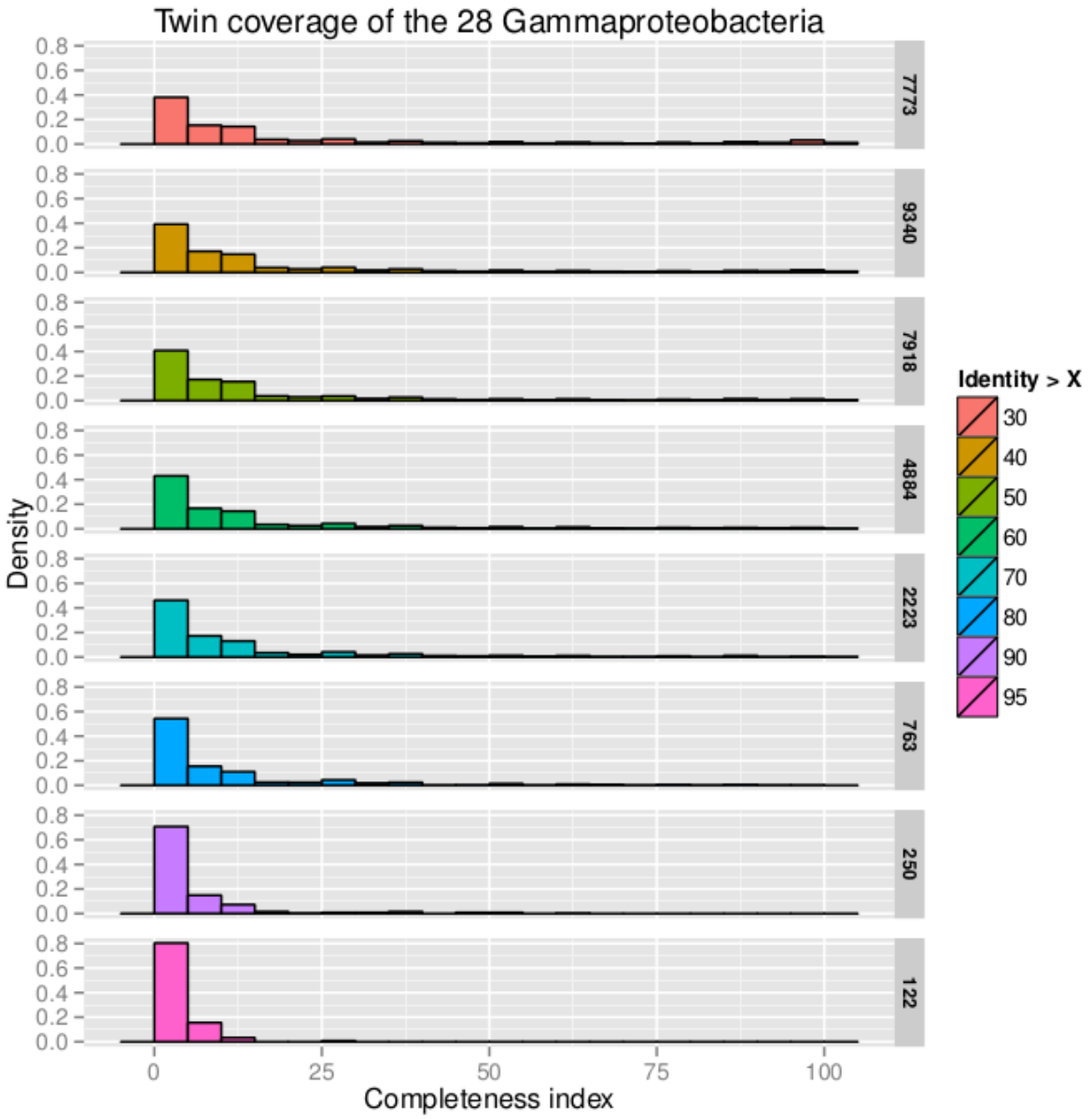

Twin coverage of the 8 Chlamydiae
Density
0.6
0.4
0.2
0.0
2976
2910
1830
711
186
34
7
2
0
25
50
75
100
Completeness index
Identity > X
30
40
50
60
70
80
90
95

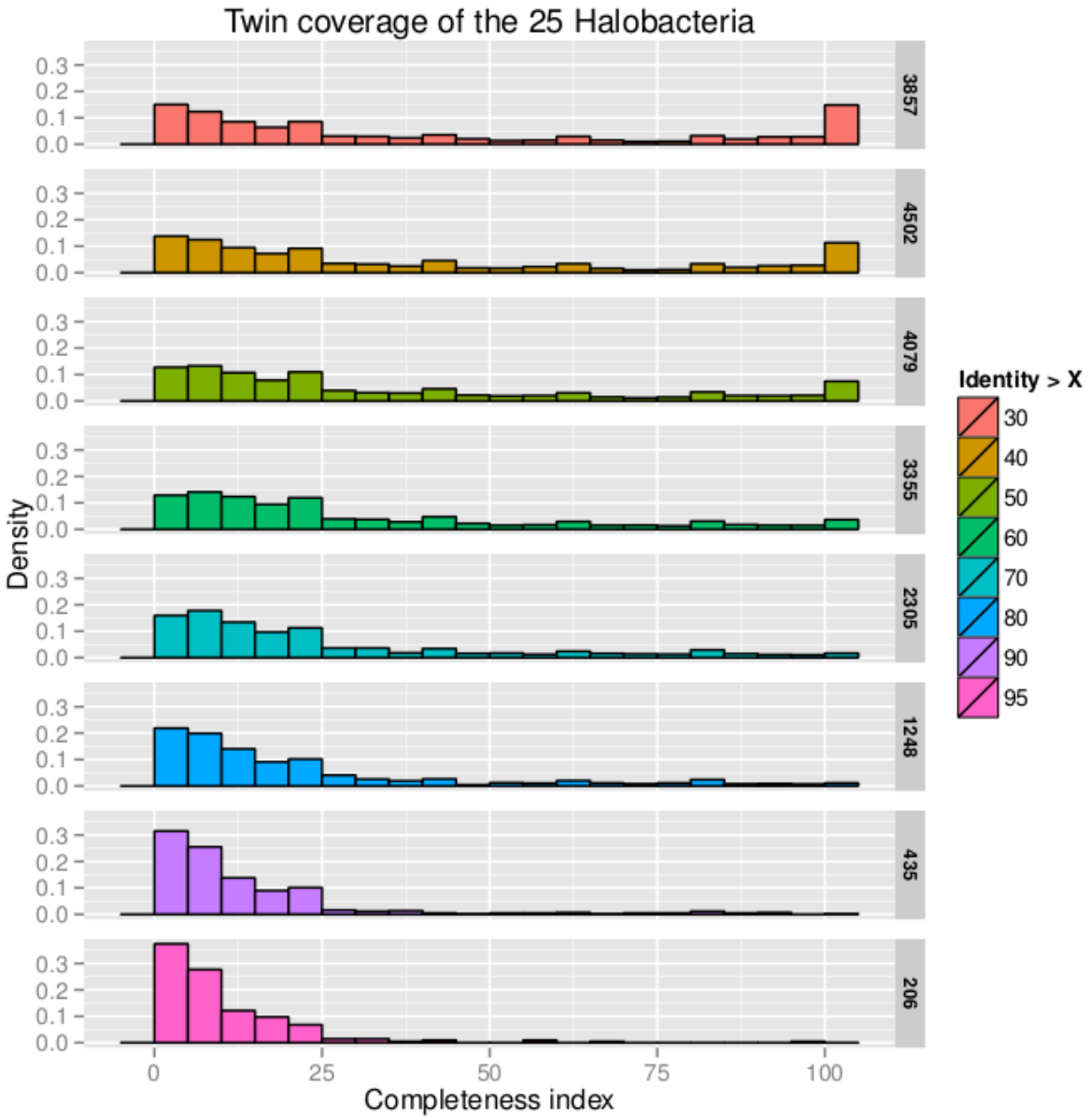


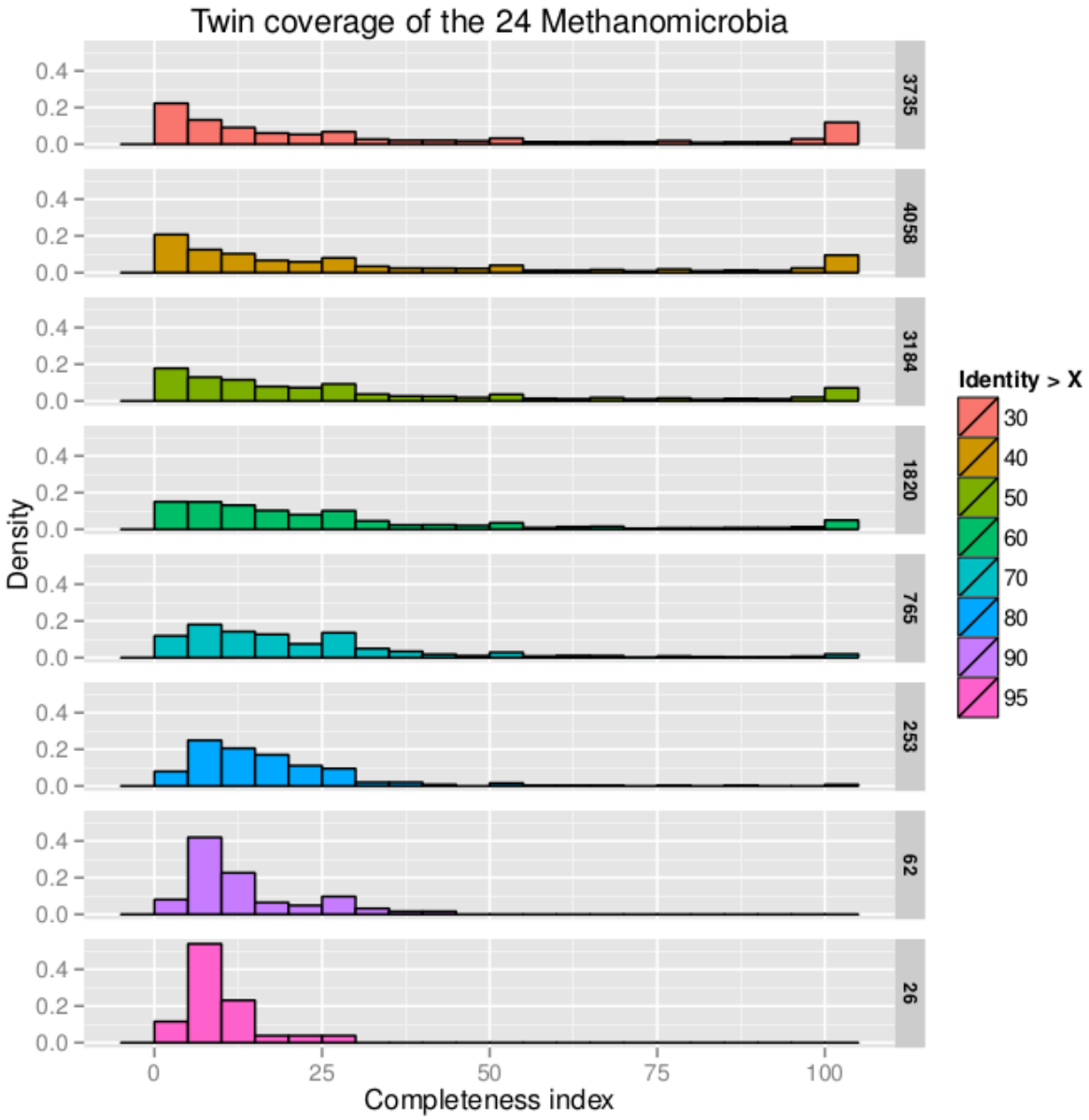
Suppl. Fig. 4-1 to 4-14. Completeness distributions for the twin supports with respect to class. For each of the 14 classes represented by at least 9 genomes, we display for each identity threshold the distribution of the completeness indices of twins (total number of twins reported on the side bar).

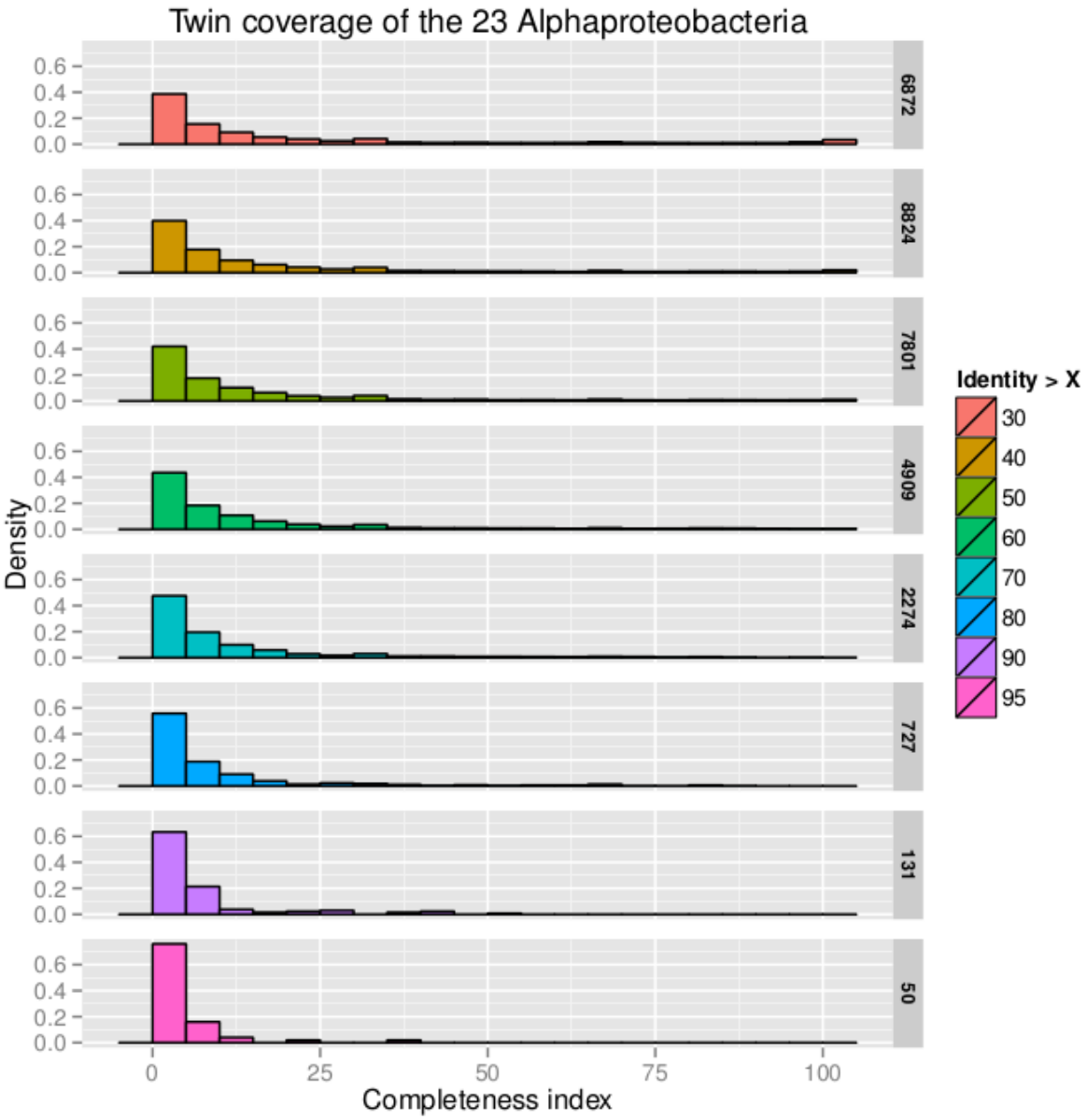

Twin coverage of the 48 Thermoprotei
Density
0.3
0.2
0.1
0.0
3316
3544
2969
2112
1330
881
592
448
0
25
50
75
100
Completeness index
Identity > X
30
40
50
60
70
80
90
95

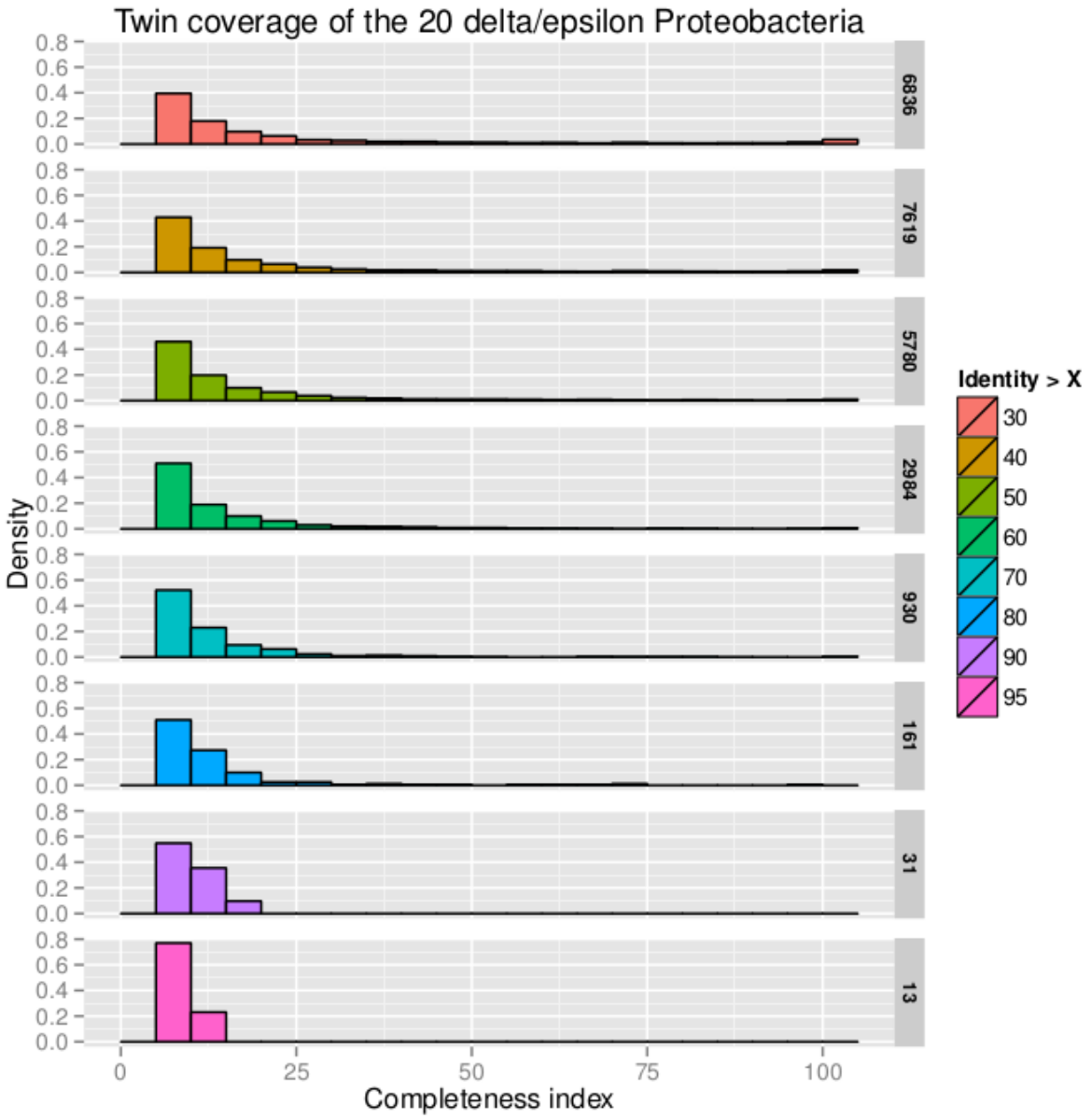

Twin coverage of the 37 Actinobacteria
Density
0.4
0.3
0.2
0.1
0.0
7935
10127
8826
5410
2394
711
151
64
0
25
50
75
100
Completeness index
Identity > X
30
40
50
60
70
80
90
95

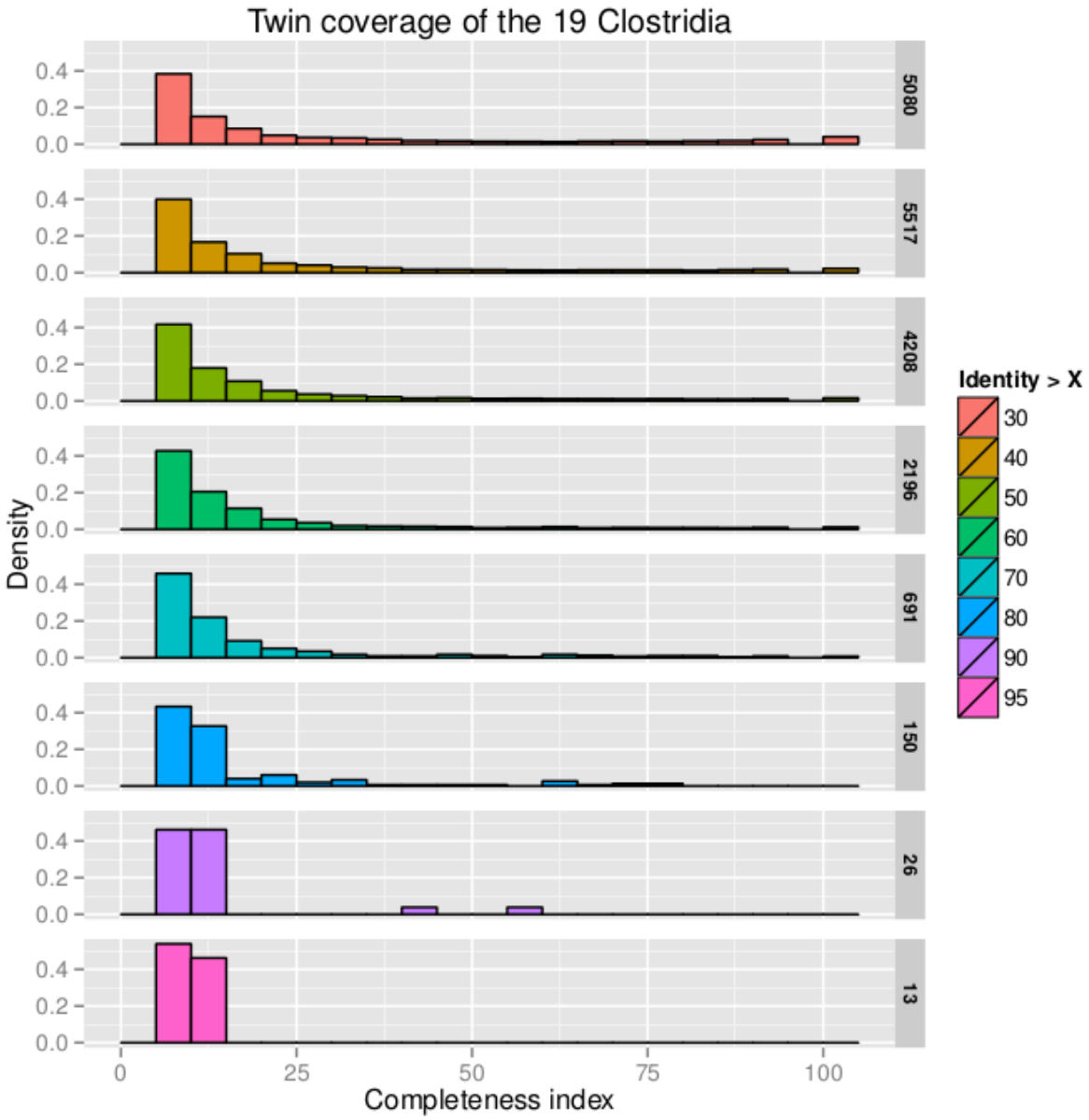

Twin coverage of the 28 Gammaproteobacteria
Density
0.8
0.6
0.4
0.2
0.0
7773
9340
7918
4884
2223
763
250
122
0
25
50
75
100
Completeness index
Identity > X
30
40
50
60
70
80
90
95

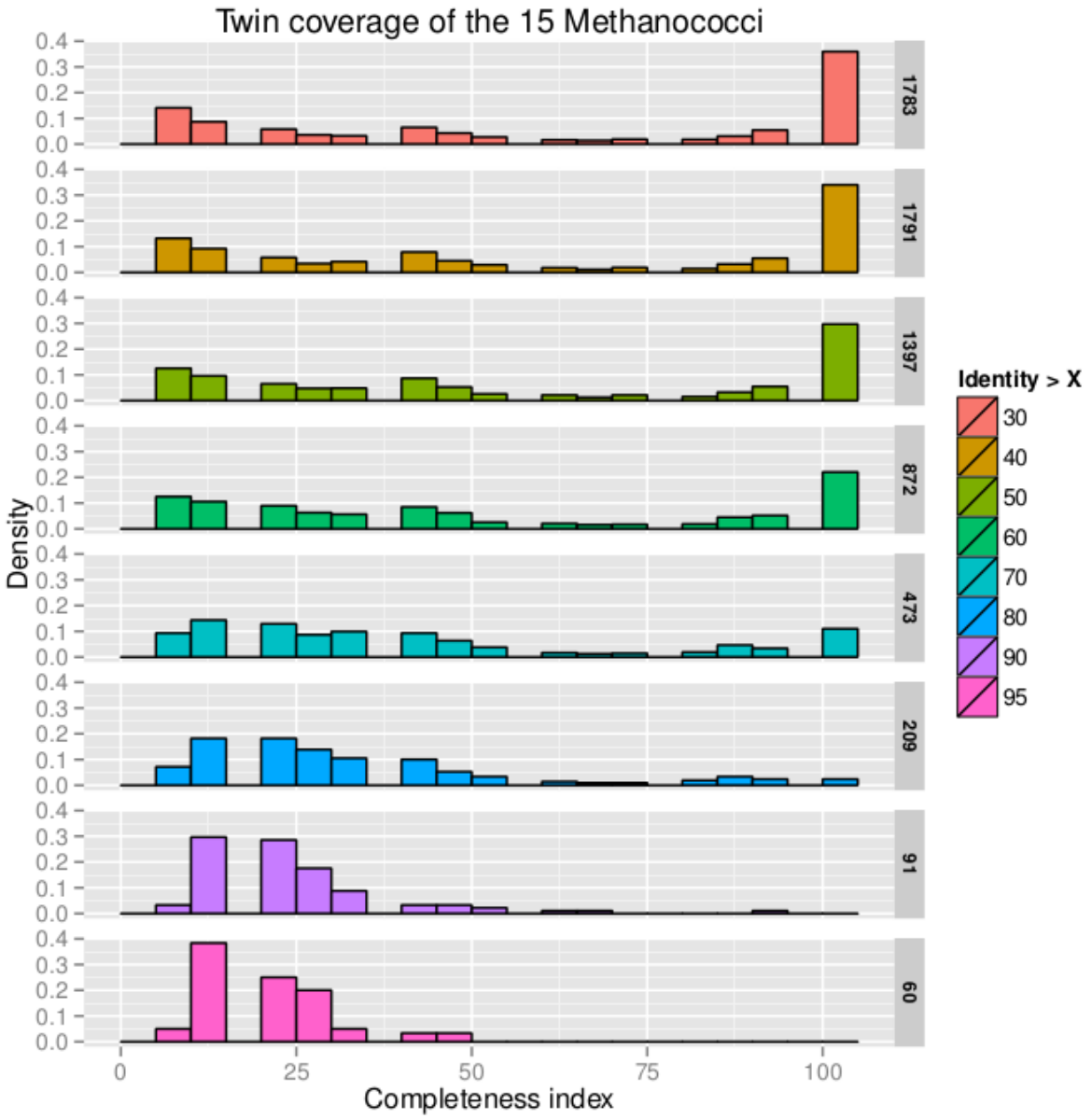

Twin coverage of the 25 Halobacteria
Density
0.3
0.2
0.1
0.0
3857
4502
4079
3355
2305
1248
435
206
0
25
50
75
100
Completeness index
Identity > X
30
40
50
60
70
80
90
95

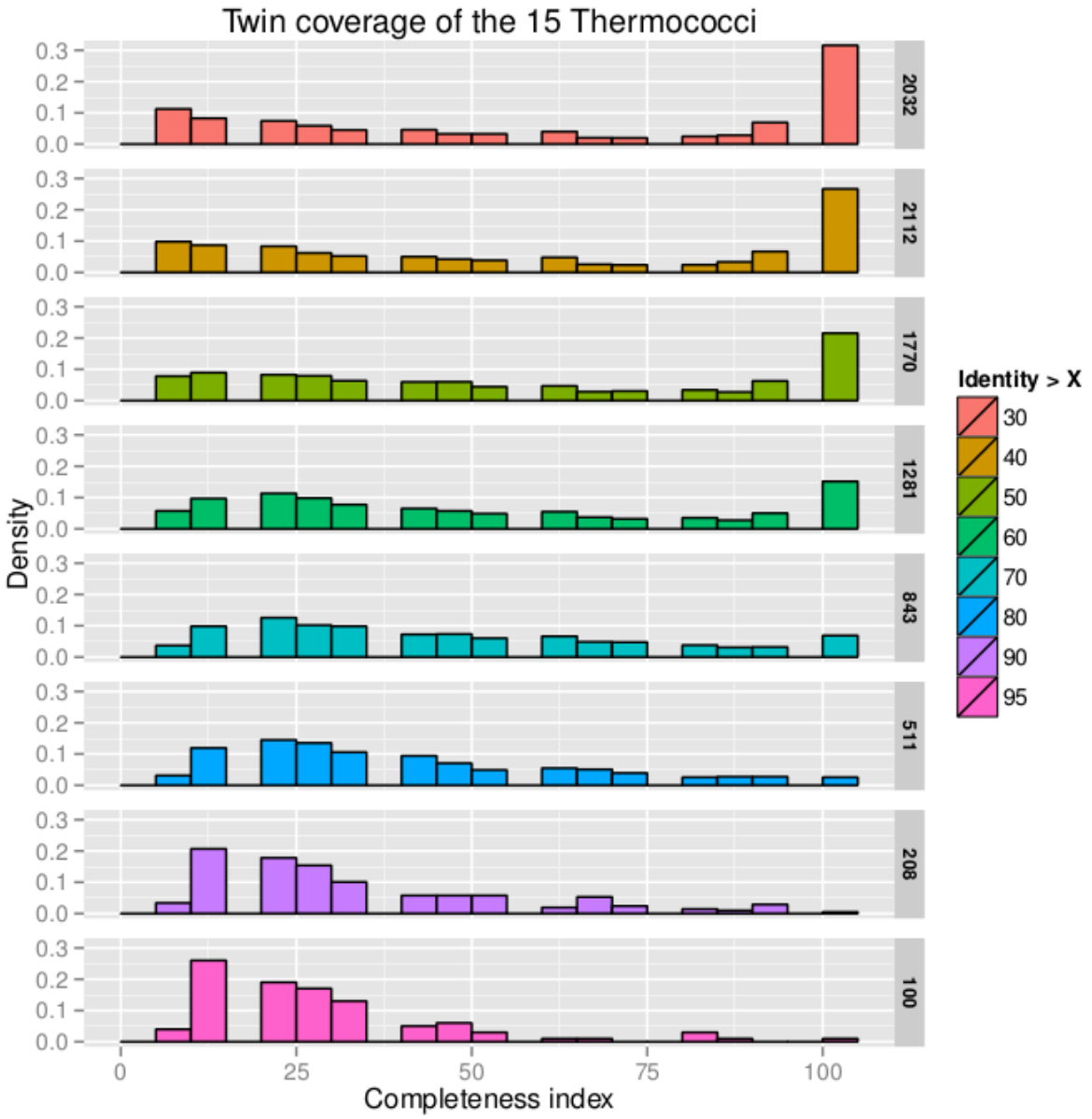

Twin coverage of the 24 Methanomicrobia
Density
0.4
0.2
0.0
3735
4058
3184
1820
765
253
62
26
0
25
50
75
100
Completeness index
Identity > X
30
40
50
60
70
80
90
95

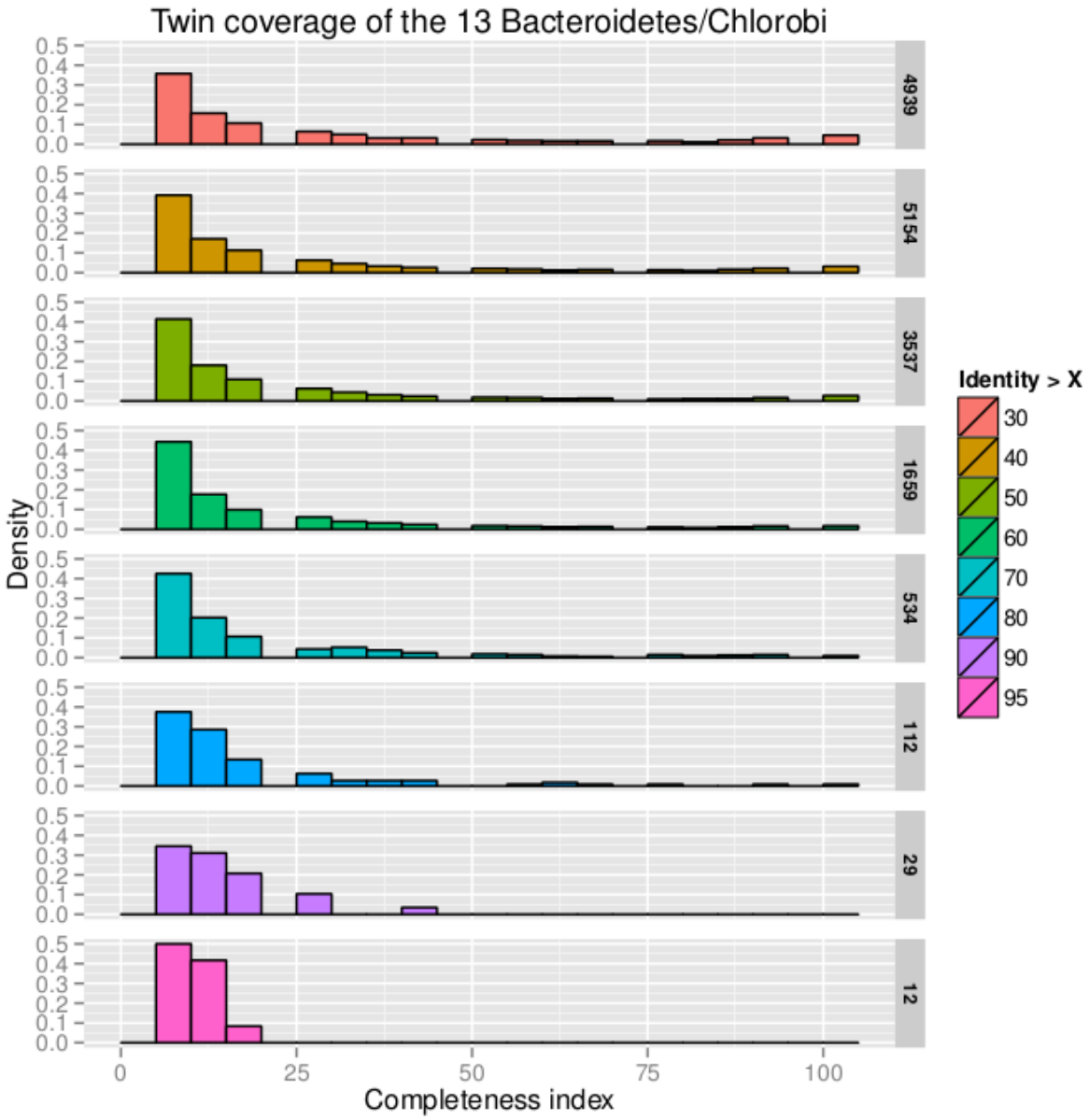

Twin coverage of the 23 Alphaproteobacteria
Density
0.6
0.4
0.2
0.0
6872
8824
7801
4909
2274
727
131
50
0
25
50
75
100
Completeness index
Identity > X
30
40
50
60
70
80
90
95

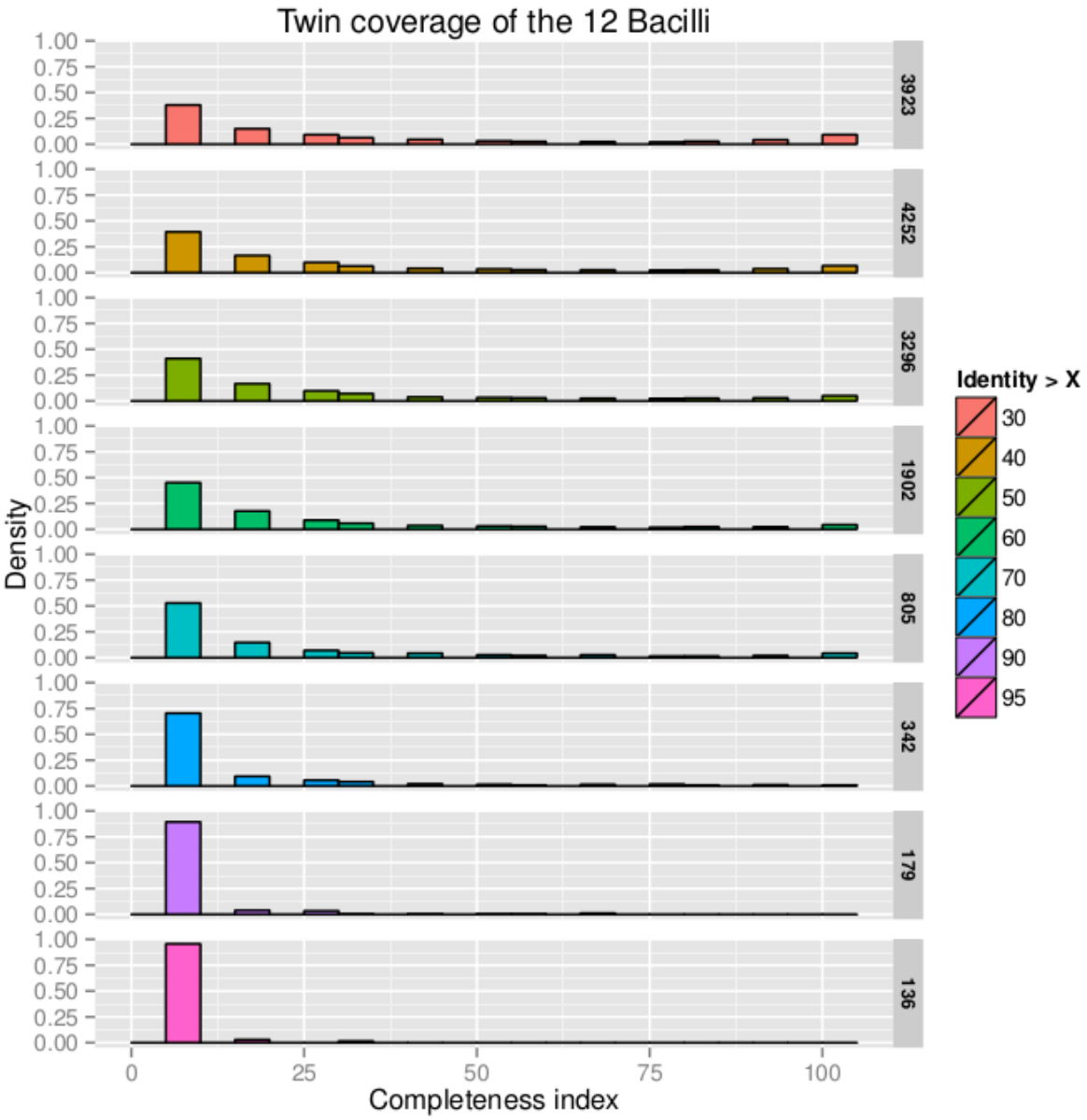

Twin coverage of the 20 delta/epsilon Proteobacteria
Density
0.8
0.6
0.4
0.2
0.0
6836
7619
5780
2984
930
161
31
13
0
25
50
75
100
Completeness index
Identity > X
30
40
50
60
70
80
90
95

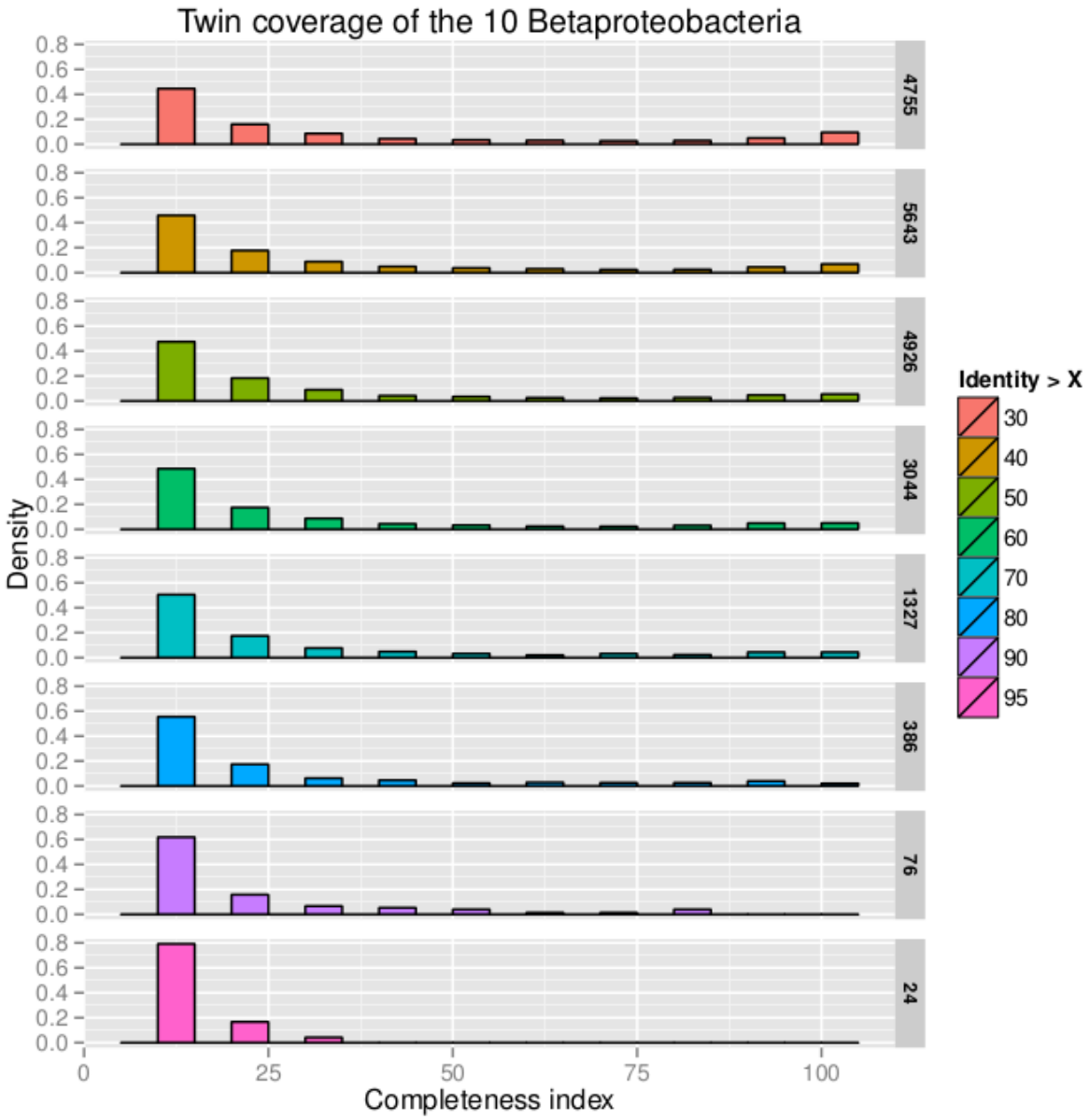

Twin coverage of the 19 Clostridia
Density
0.4
0.2
0.0
5080
5517
4208
2196
691
150
26
13
0
25
50
75
100
Completeness index
Identity > X
30
40
50
60
70
80
90
95

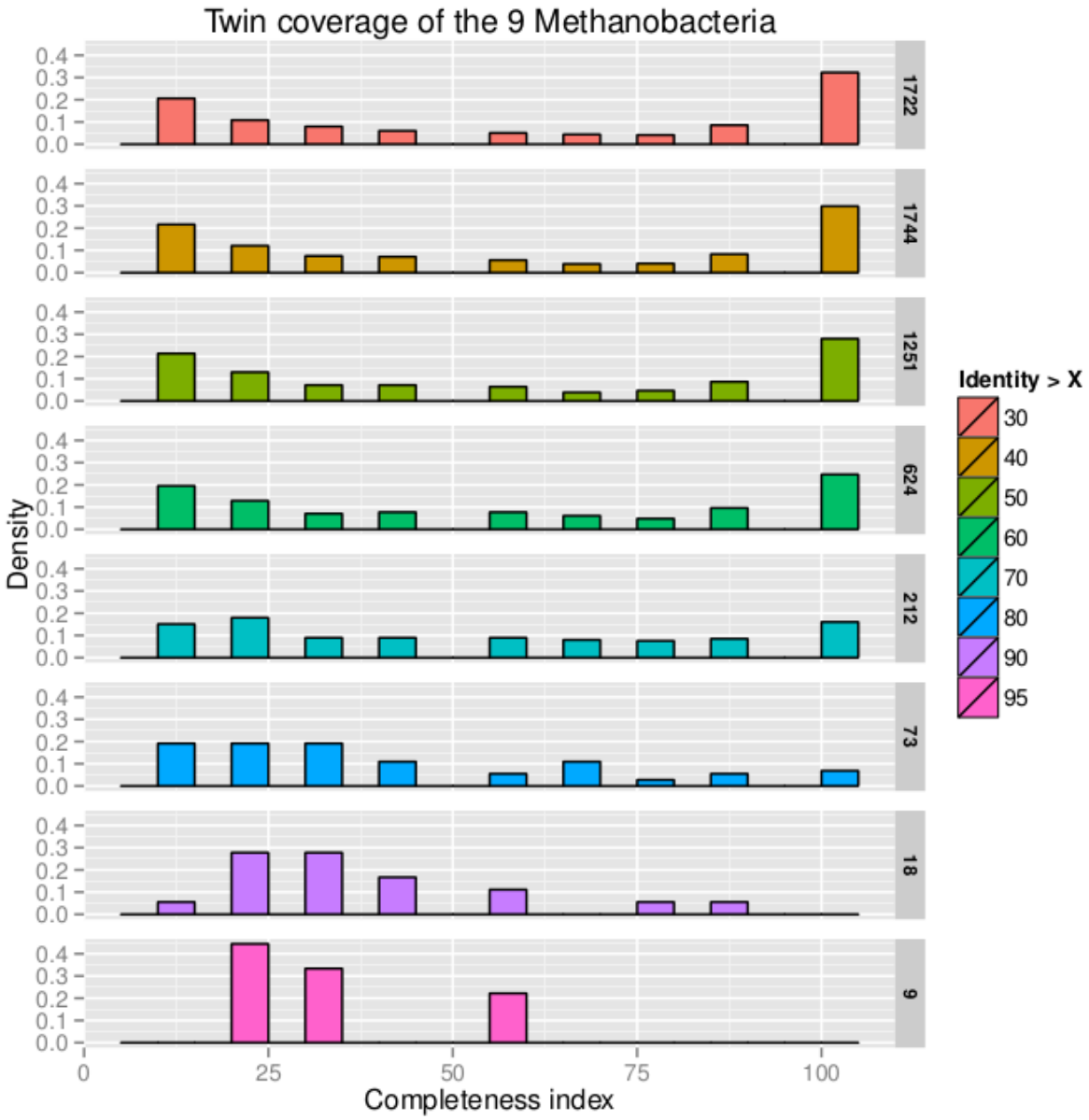

Twin coverage of the 15 Methanococci
Density
0.4
0.3
0.2
0.1
0.0
1783
1791
1397
872
473
209
91
60
0
25
50
75
100
Completeness index
Identity > X
30
40
50
60
70
80
90
95


Twin coverage of the 15 Thermococci
Density
0.3
0.2
0.1
0.0
2032
2112
1770
1281
843
511
208
100
0
25
50
75
100
Completeness index
Identity > X
30
40
50
60
70
80
90
95

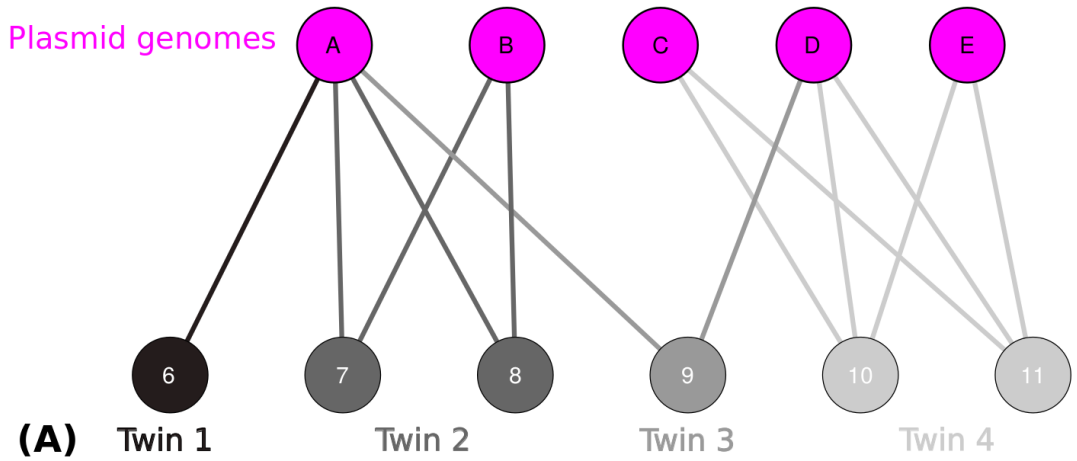

Twin coverage of the 13 Bacteroidetes/Chlorobi
Density
0.5
0.4
0.3
0.2
0.1
0.0
4939
5154
3537
1659
534
112
29
12
0
25
50
75
100
Completeness index
Identity > X
30
40
50
60
70
80
90
95


Twin coverage of the 12 Bacilli
Density
1.00
0.75
0.50
0.25
0.00
3923
4252
3296
1902
805
342
179
136
0
25
50
75
100
Completeness index
Identity > X
30
40
50
60
70
80
90
95


Twin coverage of the 10 Betaproteobacteria
Density
0.8
0.6
0.4
0.2
0.0
4755
5643
4926
3044
1327
386
76
24
0
25
50
75
100
Completeness index
Identity > X
30
40
50
60
70
80
90
95



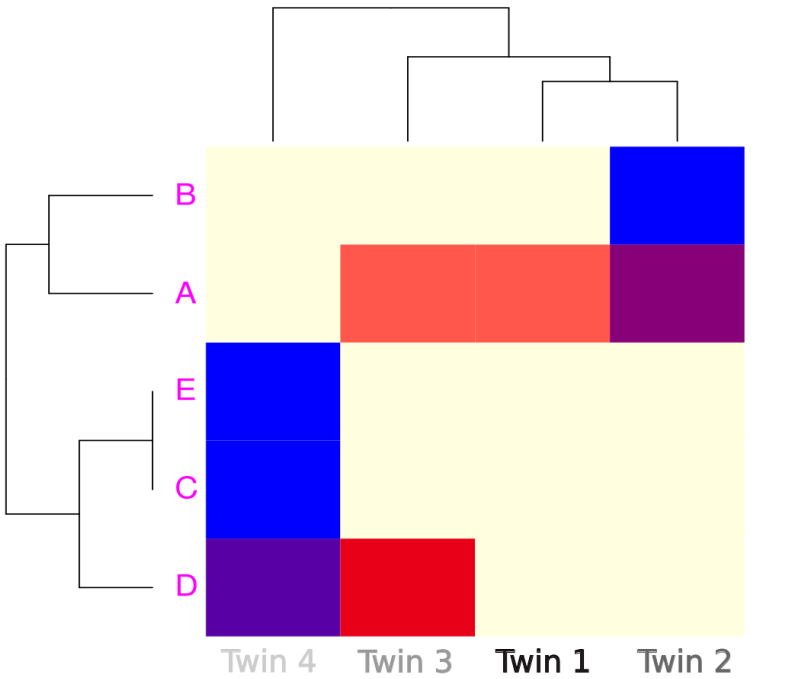
Suppl. Fig. 5. Construction of the heatmap for plasmids from the reduced genome-gene families bipartite graph.



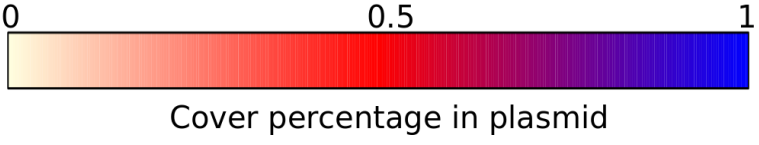
(A)

	Twin 1	Twin 2	Twin 3	Twin 4
Plasmid A	*	*	*	-
Plasmid B	-	*	-	-
Plasmid C	-	-	-	*
Plasmid D	-	-	*	*
Plasmid E	-	-	-	*

(B)



(C)

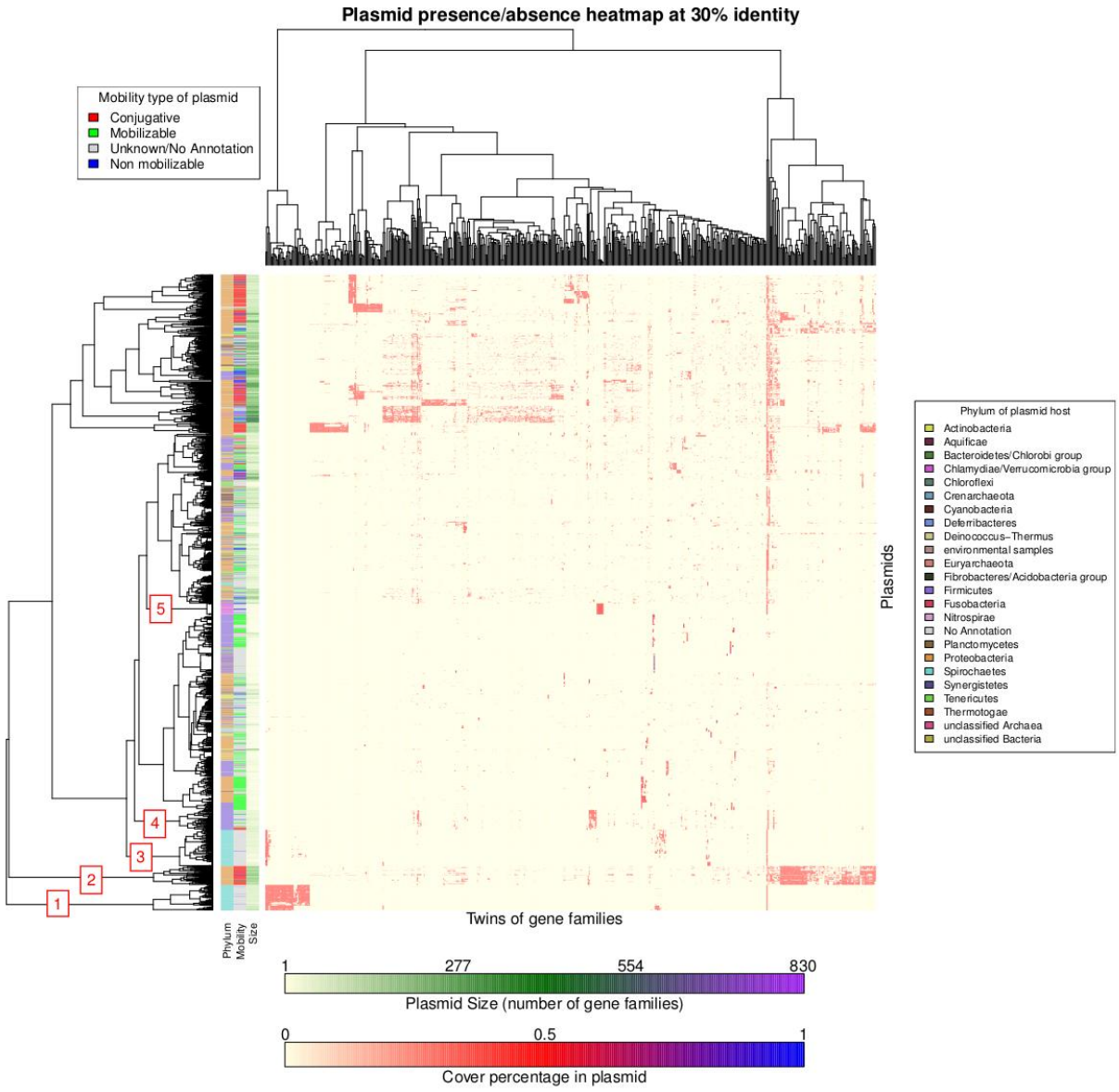


(A) Plasmid-gene families bipartite graph. The size of a plasmid is taken to be the degree of the top node representing it. The color of the edges highlights the detection of the twins.

(B) The matrix of presence/absence of twins in the plasmids. An asterisk represents the value 1. The cardinality of the twin is not taken into account.

(C) The heatmap is clustered according to the matrix in (B). However, its display is colored according to the coverage of the plasmid by the twin (as explained in the color bottom bar). In practice, we filter out twins contained in less than 1% of the plasmids.

Suppl. Fig. 6. Heatmap of the twin presence for plasmids in the gCC at ≥ 30% ID.



Plasmids are displayed in rows, and twin in columns. The color dots in the heatmap correspond to the coverage of the plasmid by the twin, as explained in the lower horizontal color bar. On the left side, the three bars represent respectively the phylum of the plasmid host, its mobility class, and its size (counted in number of gene families at ≥ 30% ID, color code in the upper horizontal bar).

The analysis of exclusive gene sharing between plasmid genomes reveals that the distribution of BTs across plasmids is not tree-like, in the sense that BT (*i.e.* shared gene families, exclusively found in sets of plasmid genomes) do not perfectly map onto a tree of genomes.

Therefore, it is difficult to classify plasmid genomes using their distribution of gene families [29]. A first step toward this direction can however be achieved from our bipartite gene families-plasmid genomes graphs (see also [30]), providing an appealing alternative to simple gene sharing networks (such as those occasionally used to classify viruses [31]). Despite the highly mosaic nature of their genomes, major groups of plasmids can be first defined at the level of CC in such networks. There are however multiple CCs, at most % ID thresholds, which confirms the absence of core gene families in plasmids, and recovers known aspects of genetic isolation of some plasmids. Typically, Borrelia plasmids only join the gCC in graphs with $\geq T$ % ID, where $T \leq 60$, while Phytoplasma plasmids enter the gCC for $T \leq 50$, and Spiroplasma plasmids only group in the gCC for $T \leq 40$. Thus, a stringency threshold of ≥ 30 % ID must be used to group all plasmids together in the same CC, but then this very large CC needs to be decomposed into subgroups to produce a classification of plasmids. We did this by analyzing all the BTs within the gCC in the gene family-plasmid graph at $\geq$ 30% ID (see Figures S5 and S6), producing the most inclusive matrix of gene families exclusively shared by the 4350 plasmid genomes of our dataset.

This matrix is very large (13,755 BT for 4020 genomes), in particular because many BTs are only present in few genomes (*i.e.* 96,10% of BTs are distributed in less than 1% of the genomes). 74,18% gene families (66,767 BTs out 90,006) are even found in a single plasmid, hence impossible to use to establish groups of plasmids. We discarded this comet-tail of "rare" gene families to focus on BT associated with $\geq$ 40 plasmid genomes (*i.e.* with over $\geq$ 1% of the plasmid genomes of the dataset). We sorted this matrix of BTs present in over 40 genomes using simple hierarchical clustering to cluster together plasmid genomes with the same BT. This simple protocol is sufficient to report multiple properties of plasmids evolution. We did not observe BTs spanning over large portions of plasmid genomes, which would be exclusively shared by these genomes. Abundant common gene families, constituting

strong genetic signatures to group plasmids together in a same class are thus lacking. Nonetheless, there were groups of plasmid genomes sharing multiple BTs, but each of these BT, represented by a red tick in Suppl. Fig. 6, corresponded to a reduced portion of the genomes. This proportion is equal to the ratio of the number of gene families of the shared BT divided by the plasmid genome size (estimated in number of gene families). Even though groups of plasmid genomes (defined by the tree on the left of Suppl. Fig. 6) with exclusive and relatively limited common pools of gene families were clearly identifiable, it is difficult to make effective generalizations about the overall gene family content of these classes. These groups of plasmids were not caused by an artificial grouping of plasmids of similar sizes (as summarized by the third column along that tree). Most of these groups clearly mixed plasmids found in prokaryotic hosts from different phyla (see the majority of the top clusters). A few other groups were consistent with the prokaryotic taxonomy of their hosts, for instance among the lowermost clusters in Supplementary Figure 6 (identified by a boxed number on their parent branch): the basal spirochaetes (1) correspond to (mainly circular) plasmids of *Borrelia*, followed by one group (2) exclusively composed of plasmids of Gammaproteobacteria, another spirochaetes group (3), this one mainly composed of linear plasmids), and a group of plasmids (4) found in firmicutes; also note the small but very conspicuous group of *Chlamydiae* plasmids (5). Generally plasmid mobility was also heterogeneous across groups of genomes sharing exclusive gene families. These features indicate that plasmids from different hosts and with different mobility can share exclusive gene families.

Supplementary tables 1-4 present the analysis of the composition of bipartite graphs at different % ID in terms of connected components (CC) and twins. We will refer to these collectively as “groups”.

Suppl. Tables 1 and 2 summarize information about the genome content of CCs and twins whose support satisfies certain constraints.

We distinguish mainly between heterogeneous and homogeneous groups, in terms of type (that is, cellular, viral and plasmidic), of phylum and of class. As a general rule, we say that a group is *homogeneous* with respect to a certain classification if all the genomes contained in that group are in the same category.

Under *type*, we consider the three categories "Prokaryote", "Virus" and "Plasmid". We distinguish here further between "prokaryote only" and "MGE only" (and in parenthesis some additional sub-cases).

Under *phylum* and *class*, we do not include the MGEs, so that groups that do not contain any prokaryote are excluded from these counts. A group will be counted as *homogeneous* if all prokaryotes contained are in the same phylum/class, even when MGEs are present in the same group: in parentheses, we specify how many such groups actually contain only one prokaryote (and are thus counted as homogeneous, yet trivially so). Accordingly, the percentages shown are relative to the number of groups that contain at least one prokaryote ("Cont. prok'").

These constraints are particularly weak or awkward in the case of the giant connected component (gCC), so we mark with an asterisk (*) whenever the count includes the gCC, and we reported the size of the gCC for each threshold.

Suppl. Table 1. Connected components in the genome-gene families bipartite graphs for cellular and MGEs at different stringency thresholds.

%ID	Number of connected components										
	All	Nodes in gCC	Prok only (a=archaea, b=bacteria)	MGE only v=virus only p=plasmid only m=mixed	Heterog. Type	Homog. Type	Cont. prok.	Heterog. Prok. phylum	Homog. Prok. phylum (only 1 prok)	Heterog. Prok. class	Homog. Prok. class (only 1 prok)
30	156	86498 (97.31%)	0	155 (130v, 25p, 0m)	1* (0,6 %)	155	1*	1* (100%)	0 (0)	1*	0 (0)
40	208	115978 (97.34%)	0	207 (165v, 41p, 1m)	2* (0,48 %)	206	1*	1* (100%)	0 (0)	1*	0 (0)
50	287	137328 (97.23%)	0	286 (226v, 59p, 1m)	2* (0,7 %)	285	1*	1* (100%)	0 (0) 0 % (0%)	1* (100%)	0 (0) 0 % (0%)
60	375	139139 (96.11%)	0	374 (299v, 73p, 2m)	3* (0,8 %)	372	1*	1* (100%)	0 (0) 0 % (0%)	1* (100%)	0 (0) 0 % (0%)
70	439	122491 (92.40%)	0	438 (340v, 95p, 3m)	4* (0,91 %)	435	1*	1* (100%)	0 (0) 0 % (0%)	1* (100%)	0 (0) 0 % (0%)
80	454	96358 (85.29%)	1 (1a)	452 (326v, 120p, 6m)	7* (1,54 %)	447	2*	1* (50 %)	1 (0) 50 % (0%)	2* (100%)	0 (0) 0 % (0%)
90	488	32636 (36.12%)	8 (3b 5a)	460 (297v, 155p, 8m)	28* (5,74 %)	460	28*	6* (21,43%)	22 (8) 78,51 % (28,57%)	8* (28,57%)	20 (8) 71,43 % (28,57%)
95	522	23071 (29.80%)	21 (6b 15a)	470 (275v, 188p, 7m)	38* (7,28 %)	484	52*	1* (1,92%)	51 (18) 98,08 % (32,62%)	6* (11,54%)	46 (18) 88,46 % (34,62%)

Suppl. Table 2. BTs in the genome-gene families bipartite graphs for cellular and MGEs at different stringency thresholds.

%ID	Number of twins								
	All	Prok only (a=archaea, b=bacteria, m=mixed)	MGE only v=virus only p=plasmid only m=mixed	Heterog. Type	Homog. Type	Heterog. Prok. phylum	Homog. Prok. phylum (only 1 prok)	Heterog. Prok. class	Homog. Prok. class (only 1 prok)
30	41646	15982 (38,38%) (4415a, 9977b, 1590m)	12657 (30,39%) (5351v, 6776p, 530m)	13537 (32,5%)	28109	12864 (30,89%)	16125 (38,72%) (3639 :8,74 %)	17059 (40,96%)	11930 (28,65 %)
40	51287	21379 (41,69%) (5840a, 13612b, 1927m)	13886 (27,8%) (5408v, 7992p, 486m)	16508 (32,18%)	34779	15445 (30,11%)	21956 (42,81%) (4376 : 5,53 %)	21260 (41,45%)	16141 (31,47%)
50	51101	22138 (43,32%) (6458a, 14263b, 1417m)	13459 (26,33%) (5129v, 8831p, 375m)	15003 (29,36%)	36098	12088 (23,66%)	24678 (48,29%) (4532 : 8,87 %)	18308 (35,82%)	18458 (36,12%)
60	41647	17122 (41,11%) (5694a, 10745b, 683m)	14208 (34,11%) (4692v, 9239p, 277m)	10594 (25,44 %)	31053	6059 (14,55%)	21380 (51,16%) (3878 : 9,31 %)	10959 (26,31%)	16480 (39,58%)
70	29052	9840 (33,87%) (4002a, 5661b, 177m)	13459 (46,33%) (4194v, 9054p 211m)	5964 (20,53%)	23088	1757 (6,05%)	13836 (47,62%) (2803 : 9,65 %)	4379 (15,07%)	11214 (38,60%)
80	19296	4320 (22,39%) (2303a, 1993b, 24m)	12314 (63,82%) (3700v, 8464p, 150m)	2812 (14,57%)	16484	259 (1,34%)	6723 (34,84%) (1690 : 8,76 %)	1138 (5,90%)	5844 (30,29%)
90	13391	1602 (11,96%) (1093a, 508b, 1m)	10826 (80,85%) (3221v, 7498p, 107m)	1070 (7,99%)	12321	20 (0,15%)	2545 (19,01%) (760 : 5,68 %)	238 (1,78%)	2327 (17,38%)
95	11288	870 (7,70%)(695a, 175b, 0m)	9767 (86,53%)(2884v, 6903p, 80m)	631 (5,59%)	10657	4 (0,035%)	1417 (12,55%) (479 : 4,24 %)	80 (0,71%)	1341 (11,88%)

In the second group of tables, we present the results on the groups that contain transposases. Both types of groups are said to *contain a transposase* if there is a node corresponding to a gene family at the given % ID containing a gene annotated as a transposase. If the group contains *another* gene family that is *not* annotated as a transposase, we label the group as

« hitch-hiking transposase ». Similarly as above, we have marked with an asterisk whenever the count includes the gCC. The homogeneity is defined in these groups as in the previous tables. The percentages are given relatively to the groups with same homogeneity requirements (column *All*).

Suppl. Table 3. Transposases in the connected components of the genome-gene families bipartite graphs for cellular and MGEs at different stringency thresholds.

%ID	Connected Components											
	Heterogeneous Type			Homogeneous Type			Heterogeneous phylum			Homogeneous phylum		
	All	With transposase	Hitch-hiking transposase	All	With transposase	Hitch-hiking transposase	All	With transposase	Hitch-hiking transposase	All	With transposase	Hitch-hiking transposase
30	1*	1* (100.00%)	1* (100.00%)	155	0 (0.00%)	0 (0.00%)	1*	1* (100.00%)	1* (100.00%)	0	0 (0.00%)	0 (0.00%)
40	2*	1* (50.00%)	1* (50.00%)	206	1 (0.49%)	0 (0.00%)	1*	1* (100.00%)	1* (100.00%)	0	0 (0.00%)	0 (0.00%)
50	2*	1* (50.00%)	1* (50.00%)	285	0 (0.00%)	0 (0.00%)	1*	1* (100.00%)	1* (100.00%)	0	0 (0.00%)	0 (0.00%)
60	3*	1* (33.33%)	1* (33.33%)	372	2 (0.54%)	1 (0.27%)	1*	1* (100.00%)	1* (100.00%)	0	0 (0.00%)	0 (0.00%)
70	4*	1* (25.00%)	1* (25.00%)	435	3 (0.69%)	1 (0.23%)	1*	1* (100.00%)	1* (100.00%)	0	0 (0.00%)	0 (0.00%)
80	7*	1* (14.29%)	1* (14.29%)	447	4 (0.89%)	3 (0.67%)	1*	1* (100.00%)	1* (100.00%)	1	0 (0.00%)	0 (0.00%)
90	28*	6* (21.43%)	2* (7.14%)	460	14 (3.04%)	10 (2.17%)	6*	2* (33.33%)	1* (16.67%)	22	4 (18.18%)	1 (4.55%)
95	38*	12* (31.58%)	6* (15.79%)	484	26 (5.37%)	20 (4.13%)	1*	1* (100.00%)	1* (100.00%)	51	10 (19.61%)	4 (7.84%)

* : includes gCC

Suppl. Table 4. Transposases in the BTs of the genome-gene families bipartite graphs for cellular and MGEs at different stringency thresholds.

%ID	Twins											
	Heterogeneous Type			Homogeneous Type			Heterogeneous phylum			Homogeneous phylum		
	All	With transposase	Hitch-hiking transposase	All	With transposase	Hitch-hiking transposase	All	With transposase	Hitch-hiking transposase	All	With transposase	Hitch-hiking transposase
30	13537	193 (1.43%)	17 (0.13%)	28109	267 (0.95%)	49 (0.17%)	12864	113 (0.88%)	14 (0.11%)	16125	173 (1.07%)	23 (0.14%)
40	16508	356 (2.16%)	20 (0.12%)	34779	434 (1.25%)	106 (0.30%)	15445	125 (0.81%)	6 (0.04%)	21956	378 (1.72%)	71 (0.32%)
50	15003	433 (2.89%)	32 (0.21%)	36098	658 (1.82%)	161 (0.45%)	12088	106 (0.88%)	7 (0.06%)	24678	502 (2.03%)	112 (0.45%)
60	10594	461 (4.35%)	50 (0.47%)	31053	855 (2.75%)	218 (0.70%)	6059	38 (0.63%)	1 (0.02%)	21380	594 (2.78%)	139 (0.65%)
70	5964	375 (6.29%)	53 (0.89%)	23088	949 (4.11%)	240 (1.04%)	1757	14 (0.80%)	0 (0.00%)	13836	491 (3.55%)	135 (0.98%)
80	2812	291 (10.35%)	55 (1.96%)	16484	943 (5.72%)	239 (1.45%)	259	3 (1.16%)	0 (0.00%)	6723	379 (5.64%)	109 (1.62%)
90	1070	192 (17.94%)	37 (3.46%)	12321	888 (7.21%)	277 (2.25%)	20	2 (10.00%)	0 (0.00%)	2545	256 (10.06%)	83 (3.26%)
95	631	150 (23.77%)	28 (4.44%)	10657	827 (7.76%)	278 (2.61%)	4	1 (25.00%)	0 (0.00%)	1417	187 (13.20%)	59 (4.16%)

Suppl. Table 5: Mean externalization rates for cellular genomes at different stringencies.

% ID	Bacteria	Haloarchaea	Other Archaea
30	43,3	46,1	27,4
40	26,1	32,7	11,6
50	12,7	21,9	3,3
60	5,2	13,7	1,0

Résumé

L'évolution des organismes, des génomes et des gènes n'est pas strictement arborescente; les symbioses, les transferts horizontaux de gènes ou encore la fusion de gènes créent des objets composites formés de parties dont les histoires évolutives sont différentes. Ces processus non arborescents sont appelés introgressifs et ont un impact non négligeable en évolution. Ils sont à l'origine de transitions évolutives majeures comme l'émergence des eucaryotes, des eucaryotes photosynthétiques ou encore de nombreux groupes d'Archaea. Dans le cas des eucaryotes, l'association et la stabilisation d'une Archaea et d'une alpha-protéobactérie a permis l'émergence d'un nouveau groupe d'organismes composites aux propriétés émergentes. L'acquisition de la photosynthèse chez les eucaryotes s'est faite *via* l'endosymbiose d'une cyanobactérie et, bien que débattue, l'apparition des grands groupes d'Archaea semble être concomitante avec l'acquisition de nombreux gènes d'origine bactérienne. Ces superorganismes ont la particularité d'avoir des génomes composés de gènes de différents partenaires symbiotiques. L'objectif de mon travail de thèse a constitué à étudier l'aspect introgressif de l'évolution par des méthodes de réseaux de similarité de séquence et des méthodes phylogénétiques. Je me suis particulièrement focalisé sur la détection de nouveaux gènes chimériques nommés gènes symbiogénétiques (S-gènes) car composés de parties originaires des différents partenaires symbiotiques. De tels gènes existent dans les génomes et plusieurs règles d'association ont pu être mises en évidence. Plus généralement, la présence de S-gènes étend la notion de mosaïcisme génomique au niveau infra-génique.

Abstract

Using network-based methods to analyze introgressive events in evolution.

Evolution of organisms, genomes and genes does not strictly follow a tree-like process; symbiosis, horizontal gene transfers and gene fusions build high level composite objects with components of phylogenetically distinct origins. Such processes have been called introgressive events and are significant in evolution. They are involved in some major evolutionary transitions such as eukaryogenesis, photosynthesis acquisition in eukaryotes and the origins of major archaeal clades. Eukaryogenesis would have involved (at least) two kinds of partners: an archaeon and an alpha-proteobacterium. Photosynthetic eukaryotes arose from the integration of a cyanobacterium into a eukaryotic cell and recent findings suggested that most archaeal lineages emerged after massive acquisitions of bacterial genes. These composite lineages carry highly chimeric genomes where genes from symbiotic partners co-localize into the same genome. During my PhD thesis, I used sequence similarity networks and phylogenetic methods in order to study reticulate evolution. My research specifically focused on a previously hidden component of composite genomes: symbiogenetic genes (S genes). These chimeric genes are found in genetic mergers, and originate from the association of genes of symbiotic partners. Some association rules have been discovered. In a broad perspective, the discovery of S-genes extends the concept of genome chimerism to the within-gene level.