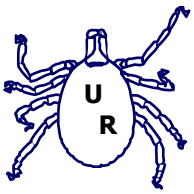


UNIVERSITE DE LA MEDITERRANEE-AIX-MARSEILLE II
FACULTE DE MEDECINE - LA TIMONE
ECOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTE

THESE DE DOCTORAT

Présentée par

Kalliopi GEORGIADES

En vue de l'obtention du grade de Docteur de l'Université de la Méditerranée
Spécialité : Maladies Transmissibles et Pathologies Tropicales

Phylogénomique des bactéries pathogènes

Soutenue le 08 Septembre 2011

COMPOSITION DU JURY

M. le Professeur Jean-Louis MEGE	**Président du Jury**
M. le Professeur Didier RAOULT	**Directeur de Thèse**
M. le Docteur Pierre PONTAROTTI	**Co-directeur de Thèse**
M. le Docteur Gilbert GREUB	**Rapporteur**
M. le Professeur Xavier NASSIF	**Rapporteur**
M. le Docteur Etienne DANCHIN	**Examinateur**

Unité de Recherche sur les Maladies Infectieuses et Tropicales Emergentes,
UMR CNRS 6236

Avant Propos :

Le format de présentation de cette thèse correspond à une recommandation de la spécialité Maladies Infectieuses et Microbiologie, à l'intérieur du Master de Sciences de la Vie et de la Santé qui dépend de l'Ecole Doctorale des Sciences de la Vie de Marseille. Le candidat est amené à respecter des règles qui lui sont imposées et qui comportent un format de thèse utilisé dans le Nord de l'Europe permettant un meilleur rangement que les thèses traditionnelles. Par ailleurs, la partie introduction et bibliographie est remplacée par une revue envoyée dans un journal afin de permettre une évaluation extérieure de la qualité de la revue et de permettre à l'étudiant de le commencer le plus tôt possible une bibliographie exhaustive sur le domaine de cette thèse. Par ailleurs, la thèse est présentée sur article publié, accepté ou soumis associé d'un bref commentaire donnant le sens général du travail. Cette forme de présentation a paru plus en adéquation avec les exigences de la compétition internationale et permet de se concentrer sur des travaux qui bénéficieront d'une diffusion internationale.

Professeur Didier RAOULT

Sommaire

REMERCIEMENTS …….

ABSTRACT

The virulence of pathogenic bacteria has been attributed to virulence factors and pathogenic bacteria are considered to have more genes compared to bacteria that do not cause disease.

According to the first genomic studies, removing a certain number of genes from pathogenic bacteria impairs their capacity to infect hosts. However, more recent studies have demonstrated that the specialization of bacteria in eukaryotic cells is associated with massive gene loss, especially for allopatric endosymbionts that have been isolated for a long time in an intracellular niche. Indeed, bacteria living in sympatry often have bigger genomes and exhibit greater resistance and plasticity and constitute species complexes rather than true species. Specialists, including specific pathogenic bacteria, escape these bacterial complexes and colonize a niche; thereby gaining a species name. Their specialization allows them to adopt allopatric lifestyle and experience reductive genome evolution.

These observations led us to design a study to quantify the rate of gene losses during the evolution of free-living bacteria to intracellular specialists. Our objective was to verify that what characterizes the evolution of intracellular bacteria is genomic reduction, taking under consideration all possible gene gain events.

Furthermore, in another neutral study comparing the 12 most dangerous pandemic bacteria to Humans to their closest non-epidemic species, we wished to identify any genomic specificities associated to the virulent capacity of pathogenic bacteria and demonstrate that, besides toxins and surprisingly, toxin-antitoxin modules, pathogenic bacteria are not characterized by more virulence factors, but rather by a loss of regulatory genes. Finally, virulent bacteria exhibit a genomic repertoire in which absent genes are as important as present ones.

Keywords: pathogenic bacteria, specialization, genomic reduction, virulence factors

RESUME

La pathogénicité des bactéries a toujours été attribuée à des facteurs de virulence et les bactéries pathogènes sont considérées comme étant mieux armées, comparé à des bactéries ne provoquant pas de maladies. Selon les premières études génomiques, le fait de supprimer un certain nombre de gènes des bactéries pathogènes, limiterait leur capacité à infecter leurs hôtes. Au contraire, des études de génomique comparatives récentes, démontrent que la spécialisation des bactéries dans les cellules eucaryotes est associée à une perte de gènes massive, en particulier pour les endosymbiontes allopatriques qui sont isolés depuis longtemps dans une niche intracellulaire. En effet, les bactéries sympatriques, extracellulaires, ont souvent des génomes plus grands et présentent une résistance et une plasticité plus importante. Ces bactéries constituent, de fait, plutôt des complexes d'espèces que de vraies espèces. Certaines bactéries spécialistes, comme les bactéries pathogènes, arrivent à s'échapper de ces complexes et à coloniser une niche, bénéficiant alors d'un nom d'espèce. Leur spécialisation leur permet de devenir allopatriques et leurs pertes de gènes favorisent une évolution réductive. Ces observations nous ont conduits à réaliser une étude afin de quantifier le taux de perte de gènes lors de l'évolution de ces bactéries extracellulaires vers celle de bactéries spécialistes intracellulaires. Notre objectif était de vérifier que ce qui caractérise l'évolution des bactéries intracellulaires est bien la réduction génomique, en prenant en compte tous les événements possibles de gains de gènes.

Par ailleurs, dans une étude neutre comparant les 12 espèces pandémiques les plus dangereuses pour l'homme avec les espèces non-épidémiques les plus proches, nous avons voulu identifier des spécificités génomiques associées à la capacité virulente de bactéries pathogènes et démontrer que, à part les toxines et les modules toxine-antitoxine, ce qui caractérise ces espèces ce ne sont pas les facteurs de virulence, mais la perte des gènes de régulation. Au final, les bactéries pathogènes ont un répertoire virulent dans lequel les gènes absents sont aussi importants que les gènes présents.

Mots clés : bactéries pathogènes, spécialisation, réduction génomique, facteurs de virulence

INTRODUCTION

Les définitions actuelles des espèces bactériennes, et en particulier des espèces pathogènes, sont limitées à cause de l'utilisation d'outils génétiques restrictifs. Par exemple, les séquences de 16S rDNA (gènes codant pour les ARN ribosomiques) ont été largement utilisées comme marqueurs pour les analyses phylogénétiques de description et de classification de nouvelles espèces (1-3). Par contre, l'utilisation de la séquence ribosomique entraîne souvent des erreurs dans la définition des espèces et ne garantit pas une délinéarisation des espèces bactériennes (1, 4). En effet, la définition actuelle des bactéries utilise une divergence de 16S rDNA de 1,3% (5). Il a été proposé qu'une divergence de 1-2% de 16S rDNA corresponde à 50 millions d'années de divergence (6-7). Pourtant, les bactéries spécialisées dans des hôtes mammifères, qui ont évolué dans une niche spécifique, ne sont pas définies comme des espèces différentes en utilisant ce critère. La première bactérie pathogène spécialisée chez l'homme, *Mycobacterium tuberculosis*, est apparue il y a seulement 20 000 ans (8), beaucoup plus tard que *Homo sapiens*, qui est apparu il a environ 250 000 à 400 000 ans (9). Donc, la définition des bactéries ne peut pas être basée sur le pourcentage de divergence du 16S rDNA, parce que les bactéries ayant une divergence inférieure à 1,3% correspondent à des complexes bactériens et non pas à des espèces spécifiques (10). Nous considérons qu'un complexe bactérien est composé de bactéries

« généralistes » ayant un mode de vie sympatrique, un niveau élevé de transferts horizontaux, un grand génome et un nombre significatif d'opérons ribosomaux (11). Une espèce *bona fide* résulte de l'isolement d'une bactérie spécialiste dans une niche spécifique. Cette espèce aura alors un mode de vie allopatrique, un génome plus petit et un nombre d'opérons ribosomaux réduit (12).

Des études de génomique comparatives plus récentes ont démontré que les bactéries sont capables d'obtenir des gènes à partir d'organismes distants. L'échange de gènes chez les bactéries se produit par des mécanismes de transferts horizontaux par l'intermédiaire soit des virus, soit des gènes égoïstes, soit de mécanismes de transformation et de conjugaison. La capacité d'échange de gènes dépend de l'environnement naturel des bactéries (12). Les bactéries sympatriques, comme les bactéries aquatiques, peuvent échanger facilement leurs gènes en interagissant avec des bactéries de phylum différents. Ces bactéries ont souvent des génomes plus grands, plus de gènes, plus d'opérons ribosomaux, de meilleures capacités métaboliques et une résistance plus importante aux agents physicochimiques (11). Par contre, les espèces allopatriques, (qui sont isolées génétiquement), n'ont pas la possibilité d'échanger des gènes ni de redevenir sympatriques. Les bactéries sympatriques créent souvent de nouveaux répertoires qui sont le réservoir de futures bactéries spécialistes et dont la spécialisation conduira à l'allopatrie et

à une absence d'acquisition de gènes par transferts horizontaux et un déséquilibre de la balance gain/perte de gènes. Le résultat est une réduction génomique considérable du fait de la perte de gènes (12).

Des études sur la pathogénicité des bactéries ont démontré qu'enlever un certain nombre de gènes des bactéries pathogènes, éliminait leur capacité à infecter leurs hôtes. La suppression d'un certain nombre de gènes diminuait la « fitness » de ces espèces dans l'écosystème dans lequel elles ont évolué et sont devenues spécialisées. Chacun des gènes enlevés a donc été appelé « facteur de virulence » (14). Donc, il est apparu que les bactéries dangereuses pour l'homme sont armées et que les bactéries inoffensives manquent de « facteurs de virulence » (13,15). Mais des études comparatives récentes ont prouvé que la spécialisation de bactéries dans les cellules eucaryotes est associée à une perte massive de gènes (12,16) et même à une perte de « facteurs de virulence » (11). La perte de gènes, associée à l'isolement génétique et à la multiplication dans un nouvel écosystème, est à la base du principe de vaccination mis en place par Pasteur (17). Alors qu'il étudiait l'agent du cholera, il a observé que la bactérie en cause, *Pasteurella multocida,* cultivée plusieurs fois dans un milieu axénique, perdait progressivement sa virulence pour les poulets sans pour autant perdre sa capacité à protéger ceux-ci contre des infections futures. L'inoculation d'une bactérie de virulence atténuée a constitué la première vaccination, suite à une

expérimentation de laboratoire (17-18). Une culture comprenant des passages multiples, imite la spécialisation et l'isolement allopatrique de bactéries qui conduit à la spéciation et à la perte de gènes plutôt qu'au gain de gènes. Ces changements font que la bactérie devient incapable d'infecter son hôte précédent. Par conséquent, les mutants sélectionnés sont plus « fit » (virulents) dans leur nouvelle niche qui est le milieu axénique.

Un des meilleurs exemples de réduction génomique est celui de *Rickettsia prowazekii*, l'espèce la plus virulente du genre *Rickettsia*. Le séquençage de son génome n'a pas trouvé de gènes correspondant à des facteurs de virulence connus (19). La mobilité intracellulaire est considérée comme un facteur de virulence pour *Shigella* (20) et *Listeria monocytogenes* (21-22). Or, *R. prowazekii* ne possède pas ce facteur et n'est pas mobile dans le cytoplasme (23-24). En plus, *R. typhi*, une espèce moins virulente que *R. prowazekii*, est mobile grâce à une protéine, Sca2 (7) qui pourrait être considérée comme un facteur de virulence. Malgré tout, cette protéine est absente chez *R. prowazekii* (24-25). Cette observation suggère que la mobilité n'est pas un facteur de virulence en soi, mais peut être trouvée dans quelques pathogènes comme faisant partie d'un répertoire virulent. Un autre exemple de réduction génomique est celui des Mycobactéries. En effet, le bacille de la lèpre a perdu 2 000 gènes après sa divergence depuis le dernier ancêtre commun mycobactérien (26-31). Enfin un autre pathogène extraordinaire est

Shigella dysenteriae, un clone qui sort du complexe *d'Escherichia coli*, a été classifié comme un genre à cause de son rôle dans la dysenterie humaine (32). Les *Shigella* diffèrent d'*E. coli* par leurs traits phénotypiques pauvres, comme une immobilité extracellulaire et une incapacité à fermenter le lactose et autres sucres (33-34). *Shigella* et *E. coli* ont toujours été considérées comme étant très proches et elles ont même été classées comme une seule espèce (34). Néanmoins, les *shigelles* provoquent des maladies chez l'homme, alors que les souches d'*E. coli*, à l'exception de quelques clones pathogènes, sont des commensaux humains (35). Il a été démontré que la perte de l'enzyme lysine decarboxylase (LDC) induit une augmentation de la virulence d'une souche entero-invasive d'*E. coli* et de *S. dysenteriae* (35). *S. dysenteriae* et les souches shigatoxigéniques d'*E. coli* n'ont pas plus de facteurs de virulence que des bactéries non pathogènes phylogénétiquement proches, excepté la toxine Shiga (36).

Il semblerait que les bactéries spécialistes s'échappent régulièrement de complexes bactériens, (qui constituent des « melting pots »), colonisent de nouvelles niches et acquièrent un niveau suffisant de spécialisation leur donnant une légitimité à être considérées comme une nouvelle espèce. Leur spécialisation va conduire à l'allopatrie et au déséquilibre de la balance entre le gain et la perte de gènes, en favorisant la perte, et finalement à une réduction génomique stochastique.

Suite à ces observations, il semblait nécessaire de réaliser une étude de génomique comparative neutre pour examiner tous les paramètres précédemment décrits comme jouant un rôle dans la pathogenicité. En neutralisant les biais d'observation, nous avons voulu identifier des spécificités génomiques associées à la capacité virulente de bactéries pathogènes. Nous avons aussi déterminé si la virulence est dictée par des règles ou si elle résulte d'histoires évolutives différentes. Pour ne pas influencer l'analyse nous avons utilisé les 12 espèces pandémiques les plus dangereuses pour l'homme et nous les avons comparées aux espèces les plus proches phylogénétiquement, non pathogènes ou au moins non épidémiques, dont le génome était disponible.

Nous avons aussi voulu étudier les mécanismes participant à l'évolution d'un mode de vie extracellulaire vers un mode de vie intracellulaire, en prenant en compte tous les événements de gain et de perte de gènes. Le meilleur modèle pour une telle étude est l'ordre des *Rickettsiales* dont la réduction génomique est toujours en cours, comme le montre l'exemple de *R. prowazekii* (7,19). Nous avons utilisé les Rickettsies et les *Orientia* et nous avons identifié des événements de transferts horizontaux de et vers plusieurs organismes distants. Nous avons aussi pu reconstruire les « gene-sets » ancestraux des ancêtres des *Rickettsiales*, des *Rickettsiaceae* et des *Rickettsies*,

et nous avons pu prédire le nombre de gènes perdus pendant le passage d’un mode de vie extracellulaire à un mode de vie intracellulaire.

Revue

Defining pathogenic bacterial species in the genomic era

Kalliopi Georgiades[1] and Didier Raoult[1*]

[1] Unité des Rickettsies, URMITE-CNRS UMR 6236 IRD 198 IFR 48, Faculté de Médecine, Université de la Méditerranée, 27, Bd Jean Moulin, 13005, Marseille, France

[*] Corresponding author: Email: didier.raoult@gmail.com

Avant-propos de l'article 1

Genomes of the most dangerous epidemic bacteria have a virulence repertoire characterized by fewer genes but more Toxin-Antitoxin modules

Kalliopi Georgiades[1] and Didier Raoult[1*]

[1] Unité des Rickettsies, URMITE-CNRS UMR 6236 IRD 198 IFR 48, Faculté de Médecine, Université de la Méditerranée, 27, Bd Jean Moulin, 13005, Marseille, France

[*] Corresponding author: Email: didier.raoult@gmail.com

La virulence des bactéries pathogènes à été attribuée aux facteurs de virulence et ces bactéries sont considérées comme mieux armées que celles qui ne provoquent pas de maladies (37). Par contre, des études comparatives plus récentes ont révélé que les génomes de bactéries telles que les Rickettsies (27, 38-39) ou les Mycobactéries, sont réduits (7, 12, 27, 40). Nous avons réalisé une étude génomique comparative basée sur une approche neutre pour identifier des spécificités génomiques associées à la capacité de virulence des bactéries pathogènes. Nous avons comparé les 12 espèces bactériennes pandémiques les plus dangereuses pour l'homme (« bad bugs ») à leurs espèces les plus proches phylogénétiquement mais non-épidémiques (« témoins »). Notre analyse a révélé plusieurs caractéristiques significativement différentes dans les « bad bugs », dont un génome plus petit qui résulte probablement d'une dégradation du système de recombinaison et de réparation. Dix catégories fonctionnelles de COGs contiennent significativement moins de gènes dans les « bad bugs » qui manquent de gènes de transcription, de mécanismes de transduction de signal, de motilité cellulaire, de production d'énergie et de fonctions de métabolisme et de régulation. Cinq « bad bugs » présentent un plus grand nombre de queues poly(A) comparé aux « témoins » alors qu'un nombre élevé de queues poly(A) est fortement corrélé à un GC% faible. Par ailleurs, les « bad bugs » possèdent significativement plus de modules Toxine-Antitoxine (TA). En plus, une Analyse des Composantes Principales a montré que ce qui semble caractériser les « bad bugs » est un petit génome et un nombre élevé de modules TA.

Nos résultats montrent que la capacité pathogénique n'est pas le résultat de « facteurs de virulence » mais plutôt la conséquence d'un répertoire génomique virulent qui résulte de génomes réduits. Les modules TA peuvent faire partie d'un tel répertoire virulent, mais ils ont probablement évolué comme des éléments égoïstes de façon indépendante.

Avant-propos de l'article 2

Gene gain and loss events in *Rickettsia* and *Orientia* species

Kalliopi Georgiades[1,2], Vicky Merhej[1], Khalid El Karkouri[1], Didier Raoult[1] and Pierre Pontarotti[2*]

[1] Unité des Rickettsies, URMITE-CNRS UMR 6236 IRD 198 IFR 48, Faculté de Médecine, Université de la Méditerranée, 27, Bd Jean Moulin, 13005, Marseille, France

[2] Evolutionary biology and Modeling, LATP UMR CNRS 6632 FR 3098 IFR 48, University of Provence, Marseilles, France

[*] Corresponding author: Email: pierre.pontarotti@univ-provence.fr

La réduction de génomes est un processus continu dans tous les membres de l'ordre de *Rickettsiales* ce qui fait que ces bactéries constituent un modèle idéal pour étudier l'évolution réductive à travers les variations de taille de génomes et de contenu génomique inter-espèces (7,19). Dans cette étude nous avons évalué le degré dans lequel les pertes de gènes ont façonné le contenu génomique de quelques *Rickettsiales*. Notre arbre phylogénomique, basé sur la présence/absence de gènes, présente une topologie différente de celle de l'arbre phylogénétique, basé sur une concaténation des séquences du « core-génome », ce qui suggère que les répertoires génomiques étudiés ont des histoires évolutives différentes qui résultent de différents événements de perte ou de gain de gènes. Nous avons identifié 3 événements de transferts horizontaux vers les *Orientia* et 6 vers les Rickettsies. Nous avons aussi identifié 3 autres événements de transferts horizontaux de Rickettsies et d'*Orientia* vers d'autres organismes. Par ailleurs, nous avons identifié 17 gènes de Rickettsies qui sont probablement le résultat d'une création *de novo*; 2 de ces gènes semblent être fonctionnels. En se basant sur ces résultats nous avons reconstruit les répertoires génomiques des ancêtres des *Rickettsiales* (proto-*Rickettsiales*) et des *Rickettsiaceae* (proto-*Rickettsiaceae*). Nous avons pu démontrer que pendant le changement de mode de vie des *Rickettsiaceae* d'extracellulaire en intracellulaire 2 135 gènes ont été perdus. Cette étude nous a permis de suivre les événements de perte et de gain de gènes qui se sont produits pendant l'évolution de ces espèces bactériennes libres en espèces bactériennes intracellulaires. Nous avons prouvé que le mécanisme principal d'évolution et de spécialisation des bactéries intracellulaires strictes est la perte de gènes. Pourtant, malgré un mode de vie intracellulaire, quelques événements de transferts horizontaux entre les *Rickettsiales* et autres organismes procaryotes, viraux et eucaryotes ont eu lieu et ont participé à la formation des répertoires génomiques de ces espèces.

Ma contribution à ce travail s’élève à 75%.

CONCLUSION

De nombreuses études ont démontré que la perte de gènes est une force majeure d'évolution des bactéries intracellulaires (41-45). Notre étude apporte un complément important à d'autres études sur l'évolution des alphaproteobactéries (46). En effet, nous avons réussi à quantifier le nombre de gènes perdus par les *Rickettsiales* pendant leur évolution vers un mode de vie intracellulaire (2 135 gènes), et décrit le taux de pertes de gènes dans 11 espèces de Rickettsies. *Rickettsia bellii* par exemple, a conservé tous les gènes des « proto-Rickettsies », alors que *R. prowazekii* et *R. typhi*, qui présentent un éventail d'hôtes limité, ont perdu le plus grand nombre de gènes ancestraux. En plus, 9 cas de transferts horizontaux vers les Rickettsies et les *Orientia* ont été identifiés. La plupart des gains ont eu lieu avant la séparation du groupe « Spotted Fever Group (SFG) » et du groupe « Typhus Group (TG) », alors que quelques gènes ont été gagnés indépendamment, par différentes espèces modernes de Rickettsies, après leur spéciation. Malgré un habitat intracellulaire et une évolution caractérisée par une perte massive de gènes, un certain nombre de transferts horizontaux a été identifié entre les *Rickettsiales* et différentes espèces distantes. En conclusion, les différences de taille de génomes observées au niveau des bactéries intracellulaires strictes (spécialisées) et des bactéries extracellulaires sont principalement le résultat d'une réduction génomique des intracellulaires strictes.

Effectivement, ce résultat était confirmé par notre étude sur les 12 bactéries pandémiques, spécialistes de l'homme (« bad bugs »). Une réduction significative de leur génome a été révélée par rapport aux espèces les plus proches non pathogènes, ou non épidémiques (« témoins »). Cette réduction est accompagnée par une diminution importante du nombre d'ORFs, ce qui montre que beaucoup de gènes disparaissent progressivement des génomes de « bad bugs ». Le mécanisme de pseudogénisation commence par un changement vers une composition plus élevée en nucléotides AT et par conséquence vers un excès

d'homopolymères (queues poly(A)). Cet excès de queues poly(A) conduit à une accumulation des codons stop, ce qui va créer des gènes coupés ou dégradés, qui seront ensuite éliminés *via* de grandes délétions (47). Néanmoins, des gènes coupés peuvent être restaurés et persister dans le génome. Le pourcentage codant est en effet plus petit dans les « bad bugs » qui en plus, possèdent un grand nombre de queues poly(A). Notre travail sur le système de réparation et de recombinaison a démontré que parmi les génomes de « bad bugs » étudiés, beaucoup de gènes de recombinaison et de réparation sont perdus. Ces espèces ont alors un système de réparation déficient qui les rend incapables de réparer des mutations et de surmonter les dégradations de gènes, les pseudogénisations et éventuellement les pertes de gènes. Les fonctions manquantes au niveau des « bad bugs » sont principalement les fonctions métaboliques, les fonctions de production d'énergie, de motilité cellulaire et de transcription.

Nous avons aussi observé que le nombre de gènes ou d'éléments considérés comme étant des « facteurs de virulence » est statistiquement plus élevé dans les « témoins » que dans les « bad bugs ». Enfin, les modules de Toxine-Antitoxine (TA) sont significativement plus nombreux dans les « bad bugs ». Les systèmes de TA ont été initialement identifiés comme des facteurs de stabilisation des plasmides et comme ayant un rôle dans la stabilisation des intégrons au niveau des chromosomes bactériens (48). Les modules TA sont des gènes égoïstes qui inhibent le détachement de l'opéron des organismes et ils sont décrits comme des molécules d'addiction (49-50). La possibilité d'éliminer seulement la toxine et pas l'antitoxine est limitée, due à leur positionnement l'une à côté de l'autre dans le génome. Toute tentative d'éliminer l'opéron conduit à la mort de la bactérie (51). Sous ces conditions, les molécules d'addiction sont sélectionnées non pas parce qu'elles sont indispensables, ou essentielles à l'organisme, mais parce que l'organisme ne peut pas s'en séparer (51). C'est intéressant de remarquer que les « bad bugs » n'ont pas seulement plus de modules TA, mais ont aussi des génomes plus petits que les « témoins ». Il est possible que les TA puissent jouer

un rôle dans la virulence de bactéries pathogènes, étant donné la pathogénicité induite après tentative de limiter leur traduction dans les bactéries les contenant.

Dans le cas de l'adaptation d'un mode de vie intracellulaire, la spéciation de bactéries est accompagnée par une perte massive de gènes qui peut conduire à une perte totale d'autonomie. La spéciation peut être définie comme la capacité réduite d'une bactérie à obtenir de nouveaux caractères, comme conséquence de son isolement dans un environnent donné. Ce phénomène est associé à une capacité limitée d'adaptation aux changements écologiques. Pour cette raison, tout changement significatif dans l'écosystème, ou tout changement de la niche peut provoquer la disparition de la bactérie (52). Contrairement à la situation prédite par la théorie de la Reine Rouge (Red Queen theory; 53-54), la plupart des pathogènes humains du XIX^{ème} siècle, comme ceux responsables du typhus, de la lèpre, de la typhoïde et de la dysenterie, sont retrouvés à un stade de déclin considérable qui résulte de leur faible capacité à s'adapter. C'est pour cela que les bactéries pathogènes sont probablement en train de disparaître mais seront sûrement remplacées par de nouveaux pathogènes spécialisés chez l'homme. Ces espèces vont émerger de complexes qui sont déjà en contact avec nous, surgissant de commensaux humains (muqueuses), d'animaux (agents de zoonoses), ou de l'environnement.

En conclusion, une espèce pathogène n'est pas caractérisée par ses « facteurs de virulence » mais par un répertoire virulent comprenant une réduction génomique considérable, résultant de sa spécialisation extrême à un environnement donné et stable. Dans le répertoire génomique virulent, les gènes absents sont aussi importants que les gènes présents. A l'exception de toxines et de modules TA qui ont un effet direct et mesurable, d'autres facteurs appelés « de virulence » sont en réalité des facteurs associés à la « fitness » dans un modèle expérimental testé.

PERSPECTIVES

Lors de mon travail de thèse sur la capacité virulente des bactéries, j'ai pu démontrer que les bactéries pathogènes ne contiennent pas plus de facteurs de virulence que les bactéries non-pathogènes, à l'exception des modules TA. Ces modules n'ont pas été considérés comme jouant un rôle dans la capacité pathogénique des bactéries avant, surtout parce qu'ils ne sont pas sécrétés dans les cellules eucaryotes (55-57). Au contraire, les toxines bactériennes ont été reconnues depuis 1888 comme les facteurs de virulence par excellence ayant un effet toxique direct et mesurable (58-59). A part les toxines et les modules TA, quelques enzymes de restriction semblent avoir une activité toxique, exprimée par des effets mutagènes sur l'ADN cellulaire (60-62). Leur rôle est de former un mécanisme de défense pour les bactéries contre l'ADN de bactériophages en le coupant au niveau de sites différents. Ils sont aussi considérés comme des éléments égoïstes (63). Il serait alors intéressant de confirmer si effectivement, les toxines sont, avec les modules TA, les seuls vrais facteurs de virulence bactériens et si les enzymes de restriction pourront aussi être considérés comme faisant partie d'un répertoire génomique virulent au même titre que les toxines et les modules TA.

Dans une vue classique, les toxines sont considérées comme des molécules provoquant une intoxication quand elles sont libérées par les bactéries dans les cellules eucaryotes. A cause de cette définition, la virulence de bactéries telles que la *Salmonella*, et *Yersinia*, qui ne secrètent pas de toxines protéiques, ne pouvait pas être expliquée. Or, nous savons aujourd'hui que ces bactéries peuvent intoxiquer leurs hôtes (64). Ces toxines, comme les modules TA, ont un relargage intracellulaire. Ce phénomène mérite d'être exploré pour voir si, ce ne sont pas plutôt les toxines intracellulaires qui sont plus liées à la capacité pathogénique de bactéries et non pas les extracellulaires, comme nous l'avons pensé jusqu'à maintenant.

Enfin, la plupart des toxines protéiques sont associées à des bactériophages ou à d'autres éléments mobiles, tels que les plasmides ou les transposons (65-66). Les toxines et les modules TA sont des gènes transférables et font partie du « mobilome ». Il faudrait donc essayer de voir quelles sont les origines des toxines et des modules TA, en prenant en compte des événements de transferts horizontaux probables, et voir aussi à quel moment ils ont fait partie du répertoire virulent. Est-ce que leur acquisition précède l'émergence de pathogènes?

Les modules TA et les toxines font partie d'un répertoire génomique, en l'occurrence pour les bactéries pathogènes, un répertoire génomique virulent. Un répertoire génomique, en général, est composé par des gènes stables (« core genome »), des gènes transférables et des gènes créés *de novo* (« orphans »). Les génomes ont donc une structure chimérique contenant des gènes de différentes origines. Il est donc impossible d'identifier un ancêtre commun pour tout le répertoire génomique d'un organisme. J'ai voulu alors mesurer le degré de mosaïsisme des génomes en prenant comme exemple les mitochondries. Nous considérons que les mitochondries sont apparues suite à un événement endosymbiotique entre un proto-*Rickettsiales* et une cellule proto-eucaryote (67-68). La plupart des analyses phylogénomiques placent les mitochondries parmi les Alpha-protéobactéries (69-71) et d'autres études ont montré que les espèces les plus proches des mitochondries font partie de l'ordre de *Rickettsiales* (72). Néanmoins, cette relation est contestée par une étude proposant que *Rhodospirillum rubrum* soit plus proche des mitochondries que les *Rickettsiales* (69). D'autres études plus récentes, ont associé les mitochondries aux *Rhizobiales* et aux *Rhodobacteriales* (71), alors qu'une étude a démontré que d'autres bactéries, en dehors de *Rickettsiales*, ont joué un rôle significatif dans la formation des mitochondries (73). Ainsi, les mitochondries constituent un bon exemple pour étudier la formation de la structure mosaïque des génomes et en particulier la mitochondrie de *Reclimonas americana* dont le

génome est probablement le plus proche du génome mitochondrial ancestral (74). Le but est de déterminer le degré de la participation d'un ancêtre commun lointain à la formation du génome mitochondrial et aussi le taux, les origines et le moment auquel d'autres éléments ont été ajoutés pour créer le répertoire chimérique de ces organelles. Bien sûr, la représentation d'un tel répertoire, provenant d'une variété d'ancêtres, ne pourra pas se faire avec un arbre Darwinien classique, mais sur un arbre généalogique, sur lequel nous pourrions suivre l'origine de chaque gène, ainsi que le temps de la divergence des ancêtres et le temps de l'acquisition de tous les éléments.

Des transferts horizontaux, initialement observés chez les bactéries, ont été identifiés dans tous les organismes vivants. Même le génome humain est une mosaïque de gènes d'origines eucaryotes, procaryotes et virales (75), et les virus géants ont aussi des génomes chimériques d'origines différentes (76). L'arbre généalogique constitue, actuellement, le seul modèle sur lequel nous pouvons intégrer les occurrences des origines multiples des répertoires génomiques des organismes. De nouveaux gènes et de nouvelles espèces se créent perpétuellement. Leurs répertoires génomiques d'origines variables leur permettront d'évoluer, si les conditions environnementales sont favorables et vont remplacer les espèces qui sont en train de régresser.

REFERENCES

1. Rossello-Mora R & Amann R (2001) The species concept for eukaryotes. *FEMS Microbiol Rev* **25**: 39-67.

2. Drancourt M, Berger P & Raoult D (2004) Systematic 16S rRNA gene sequencing of atypical clinical isolates identified 27 new bacterial species associated with humans. *J Clin Microbiol* **42**: 2197-2202.

3. Roux V, Drancourt M, Stein A, Riegel P & Raoult D (2004) Corynebacterium species isolated from bone and joint infection identified by 16S rRNA gene sequence analysis. *J Clin Microbiol* **42**: 2231-2233.

4. Fox GE, Wisotzkey JD & Jurtshuk P (1992) How close is close-16S ribosomal-RNA sequence identity may not be sufficient to guarantee species identity. *Int J Syst Bacteriol* **42**: 166-170.

5. Stackebrandt E & Ebers J (2006) Taxonomic parameters revisited: tarnished gold standards. *Microbiol Today* **33**: 152-155.

6. Ochman H, Elwyn S & Moran NA (1999) Calibrating bacterial evolution. *Proc Natl Acad Sci U S A* **96**: 12638-12643.

7. Ogata H, Renesto-Audiffren P, Fournier PE, Barbe V, Samson D *et al*. (2001) Mechanisms of evolution in *Rickettsia conorii* and *Rickettsia prowazekii*. *Science* **293**: 2093-2098.

8. Wirth T, Hilderbrand F, Allix-Béguec C, Wolbeling E, Kubica T *et al*. (2008) Origin, spread and demography of the *Mycobacterium tuberculosis* complex. *Plos Pathog* **4**: e1000160.

9. Feng DF, Cho G & Doolittle RF (1997) Determing divergence times with a protein clock: update and reevaluation. *Proc Natl Acad Sci U S A* **94**: 13028-13033.

10. Doolittle WF & Papke RT (2006) Genomics and the bacterial species problem. *Genome Biol* **7**: 116-123.

11. Audic S, Robert C, Campagna B, Parinello H, Claverie JM *et al*. (2007) Genome analysis of *Mycobacterium massiliensis* highlights the convergent evolution of water-living bacteria. *Plos Genet* **3**: e138.

12. Merhej V, Royer-Carenzi M, Pontarotti P & Raoult D (2009) Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. *Biol Direct* **4**: 13.

13. Ochman H, Lerat E & Daubin V (2005) Examining bacterial species under the specter of gene transfer and exchange. *Proc Natl Acad Sci USA* **102**: 6595-6599.

14. Dobrindt U, Hochhut B, Hentschel U & Hacker J (2004) Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol* **2**: 414-424.

15. Lawrence JG (1999) Gene transfer, speciation and the evolution of bacterial genomes. *Curr Opin Microbiol* **2**: 519-523.

16. Nierman WC, DeShazer D, Kim HS, Tettelin H, Nelson KE *et al*. (2004) Structural flexibility in the *Burkholderia mallei* genome. *Proc Natl Acad Sci U S A* **101**: 14246- 14251.

17. Chambeland C (1883) *Le charbon et la vaccination charbonneuse*. Paris: B. Tignol.

18. Walsh JJ (1913) *Louis Pasteur*: *Catholic Encyclopedia*. New York: Robert Appleton Company.

19. Bechah Y, Karkouri EK, Mediannikov O, Leroy Q, Pelletier N *et al*. (2010) Genomic, proteomic and transcriptomic analysis of virulent and avirulent *Rickettsia prowazekii* reveals its adaptive mutation capabilities. *Genome Res* **20**: 655-663.

20. Goldberg MB & Theriot JA (1995) *Shigella flexneri* surface protein IcsA is sufficient to direct actin-based motility. *Proc Natl Acad Sci U S A* **92**: 6572-6576.

21. Tilney LG & Portnoy DA (1989) Actin filaments and the growth movement and spread of the intracellular bacterial pathogen, *Listeria monocytogenes*. *J Cell Biol* **109**: 1597-1608.

22. Mounier J, Ryter A, Coquis-Rondon M & Sansoneti PJ (1990) Intracellular and cell-to-cell spread of *Listeria monocytogenes* involves interaction with F-actin in the enterocyte like cell line Caco-2. *Infect Immun* **58**: 1048-1058.

23. Pollard TD (2003) The cytoskeleton, cellular motility and the reductionist agenda. *Nature* **422**: 741-745.

24. Kleba B, Clark TR, Lutter EL, Ellison DW & Hackstadt T (2010) Disruption of the *Rickettsia rickettsii* Sca2 autotransporter inhibits actin-based motility. *Infect Immun* **78**: 2240-2247.

25. Merhej V & Raoult D (2010) Rickettsial evolution in the light of comparative genomics. *Biol Rev* doi: 10.1111/j.1469-185X.2010.00151.x.

26. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C *et al*. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**: 537-544.

27. Cole ST, Eiglmeier K, Parkhill J, James KD, Thompson NR *et al*. (2001) Massive gene decay in the leprosy bacillus. *Nature* **409**: 1007-1011.

28. Cole ST (2002) Comparative and functional genomics of the *Mycobacterium tuberculosis* complex. *Microbiology* **148**: 2919-2928.

29. Stinear TP, Seemann T, Harrison PF, Jenkin JA, Davies JK *et al*. (2008) Insights from the complete genome sequence of *Mycobacterium tuberculosis*. *Genome Res* **18**: 729-741.

30. Demangel C, Stinear TP & Cole ST (2009) *Buruli ulcer*: reductive evolution enhances pathogenicity of *Mycobacterium ulcerans*. *Nat Rev Microbiol* **7**: 50-60.

31. Monot M, Honoré N, Garnier T, Zidane N, Shefari D *et al*. (2009) Comparative genomic and phylogeographic analysis of *Mycobacterium leprae*. *Nat Genet* **41**: 1282-1289.

32. Niyogi SK (2005) Shigellosis. *J Microbiol* **43**: 133-143.

33. Karaolis DKR, Lan RT & Reeves PR (1994) Sequence variation in *Shigella sonnei* (Sonnei), a pathogenic clone of *Escherichia coli*, over 4 continents and 41 years. *J Clin Microbiol* **32**: 796-802.

34. Pupo GM, Lan RT & Reeves PR (2000) Multiple independent origins of *Shigella clones* of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc Natl Acad Sci U S A* **97**: 10567-10572.

35. Maureli AT, Fernandez RE, Bloch CA, Rode CK & Fasano A (1998) Black holes and bacterial pathogenicity: a large genomic deletion that enhances the virulence of *Shigella* spp. and enteroinvasive *Escherichia coli*. *Proc Natl Acad Sci U S A* **95**: 3943-3948.

36. Beutin L (2006) Emerging enterohaemorrhagic *Escherichia coli*, causes and effects of the rise of a human pathogen. *J Vet Med B Infect Dis Vet Public Health* **53**: 299–305.

37. Wu HJ, Wang HJA & Jennings MP (2008) Discovery of virulence factors of pathogenic bacteria. *Curr Opin Chem Biol* **12**: 1-9.

38. Wixon J (2001) Reductive evolution in bacteria: *Buchnera* sp., *Rickettsia prowazekii*, *Mycobacterium leprae*. *Comp Funct Genom* **2**: 44-48.

39. Sakharkar RK, Dhar KP & Chow TKV (2004) Genome reduction in prokaryotic obligatory intracellular parasites of humans: a comparative analysis. *Int J Syst Evol Microbiol* **54**: 1937-1941.

40. Andersson JO & Andersson SGE (1999) Genome degradation is an ongoing process in *Rickettsia*. *Mol Biol Evol* **16**: 1178-1191.

41. Moran NA (2002) Microbial minimalism: genome reduction in bacterial pathogens. *Cell* **108**: 583-586.

42. Fournier PE, El Karkouri K, Leroy Q, Robert C, Guimelli B *et al*. (2009) Analysis of the *Rickettsia africae* genome reveals that virulence acquisition in *Rickettsia* species may be explained by genome reduction. *BMC Genomics* **10**: 166-181.

43. Fournier PE & Didier R (2007) Identification of rickettsial isolates at the species level using multi-spacer typing. *BMC Microbiology* **7**: 72.

44. Moran NA & Wernegreen JJ (2000) Lifestyle evolution in symbiotic bacteria: insights from genomics. *Trends Ecol Evol* **15**: 321-326.

45. Moran NA & Mira A (2001) The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. *Genome Biol* **2(12)**: research0054.1-0054.12.

46. Boussau B, Karlberg EO, Frank AC, Legault BA & Andersson SGE (2004) Computational inference of scenarions for α-proteobacterial genome evolution. *Proc Natl Acad Sci U S A* **101**: 9722-9729.

47. Medina M & Sachs JL (2010) Symbiont genomics, our new tangled bank. *Genomics* doi:10.1016/j.ygeno.2009.12.004.

48. Szekeres S, Dauti M, Wilde C, Mazel D & Rowe-Magnus DA (2007) Chromosomal toxin-antitoxin loci can diminish large scale genomes reductions in the absence of selection. *Mol Microbiol* **63**: 1588-1605.

49. Jensen RB & Gerdes K (1995) Programmed cell death in bacteria: proteic plasmid stabilisation systems. *Mol Microbiol* 1**7**: 205-210.

50. Fozo EM, Makarova KS, Shabalina SA, Yutin N, Koonin EV *et al*. (2010) Abundance of type I toxin-antitoxin systems in bacteria: searches for new candidates and discovery of novel families. *Nucleic Acids Res* **38**: 3743-3759.

51. D'Elia MA, Pereira MA & Brown DE (2009) Are essential genes really essential? *Trends Microbiol* **17**: 433-438.

52. Wolfe ND, Dunavan CP & Diamond J (2007) Origins of major human infectious diseases. *Nature* **447**: 279-283.

53. van Vallen LM (1973) A new evolutionary law. *Evol Theory* **1**: 1-30.

54. Benton MJ (2009) The Red Queen and the Court Jester: species diversity and the role of biotic and abiotic factors through time. *Science* **323**: 728-732.

55. Kristoffersen P, Jensen GB, Gerdes K & Piskur J (2000) Bacterial toxin-antitoxin gene system as containment control in yeast cells. *Appl Environ Microbiol* **66**: 5524–5526.

56. Picardeau M, Le DC, Richard GF & Saint GI (2003) The spirochetal chpK chromosomal toxin-antitoxin locus induces growth inhibition of yeast and mycobacteria. *FEMS Microbiol Lett* **229**: 277–281.

57. Yamamoto TA, Gerdes K & Tunnacliffe A (2002) Bacterial toxin RelE induces apoptosis in human cells. *FEBS Lett* **519**: 191–194.

58. Alouf EJ (2000) Bacterial toxins: methods and protocols. *Methods Mol Biol* **145**: 1-26.

59. Merrell DS & Falkow S (2004) Frontal and stealth attack strategies in microbial pathogenesis. *Nature* **430**: 250-256.

60. Arber W (1978) Autobiography. Nobelprize.org.2011.

http://nobelprize.org/nobel_prizes/medicine/laureates/1978-arber-autobiography.html.

61. Kinashi Y, Nagasama H & Little BJ (1993) Mutagenic effects of restriction enzymes in Chinese hamster cells: evidence for high mutagenicity of Sau3AI and the hprt locus. *Mutat Res* **285**: 251-257.

62. Price RC, Margison GP, Hendry JH & West CM (1995) Relationships between the cytotoxic effects of restriction endonucleases and radiation on mammalian cells. *Int J Radiat Biol* **67**: 327-334.

63. Bourniquel AA & Bickle TA (2002) Complex restriction enzymes: NTP-driven molecular motors. *Biochimie* **84**: 1074-1059.

64. Masignani V, Pizza M & Rappuoli R (2006) *Bacterial toxins* in *Prokaryotes A handbook on the biology of bacteria: Ecophysiology and Biochemistry*. Chapter 1.28 **2**: 893–955 Springer, Singapore.

65. Waldor MK & Mekalanos JJ (1996) Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science* **272**: 1910-1914.

66. Nataro JP & Kaper JB (1998) Diarrheagenic *Escherichia coli*. *Clin Microbiol Rev* **11**:142-201.

67. Cavalier-Smith T (2002) The phagotrophic origin of eukaryotes and phylogenetic classification of protozoa. *Int J Syst Evol Microbiol* **52**: 297-354.

68. Cavalier-Smith T (2004) Only six kingdoms of life. *Proc Biol Sci* **271**: 1251-1262.

69. Esser C, Ahmadinejad N, Wiegand C, Rotte C, Sebastiani F *et al*. (2004) A genome phylogeny for mitochondria among α–proteobacteria and a predominantly eubacterial-ancestry of yeast nuclear genes. *Mol Biol Evol* **21**: 1643-1660.

70. Esser C & Martin W (2007) Supertrees and symbiosis in eukaryote genome evolution. *Trends Microbiol* **15**: 435-437.

71. Atteia A, Adrait A, Brugiere S, Tardif M, van Lis R *et al*. (2009) A proteomic survey of *Chlamydomonas reinhardtii* mitochondria sheds light on the metabolic plasticity of the organelle and on the nature of the alphaproteobacterial mitochondrial ancestor. *Mol Biol Evol* **26**: 1533-1548.

72. Fitzpatrick DA, Creevey CJ & McInerney JO (2006) Genome phylogenies indicate a meaningful alpha-proetoabacterial phylogeny and support a grouping of the mitochondria with the *Rickettsiales*. *Mol Biol Evol* **23**: 74-85.

73. Abhishek A, Bavishi A, Bavishi A & Choudhary M (2011) Bacterial genome chimearism and the origin of mitochondria. *Can J Microbiol* **57**: 49-61.

74. Lang BF, Burger G, O'Kelly CJ, Cedergren R, Golding GB *et al*. (1997) An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature* **387**: 493-497.

75. Raoult D (2010) The post-Darwinist rhizome of life. *The Lancet* **375**: 104-105.

76. Raoult D & Forterre P (2008) Redefining viruses: lessons from Mimivirus. *Nat Rev Microbiol* **6**: 315-319.

ANNEXES

Lors de mon travail de thèse, j'ai eu l'opportunité d'approfondir sur quelques définitions très répandues de la microbiologie. En effet, beaucoup de ces définitions sont fausses: la présence de compartimentation dans les Planctomycetes met en question les définitions des eucaryotes et des procaryotes, les virus géants contestent la définition actuelle des virus basée sur leur petite taille, en outre, la pathogenicité ne nécessite pas de facteurs de virulence alors que la perte des gènes semble jouer un rôle dans l'émergence de la virulence. En fin, la théorie de l'arbre de la vie, ne peut plus représenter nos connaissances actuelles, puisque les études de génomique ont révélé un grand nombre d'orphans et une structure mosaïque des génomes. A travers la revue de l'annexe 1, nous avons démontré que l'adoption d'une vision post-moderne de la microbiologie est nécessaire à la compréhension de l'évolution des microbes.

Dans un autre travail, présenté dans l'annexe 2, je me suis intéressée à l'origine des mitochondries, et en prenant en considération des nouvelles données par le génome d'*Odyssella thessalonicensis*, nous avons démontré que la mitochondrie de *Reclinomonas americana* a émergé des *Rickettsiales* et qu'elle est proche de *Candidatus Pelagibacter ubique*, qui, comme le démontrent nos résultats, est une *Rickettsiales* qui a probablement évolué d'un mode de vie intracellulaire à un organisme extracellulaire.

Annexe 1

The influence of Rickettsiologists on post-modern microbiology

Kalliopi Georgiades[1], Vicky Merhej[1],
and Didier Raoult[1*]

Under review in *Frontiers in Cellular and infection microbiology*

[1] Unité des Rickettsies, URMITE-CNRS UMR 6236 IRD 198 IFR 48, Faculté de Médecine, Université de la Méditerranée, 27, Bd Jean Moulin, 13005, Marseille, France

[*] Corresponding author: Email: didier.raoult@gmail.com

Abstract

Many of the definitions in microbiology are factually false. We have reviewed the great denominations of microbiology and attempted to free microorganisms from the theories of the 20th century. The presence of compartmentation and a nucleoid in Planctomycetes clearly calls into question the accuracy of the definitions of eukaryotes and prokaryotes. Archaea are viewed as prokaryotes resembling bacteria. However, the name archaea, suggesting an archaic origin of lifestyle, is inconsistent with the actual history of this family. Viruses are defined as small, filterable infectious agents, but giant viruses challenge the size criteria used for the definition of a virus. Pathogenicity does not require the acquisition of virulence factors (except for toxins), and in many cases, gene loss seems to be implicated in the emergence of virulence. Species classification based on 16S rRNA is useless for taxonomic purposes of human pathogens, as a 2% divergence would classify all *Rickettsiae* within the same species and would not identify bacteria specialized for the infection of mammals. The use of metagenomics helps us to understand evolution and physiology by elucidating the structure, function and interactions of microbial communities, but it neglects the minority populations. Finally, Darwin's descent with modification theory, as represented by the tree of life (TOL), no longer matches our current genomic knowledge because genomics has revealed the occurrence of *de novo*-created genes and the mosaic structure of genomes.

1. Introduction

Post-modern philosophy, also called the French theory (Wicks, 2003), states that the majority of theories, including scientific theories, are only based on meta-narratives expressing the influence of a culture at a given time. These scientific theories can be questioned when a change in techniques creates instability in the theory, as postulated by Karl Popper (Popper, 1959; Raoult, 2010*a*). In addition (and in the direction of post-modern theory), these theories can also be called into question due to an intellectual change of paradigm (Kuhn, 1962). The study of *Rickettsiae* has been challenging for the past few years because of the great difficulty in their handling. Moreover, the ancestors of *Rickettsiae* contributed to the birth of modern eukaryotes by transferring genes to the mitochondrion and the nucleus (Koonin, 2010; Renvoisé *et al*, 2011). During the explosion of microbial genetics, the study of *Rickettsia* did not benefit from the model of *Escherichia coli*, and rickettsiologists had to develop alternative approaches that did not include the common meta-narratives (Renvoisé *et al*., 2011). Among these approaches were those based on observations of the characteristics of intracellular bacteria, whose genomes and behaviors resemble those of viruses. Thus, *Rickettsia* has been classified as intermediate bacteria between the viruses and bacteria. Currently, the genomic revolution and "multiomics" have made it possible to analyze *Rickettsia* with many new tools (Bechah *et al*., 2010), and *Rickettsia* was among the species that were sequenced most quickly (Andersson *et al*., 1998; Ogata *et al*., 2001). This

sequencing and generally all the work achieved by Rickettsiologists brought an important revision to the way of thinking with respect to *E. coli* and forced microbiologists to visualize the general theories concerning bacterial species in a different way, so several theories concerning bacteria had, or need, to be revised (Georgiades and Raoult, 2011*a*). In this work, our goal was to revise the overarching classifications and denominations used in microbiology. In particular, as postulated by post-modern philosophy (Lyotard, 1979; Williams, 1998), we know that the denomination of an object constrains it in its definition and that when the definition is inappropriate, one cannot conceive of the object in a reasonable way.

2. Definition of eukaryotes and prokaryotes

The word “microbe”, literally meaning “small life”, was introduced by the French surgeon Charles Sédillot in 1878 to define infinitely small living organisms (Vallery-Radot, 1885). One of the most important advances in our understanding of the living world was the realization by the French scientist Edouard Chatton that there are two major groups of organisms that he named the prokaryotic (bacteria) and the eukaryotic (organisms with nucleated cells) type (Chatton, 1925; Sapp, 2005; Stanier and van Niel, 1962). This classification was adopted by Stanier and van Niel (1962) and the prokaryote-eukaryote dichotomy was universally accepted as the natural order of things until the 1970s and the emergence of rRNA phylogenetics (Sapp, 2005). At that time, Woese achieved a

comprehensive understanding of bacterial phylogeny using laborious molecular sequencing methods (Woese *et al*., 1975). Those data revealed two separate lineages among prokaryotes: the Archaea (Archaebacteria) and the Bacteria (Eubacteria). The prokaryote/eukaryote system was replaced by the "three domain system" and the classification of Eukarya, Archaea and Bacteria (Woese, 1994**).** However, bacteria had always been defined largely in negative terms: they lacked a nucleus, compartmentation and sexual reproduction (Sapp, 2005). This negative description is somewhat invalid because it does not define what a prokaryote is but rather what it is not (Pace, 2006). Furthermore, recent observations of Planctomycetes prove that the definitions of eukaryotes and prokaryotes are erroneous. Planctomycetes is a distinctive phylum of the domain Bacteria, in which the cells possess a different structural plan than other prokaryotes; the cells of all cultured and some uncultured species are divided into compartments by one or more membranes **(Fig. 1)**. In addition, in one particular species, *Gemmata obscuriglobus*, the nucleoid is enveloped in two membranes to form a nuclear body that is analogous to the structure of a eukaryotic nucleus. The existence of these organisms clearly calls into question the accuracy of the actual definitions of eukaryotes and prokaryotes (Fuerst, 1995; 2005; 2010). The nucleus of these cells likely resulted from autogenous membrane development in a prokaryote lineage (Taylor, 1976; Lake and Rivera, 1994; Glansdorff *et al*., 2008), most likely in Planctomycetes and the closely related *Chlamydia* (Ward *et al*., 2000; Horn *et al*., 2004) **(Fig. 2)**. This theory

has been strengthened by the discovery of nuclear envelope fold topology in Planctomycetes, which is analogous to the eukaryotic cell structure (Fuerst, 2005; Fuerst, 2010). Moreover, the eukaryotes all harbor mitochondria, or mitochondria-related genes, inherited from *Rickettsiales* (Golding and Gupta, 1995; Lang *et al*., 1999). Therefore, eukaryotes are younger than *Rickettsia*, their other ancestors are unknown, and there is no evidence that these ancestors had a nucleus. As it turns out, the three domain system, as previously defined, does not exist (Lake, 1988).

3. Archaea

According to a basic dichotomy, that of eukaryote versus prokaryote, Archaea, or archaebacteria, have been viewed as prokaryotes resembling bacteria (Woese and Fox, 1977; Woese, 1994). The similarity between archaea and bacteria is displayed in numerous characteristics, including the size and shape of cells, the absence of a nucleus, and the organization of genes into operons (Mayr, 1998). Koonin *et al*. have shown that about 63% of the genes in *Methanococcus janaschii* are also found in other bacteria (Koonin *et al*., 1997). Many metabolic pathways are common between archaea and bacteria (Koonin *et al*., 1997*)* but also between archaea and eukarya (Podani *et al*, 2001) and many metabolic pathways are specific to the Archaea (Sato and Atomi, 2011). *Archaea* are the only life forms known to perform methanogenesis to produce biological methane (DiMarco *et al*., 1990). Several additional characteristics,

such as the presence of ether-linked isoprenoid lipids with a glycerol-1-phosphate backbone in the cell membrane (Kates, 1993), the lack of peptidoglycan in the cell wall (Woese and Fox, 1977) the presence of modified bases in tRNA and the presence of a unique form of DNA polymerase (Ishino *et al*., 1998) distinguish archaea from bacteria. Because archaea and bacteria are no more related to each other than they are to eukaryotes, the term *prokaryote*, used for archaea to mean "not a eukaryote", is not adequate (Pace, 2006; Glansdorff *et al.*, 2008). Archaea share with eukaryotes similar transcription machinery in terms of RNA polymerase structure and function that is very different from bacterial RNA polymerase (Hirata and Murakami, 2009; Todone *et al*., 2001). This finding supported the separation of prokaryotes into archaea and bacteria and the proposal of three different evolutionary domains on the basis of ribosomal RNA (Olsen *et al*., 1994; Olsen and Woese, 1997; Werner and Grohmann, 2011; Woese and Fox, 1977; Woese, *et al*., 1990).

Archaea are similar to other domains of life and have been considered models of the early evolution of cellular life forms (Romano and Conway, 1996; Yutin *et al*., 2008). Indeed, the word archaea comes from the ancient Greek, meaning "ancient things". The information processing machineries of archaea are considered ancestral forms of the more complex replication, transcription and translation machineries of the eukaryotic cell (Gribaldo *et al*., 2010). The homology of archaeal operational genes with those of bacteria has been interpreted to be the result of an extensive lateral gene transfer (LGT) between

the two prokaryotic domains (Faguy and Doolittle 1999; Makarova *et al*., 1999). Other evolutionary hypotheses about the path of life refute the ancient characteristics of the archaea. It has been suggested that the three domains of life evolved from a pre-cellular community containing different types of genes using a process that led to the fixation of specific subsets of genes in the ancestors of these domains (Woese, 1998). These pre-cellular entities have been proposed to have no stable genealogy or chromosome and also to lack a typical cell membrane, allowing for unrestricted LGT. Considering evidence from molecular sequences, envelope structure and motility mechanisms, other hypothesis suggested that the archaea evolved from Gram-positive bacteria as an adaptation to hyperthermophily or hyperacidity (Cavalier-Smith, 2002) or in response to antibiotic selection pressure (archaea are resistant to a wide variety of antibiotics that are primarily produced by Gram-positive bacteria) (Gupta, 1998*a*; Gupta, 1998*b*; Gupta, 2000). Finally, the genomic sequence analyses by Lake and coworkers provide evidence that the root of the tree of life does not lie in archaea (Skophammer *et al*., 2006).

Archaea have often been viewed as extremophile organisms, but recent results obtained using molecular approaches and metagenomic studies have changed our perspective of the nature and the diversity of archaea. Indeed, archaea were considered predominant over bacteria in all extreme environments (temperature, salinity and pH). This is indeed true for high-temperature environments, as only archaea can thrive at temperatures from 95°C to 113°C

(Huber *et al*., 2000). However, in all other situations, species of Bacteria have been found together with those of Archaea and Eukarya (Rothschild and Mancinelli, 2001). Molecular studies have revealed the diverse and widespread distribution of archaea in different environments (Aravalli *et al*., 1998). Archaea include mesophiles that grow under mild conditions (DeLong, 1998). Novel archaea have been isolated from a variety of temperate and cold environments (Cavicchioli, 2006), human mucosa (Belay *et al*., 1988; Dridi *et al*., 2009; Miller and Wolin, 1982), agricultural and forest soils, plankton, fresh water lake sediments and the deep waters of oceans (Schleper *et al*., 2005). They constitute a major part of global ecosystems; in samples collected from the marine waters of Antarctica, archaea were estimated to account for approximately 34% of the total biomass (DeLong, 1998). Mesophilic archaea of the phylum Crenarchaeota, recently renamed to Traumarchaeota (Brochier-Armanet *et al*., 2008; Spang *et al*., 2010), account for nearly 20% of the total marine picoplankton biomass worldwide (Karner *et al*., 2001). The ubiquitous abundance of archaea and their influence on biogeochemical cycles remain largely unexplored. A recent tentative to infer the ancestral conditions of life suggests that the last common ancestor of archaea has been hyperthermophilic and mesophilic species have showed adaptations to cooler environments (Groussin and Gouy, 2011).

Recent experiments and genomic approaches have suggested that the different criteria used to define archaea are not completely valid. The definition

currently used for Archaea merely cloaks our lack of knowledge of this domain of life. Genomic studies have conclusively demonstrated that archaea are not a form of “archaic” bacteria but they rather represent a distinct evolutionary domain.

4. Bacterial virulence factors

It is not surprising that many people believe that bacteria that are dangerous to us are better armed than non-pathogenic bacteria. The conclusions of most studies on bacterial virulence, driven by anthropocentric intuition and perspective, suggested, and some still suggest that non-pathogenic bacteria lack supplementary virulence factors (Lawrence, 1999; Ochman *et al*., 2005).

An outstanding example of this way of thinking is the *Shigella* paradigm. *Shigella* spp. are human pathogens associated with bacillary dysentery, or shigellosis. *Shigella dysenteriae* causes deadly epidemics in many of the world’s poorest countries. *Shigella* spp. and *E. coli* have always been considered closely related, and they have even been placed in the same species (Pupo *et al*., 2000). However, most *E. coli* strains are commensals of the human intestine (Maurelli *et al*., 1998), and *Shigella* spp. differ from *E. coli* in their lack of certain phenotypic traits, such as extracellular mobility and the ability to ferment lactose and many sugars (Karaolis *et al*., 1994; Pupo *et al*., 2000). Similar to *S. dysenteriae*, pathogenic enteroinvasive *E. coli* lack lysine decarboxylase (LDC) activity. In a study by Maurelli *et al*. (1998), the induction of LDC expression

attenuated the virulence of a transformed strain of *S. flexneri*. It seems plausible that *Shigella* evolved from the *E. coli* complex through the acquisition of a plasmid containing critical genes. Plasmids of *Shigella* spp. have been directly associated with virulence and were even named “virulence plasmids” after their discovery (Hale *et al*., 1983). Furthermore, actin-based motility initiated by the *icsA* gene has also been reported to be a virulence factor (Goldberg *et al*., 1994). However, virulence increased after massive gene deletions (Maurelli *et al*., 1998). In conclusion, *S. dysenteriae* was not found to have more virulence genes than related bacteria (Georgiades and Raoult, 2011*a*).

Many recent comparative genomics studies have demonstrated that the specialization of bacteria for the colonization of eukaryotic hosts is associated with massive gene loss (Nierman *et al*., 2004; Merhej *et al*., 2009*a*) and the loss of identified “virulence factors” (Audic *et al*., 2007). Genomic analysis has revealed that *Borellia recurrentis*, the agent of deadly louse-borne relapsing fever, encodes fewer putative virulence factors than *Borellia duttonii* (Lescot *et al*., 2008). Gene loss has also accompanied the evolution of pathogenic *Bordetella* species (Cummings *et al*., 2004) and gene deletions in *Mycobacterium tuberculosis* have resulted in a hypervirulent phenotype (Bokum AMC *et al*., 2008). Finally, in a study by Audic *et al*. (2007), the number of putative virulence factors was found to be higher in water-dwelling bacteria than in any other categories of bacteria, including specialized pathogens (Audic *et al*., 2007).

One of the best examples of genome reduction can be found in the epidemic-causing *Rickettsia prowazekii*, which is the most dangerous rickettsial species. Genome comparisons of *R. prowazekii* with the less virulent *R. conorii* have revealed that *R. prowazekii* is a subset of *R. conorii*, with only 834 open reading frames (ORFs) compared to the 1,374 ORFs of *R. conorii* (Ogata *et al*., 2001). Although intracellular motility has been considered a virulence factor of *Shigella* (Goldberg and Theriot, 1995) and *Listeria monocytogenes* (Tilney and Portnoy, 1989; Mounier *et al*., 1990), *R. prowazekii* is completely immobile in the cytoplasm (Teysseire *et al*., 1992). Actin-based motility in *R. conorii* and *R. rickettsii* requires two proteins functioning together, Sca2 and RickA, suggesting that these two proteins could be virulence factors of *R. rickettsii*. *R. typhi* possesses only the Sca2 protein and is also mobile in the cytoplasm but less than *R. conorii* (Teysseire *et al*., 1992) **(Fig. 3)**. However, none of these proteins are found in *R. prowazekii*, which lacks actin-based motility (Kleba *et al*., 2010). Consequently, motility is not a virulence factor *per se* but can be part of a virulence repertoire in some pathogens (Georgiades and Raoult, 2011*a*). Other studies have also demonstrated genome reduction to a lower extent in the extremely successful and fit *R. africae*, the agent of African tick-bite fever. In contrast with their possession of virulence factors, *R. prowazekii* and *R. africae* have the most and the least decayed genomes, respectively, among pathogenic *Rickettsiae* (Fournier *et al*., 2009). A comparison of *R. africae* with *R. rickettsii* suggested the loss of essential genes in *R. rickettsii* as a possible factor involved

in the development of virulence (Fournier *et al*., 2009). In general, pathogenic *Rickettsia* species lack what was defined as “pathogenicity islands” and that are present in other bacterial pathogens (Hacker and Kaper, 2000). It has been suggested that plasmids contain genes encoding proteins responsible for host recognition, invasion and pathogenicity. The presence of plasmids in *Rickettsia* species, however, did not show any correlation with virulence (Paddock *et al*., 2004; Ogata *et al*., 2005; Blanc *et al*., 2007*a*). The examples of *Rickettsiae* and *Shigella* spp. show that the terms “pathogenicity islands” and “virulence plasmids” are misleading. The genomic analysis of rickettsial species has revealed that the shift to pathogenicity does not require the acquisition of new genes, but in more cases, and not only in *Rickettsia*, gene loss seems to be implicated in the emergence of virulence (Moran, 1996; Andersson and Kurland, 1998; Andersson and Andersson, 1999; Blanc *et al*., 2007*a*; Darby *et al*., 2007; Merhej *et al*., 2009*a*). In a recent study in our laboratory, we demonstrated that the only features found at higher levels in extremely dangerous bacterial pathogenic species than in closely related less pathogenic species were toxins and toxin-antitoxin modules (TA) (Georgiades and Raoult, 2011*b*).

In conclusion, except for toxins, which have a direct effect and are indeed virulence factors, other products named “virulence factors” are, in reality, associated with fitness in a genomic context and in a specific environment, including in tested experimental models. Comparative genomics have shown than pathogenic bacteria have smaller genomes than non-specialized bacteria.

Therefore, it is not possible to say that supplementary virulence factors establish pathogenicity, but rather, the overall gene repertoire is more associated with virulence than specific genes. In a recent study, the deletion of four different gene clusters in fungi attenuated their virulence in plants, while deletion of the "divergence cluster 8-12" (region encoding effector genes with mow sequence conservation) caused a hypervirulent fungal phenotype (Schirawski *et al*., 2010). Under these conditions, a virulent gene repertoire is composed of both present and absent genes. The term "virulence factor" seems to be invalid, and we propose that it should be abandoned.

5. Phylogeny and taxonomy

Biological dogma states that phylogeny reflects taxonomy. Indeed, the 16S rRNA sequence has been widely used for the description of many newly classified bacterial species (Rossello-Mora and Amann, 2001; Drancourt *et al*., 2004; Roux *et al*., 2004). A 16S rRNA divergence of 1-2% is considered to correspond to approximately 50 million years of divergence (Ochman *et al*., 1999; Ogata *et al*., 2001), and a cut-off of 98.7% similarity in 16S rRNA reflects a new species (Stackebrandt and Ebers, 2006). However, an accurate delineation of bacterial species cannot be guaranteed by the use of ribosomal DNA sequence identity, which often leads to misleading species definitions (Fox *et al*., 1992; Rossello-Mora and Amann, 2001). *Bartonella henselae* has two copies of 16S rRNA in some cases, which likely emerged through recombination (Sanogo *et*

al., 2003), and these copies present a divergence higher than 1.3% (Viezens and Arvand, 2008). For several bacterial species, the presence of multiple copies of the 16S rRNA gene has been documented (Acinas *et al*., 2004). Although generally, these multiple copies in an organism are either identical or nearly identical, in some cases, they are divergent enough to overestimate the number of bacterial strains. This overestimation can be seen in the case of *Delisea pulchra* strains, in which 16S rRNA gene copies were used to illustrate the effects of 16S rRNA heterogeneity in the marine bacterial community (Dahllöf *et al*., 2000; Adékambi *et al*., 2008). The use of 16S rRNA for such analysis is limited due to its inherent heterogeneity (Dahllöf *et al*., 2000). Moreover, using the molecular clock scale based on 16S rRNA as a species definition criterion, specialized bacteria within mammalian hosts are not defined as species (Georgiades and Raoult, 2011*a*). Species definition cannot be based on the percent divergence of 16S rRNA because bacteria having a divergence less than 1.3% correspond to bacterial complexes rather than species (Doolittle and Papke, 2006).

There are 9,000 validated bacterial species and 1.5 million eukaryotic species, even though the biomass of bacteria is comparable to that of eukaryotes; this suggests that use of the 16S rDNA sequence as a taxonomic tool is not adapted to the definition of species. Furthermore, genomic contents are not represented by phylogeny. In a study based on the genomic content comparison of bacteria with different lifestyles, discrepancies between taxonomy and gene content were

observed (Audic *et al*., 2007, Merhej *et al*., 2009*b*). The phylogenomic analysis yielded a tree similar to the one produced using the 16S rDNA gene sequence. However, *γ*-proteobacteria appeared to be divided into three groups, confirming that these species were more similar to each other in terms of gene content than to their close phylogenetic relatives (Audic *et al*., 2007). Similarly, rickettsial species and relatives, such as *Wolbachia* and *Ehrlichia* species, comprise an α-proteobacterial clade characterized by small genomes; this clade is distantly related to other α-proteobacterial species with larger genomes (Moran, 2002). Phylogenetic analysis of *Rickettsia* species based on 16S rRNA sequences has been frequently performed; however, significant inferences about intragenus phylogeny are not possible because the sequences are almost identical (Roux and Raoult, 1997). In fact, the official molecular criteria used for the classification of a bacterial species, DNA/DNA hybridization > 70%, GC content < 5% and a 16S rRNA divergence < 1, 3%, cannot be applied to *Rickettsia* species. A 16S rRNA divergence < 2% alone would classify all *Rickettsiae* within the same species (Fournier and Raoult, 2009). Furthermore, based on this criterion, bacteria specialized to mammalian hosts are not defined as a species (Georgiades and Raoult, 2011*a*). *Homo sapiens* emerged approximately 250,000 to 400,000 years ago, while the first human-specialized pathogenic bacterial species, *M. tuberculosis*, emerged much later, only 20,000 years ago (Wirth *et al*., 2008). The trees produced by the concatenation of 704, 668 and 635 core gene proteins for 7, 8 and 11 available rickettsial genome

sequences gave more robust and accurate phylogenies (Blanc *et al*., 2007*b*; Fournier *et al*., 2008; Merhej *et al*., 2009*b*). Through these analyses, the typhus group (TG) and the spotted fever group (SFG) were delineated, as well as SFG subgroups (Roux and Raoult, 1997). *R. bellii* and *R. canadensis* were also found to lie outside of the TG and the SFG (Merhej *et al*., 2009*b*).

For organisms such as archaea, bacteria and some unicellular eukaryotes, the species and gene trees do not show much identity with each other on an evolutionary scale (Bapteste *et al*., 2009). This result is due to the fact that individual gene histories can be different from the history of a species. Over the past 15 years, lateral inheritance (as opposed to vertical inheritance) has been proven to be a major evolutionary force in microorganisms (Bapteste and Boucher, 2008). Examples of extensive chimerism and LGT across prokaryotes are common, and it is absolutely plausible that every gene in prokaryotes has been affected by LGT at some point in evolutionary history (Bapteste *et al*., 2009). With this in mind, whole gene content and present and absent genes should be taken into consideration when searching for a reliable species classification **(Fig. 4)**.

6. Definition of a virus

The discovery of giant viruses with large genomes has raised many questions about virus definitions and evolution. According to Lwoff, viruses have typically been defined as “filterable infectious agents” smaller than 200 nm

that are unable to undergo binary fission and have one type of nucleic acid with few protein-encoding genes (Lwoff, 1957). Giant viruses, such as mimivirus (Raoult *et al*., 2004; Raoult *et al*. 2007) and mamavirus (La Scola *et al.* 2008), challenge the size criteria used for the definition of a virus. These viruses, with an icosahedral capsid diameter of nearly 400 nm, have particle sizes comparable to that of bacteria such as *Mycoplasma* (La Scola *et al*., 2003; Raoult *et al*., 2004). Mimivirus possesses a large double-stranded DNA genome (1,181 Kb). The mimivirus genome has 1,262 putative ORFs, of which 911 are predicted to be protein-coding genes, and 298 could be associated with functional attributes (Raoult *et al*., 2004). Mamavirus has a slightly larger genome than mimivirus (1,200 Kb), and 99% of its predicted genes are orthologous to mimivirus ORFs (Colson and Raoult, 2010). The concept of the small particle (and genome) that once defined viruses is no longer valid.

The discovery of large viruses prompted a re-evaluation of the commonly used viral isolation methods and consideration of the role played by amoebae as a source of giant viruses. Because amoebae ingest any particle that is larger than 100 nm, these phagocytes represent a potential source of giant viruses with chimeric repertoires (Raoult and Boyer, 2010). Indeed, another virus, Marseillevirus, has recently been isolated from amoebae. It has a diameter of 250nm and a genome size of 368,454 bp (Boyer *et al*., 2009). Mimivirus, Mamavirus and Marseillevirus belong to the *Mimiviridae,* a family in the group of nucleo-cytoplasmic large DNA viruses (NCLDVs) (Iyer *et al*. 2006; Boyer *et*

al., 2009). Genomic analysis of the giant viruses showed that only 4.6% and 11.2% of the ORFs of mimivirus and marseillevirus, respectively, had homologs in the NCLDV core gene set. Hence, the majority of the genome is lineage-specific. In addition to the core genome, the gene repertoire of these amoeba-associated viruses contains duplicated genes, ORFans and genes likely acquired through LGT. Indeed, a substantial proportion of the genome exhibits sequence similarities to gene homologs found in bacteria, archaea, eukaryotes and viruses (Raoult *et al*., 2004). Using phylogenetic analyses, a bacterial or bacteriophage origin has been inferred for 49 genes and a eukaryotic origin for 85 genes of the marseillevirus genome (Boyer *et al*., 2009). Likewise, 60 genes from the mimivirus genome had reliable homologs in cellular species and seemed to be acquired from eukaryotes, especially from amoebae (Moreira and Brochier-Armanet, 2008). These chimeric gene contents may have resulted from acquisitions through LGT involving the eukaryotic host (amoeba) and sympatric bacteria and viruses. Amoebae may serve as a genetic mixing bowl from which giant viruses may have gathered a complex set of genes, leading to large chimeric genomes (Raoult and Boyer, 2010).

The genomes of giant viruses help to elucidate their origin and early evolution. The position of viruses within the TOL has been a subject of disagreement. Indeed, the classification of organisms into a universal TOL based on ribosomal RNA sequences (Pace, 2006) evidently excludes viruses, which lack ribosomes. As acellular organisms, viruses were intentionally not represented with other

living ribosome-encoding organisms in the TOL (Moreira and Lopez-Garcia, 2009). Like other viruses, the mimivirus genome contains genes for replication. Surprisingly, the genome of mimivirus also contains genes that code for components of translation machinery never before found in viruses, including four amino-acyl transfer RNA synthetases, peptide release factor 1, the translation elongation factor EF-TU and translation initiation factor 1 (Raoult *et al*., 2004). The presence of these genomic features has triggered a reappraisal of the definition of living beings (Raoult and Forterre, 2008) and the evolutionary implication of viruses. The phylogenetic analysis of mimivirus proteins that have closely related eukaryotic homologs support the appearance of mimivirus as representing a fourth domain of life together with bacteria, archaea and eukaryotes (Raoult *et al*., 2004). Indeed, there are some genes that allow tracing history, including DNA processing genes, even thought a whole, complete organism cannot be represented by a classic tree of life. An additional genomic study revealed the early emergence of NCLDVs whose core genome is as ancient as the three currently accepted domains of life (Boyer *et al*., 2010*a*). These findings confirm previous hypotheses stating that viruses may be at the origin of many eukaryotic genes (Villarreal and DeFilippis, 2000; Forterre, 2006) and might have contributed to nucleus formation (Bell, 2001; Takemura, 2001). Thus, the study of giant virus genomes sheds light on the origin of eukaryotes and emphasizes the possible role played by capsid-containing organisms in the evolution of ribosome-encoding organisms.

7. Metagenomics and microbial diversity

The study of many species is difficult or even impossible, mainly due to our inability to culture them in the laboratory (Zengler *et al*., 2005). Metagenomics, or the culture-independent genomic analysis of an assemblage of organisms, allows us to study microorganisms by deciphering their genetic information from DNA that is extracted directly from communities of environmental microorganisms. Metagenomics has revealed that the vast majority of microbial diversity has been missed using cultivation-based methods (Handelsman, 2004; Riesenfeld et al., 2004; Eckburg *et al*., 2005). Indeed, approximately 10% and 60% of the sequences from environmental microbial and viral metagenomes, respectively, are novel sequences; they have no significant similarity to any sequence in the GenBank non-redundant database (Tyson *et al*., 2004; Venter *et al*., 2004; Edwards and Rohwer, 2005). Thus, our knowledge has been gleaned from the relatively small number of presently culturable representatives while ignoring the "uncultured majority" (Hugenholtz *et al*., 1998).

Metagenomics has offered unprecedented insights into microbial diversity and sparked a revolution in the field of microbiology. Historically, microbiology has focused on single species in pure laboratory cultures; thus, the understanding of microbial communities has lagged behind the understanding of their individual members. In addition, limited information about physiology and functional roles

can be gained from microbes in culture. Metagenomics is a new tool to study microbes in the complex communities in which they live and to begin to understand how these communities work. Indeed, metagenomics relies on high-throughput sequencing, which permits the isolation of large portions of genomes, providing access to protein-coding genes and biochemical pathways. Metagenomics focuses on profiling the functions encoded by a microbial community in a selected environment rather than the types of organisms producing them. Analysis of the genomic content of communities of organisms sheds light on the metabolic variability of an environment and on specific physiological functions (Eckburg *et al*., 2005; Dinsdale *et al*., 2008). Metagenomic studies of the pathogen-associated microbiome have allowed for an understanding of the role of microbial communities and their clinical implications (Gill *et al*., 2006; Ley *et al*., 2006; Turnbaugh *et al*., 2006; Willner *et al*. 2009). Information from metagenomic libraries has the ability to enrich our knowledge and has applications in many aspects of industry, therapeutics, and environmental sustainability.

Metagenomic approaches have revealed insights into environmental features with important evolutionary implications. Metagenomic functional analyses of ecosystems have revealed the correlation between geochemical conditions, metabolic capacity and genetic diversity in microbial communities (Edwards *et al*. 2006; Frias-Lopez *et al*., 2008; Simon *et al*., 2009). Indeed, sequencing projects provide a means for sampling the genetic diversity of natural microbial

populations by estimating the rate of recombination and have the potential to reveal much about the evolution of these populations (Johnson and Slatkin, 2009). Moreover, this gene-centric approach to environmental sequencing suggests that the functional profile predicted from environmental sequences of a community is similar to that of other communities whose environments of origin pose similar demands. These findings have provided insight into the processes of adaptation and the evolution of life on earth.

Notably, metagenomics represents a powerful tool that can be used to access the abundant biodiversity of environmental samples, but its accuracy depends on many limitations. Technical and economic constraints limit the depth of analysis necessary for obtaining a representative picture of microbial and viral communities, their metabolic profiles and their adaptation dynamics (Morgan *et al*., 2010). Indeed, large-scale sequencing of metagenomic DNA permits the isolation of the most predominant species in the environment, while sequences from low-abundance species may go undetected. In this way, only the most frequently represented functional genes and metabolic pathways that are relevant in a given ecosystem can be identified and assessed. However, the low-abundance species and their encoded functions could also play a critical role in the ecology and physiology of the studied environment (Piganeau and Moreau, 2007).

In conclusion, metagenomics has shown that the uncultured microbial world far outnumbers the cultured world and has emphasized the extent of our ignorance

about the microbial world. Metagenomics has helped elucidate the structure, function and interactions of microbial communities; these advances were not possible in the culture-dependent era. Metagenomics constitutes a comprehensive approach for understanding evolution and physiology.

8. Orphan genes

Orphan genes constitute a class of lineage-specific genes that do not show homology to sequences in other species (Fischer and Eisenberg, 1999). They typically encode small proteins and show high non-synonymous substitution rates, but their functions are unknown (Domazet-Loso and Tautz, 2003; Daubin and Ochman, 2004). Recently, a classification of ORFans has been proposed, dividing ORFans into singletons, multipletons, and lineage ORFans (Boyer *et al*., 2010*b*). Each newly sequenced genome contains significant numbers of such genes (Toll-Riera *et al*., 2009). For example, of 60 fully sequenced microbial genomes, 14% of genes are species-specific orphans (Siew and Fischer, 2003), while 18% of genes in *Drosophila* are restricted to the *Drosophila* group (Zhang *et al*., 2007; Zhou *et al*., 2008). However, the origin of orphan genes remains a mystery (Merkeev and Mironov, 2008). One proposed scenario is that they derived from gene duplication events in which one copy accumulated so many sequence changes that the ancestral similarity is no longer detectable (Domazet-Loso and Tautz, 2003). It was recently proposed that such ORFans could also represent genes of viral or plasmid origin (Rocha *et al*., 2006), and some seem to

correspond to truly new genes formed *de novo* through diverse mechanisms of gene evolution (Boyer *et al*., 2010*b*). This mechanism has been proposed to have made a significant contribution to the formation of novel genes in mammals, specifically in primates, in which 5.5% of orphan genes could have originated *de novo* from non-coding genomic regions (Toll-Riera *et al*., 2009). The formation of novel genes has also been described in *Drosophila* (Begun *et al*., 2006; Levine *et al*., 2006; Zhou *et al*., 2008) and *Saccharomyces cerevisiae* (Cai *et al*., 2008).

In a recent study in our laboratory, we identified a small number of gene sequences in *Rickettsia* species that had no match in any database and that seem to have resulted from *de novo* creation (Georgiades *et al*., 2011). Indeed, 17 rickettsial gene sequences seem to have no homologs in the NR database. The Ka/Ks ratio revealed that 15 of these sequences were either non-functional or had adopted functionality. Of course, the probability of pseudogenization or even of a possible viral origin of these genes should not be excluded, but because these genes were not found in regions with traces of active or ancient integrated extra-chromosomal elements, we strongly believe that they are novel genes (Georgiades *et al*., 2011).

Finally, it has been reported that new genes might be essential to an organism's viability. In the case of *Drosophila*, 59 *de novo* genes were found to be as essential as the old genes in terms of viability. The observation of lethal phenotypes caused by the knockout of new genes suggested that *de novo*-created

genes may integrate a vital pathway by interacting with existing genes, and this co-evolution may lead to the new gene becoming indispensable (Chen *et al*., 2010).

In summary, gene creation is a continuous and unsettled phenomenon, and this idea is supported by the discovery of new genes, which are permanently generated and whose identification is becoming increasingly frequent (Boyer *et al*., 2010*b*; Raoult, 2010*a*). *De novo*-created genes are evidence of life's permanent creativity.

9. The tree of life

The tree of life was used by Darwin approximately 150 years ago, as a concept to explain the evolutionary relationships between different species (Doolittle, 1999; Lawton, 2009). It has been accepted as a biological fact since (Doolittle and Bapteste, 2007). According to Darwin's theory, namely the "descent with modification theory" (Penny, 2011), the common descent of species is demonstrated by similarities between species, while modifications driven by natural selection create differences in species that result in speciation (Doolittle and Bapteste, 2007). The TOL is therefore composed of a common ancestor, the root of the tree, species separated quickly and in a stable way, key branches, and branches containing the most recently arisen species (Raoult, 2010*b*; Raoult, 2010*c*). However, evidence acquired using comparative genomic analyses contradicts the existence of a single common ancestor for the gene

repertoire of any organism. Evidence obtained through genomic analyses suggests that nearly all genes have been exchanged or recombined at some point and that there are no two genes with a similar history on the phylogenetic tree (Raoult, 2010*c*).

Since the late 1990s, LGT and gene loss in bacterial genomes have been recognized as much more frequent than previously proposed (Ochman *et al*., 2000; Lawrence, 2005; Dagan and Martin, 2007). Up to 30% of the genome-to-genome variation within a species is the result of LGT and gene loss, and homologous recombination is now thought to be the first cause of sequence divergence in many bacteria (Doolittle and Bapteste, 2007). Thus, LGT had been considered a rare phenomenon in intracellular bacteria (Audic *et al*., 2007) until the discovery of the mobilome in *Rickettsiae,* suggesting that such events were possible (Merhej and Raoult, 2010). Consequently, several further studies identified candidates for LGT in *Rickettsia* species (Wolf *et al*., 1999; Ogata *et al*., 2006; Blanc *et al*., 2007*a*; Blanc *et al*., 2007*b*; Georgiades *et al*., 2011). Moreover, genetic elements invade and proliferate in rickettsial genomes and eventually integrate genes into their host's chromosomes (Merhej and Raoult, 2010). Analysis of the *R. felis* genome has provided evidence for gene transfers between the chromosome and the *R. felis* plasmid, while the plasmids themselves seem to have been acquired through conjugation (Ogata *et al*., 2005). The first evidence for LGT in *R. bellii* also indicated the role of amoebae in gene exchanges; amoebae constitute a melting pot in which species can exchange

genetic material (Ogata *et al*., 2006; Moliner *et al*., 2010). Indeed, the genome of *R. bellii* contains many genes highly similar to those of intracellular bacteria of amoebae, such as *Legionella pneumophila* and *Protochlamydia amoebophila* (Ogata *et al*., 2006). *L. pneumophila* has developed the ability to infect different species of amoebae (Rowbotham, 1980; Fields *et al*., 2002). A recent study on *L. pneumophila* provided evidence for non-vertical inheritance: 34-57% of the genome has been involved in recombination events. In this study, LGT events between *Legionella* and all bacterial groups known to be present in amoebae were detected (Coscolla *et al*., 2011). In parallel, other studies have identified eukaryotic-like genes in *Legionella* that are most likely of amoebal origin (Lurie-Weinberger *et al*., 2010; Moliner *at al*., 2010; Schmitz-Esser *et al*., 2010). The most plausible scenario for the multiple phylogenetic origins of an important fraction of *Legionella* genes is the exchange of genetic material in the common amoeba host.

These lateral transfer events do not always involve whole genes or certain gene functions. The *R. felis* paradigm is the first rickettsial genomic analysis in which random transfers of DNA sequences were found to occur independently of gene functions or sequence lengths (Merhej *et al*., 2011, *in press*). The functional vision of genes and sequences often influences scientists' analytical strategies and interpretations. Some bacterial genomes contain up to 40% of genes with no apparent function aside from duplication (selfish genes) (Raoult, 2010*b*).

Likewise, random sequences could have hybridized between species because of their sympatric lifestyle (Mayr, 1957).

In light of these post-genomic data, a post-Darwinist concept should be introduced, one that assimilates the chimerism and mosaic structure **(Fig. 5)** of all living organisms through both non-vertical inheritance and *de novo* creation (Raoult, 2010*c*). The Tree of Life is a biblical phrase (Penny, 2011) that matches well the desire to have classification reflecting the "natural order" that is inclusively hierarchical and goes back to a single ancestor (Doolittle, 1999). Our current genomic knowledge no longer matches with Darwin's representation of the tree of life. Species evolution looks much more like a rhizome (Deleuze and Guattari, 1976; Raoult, 2010*c*), reflecting all of the various origins of genomic sequences in each species (Raoult, 2010*c*). Every living organism has a variety of ancestors; exchanges between species are intense, and the creation of new genes is frequent and constant in all organisms. For example, the human genome is a chimera and viruses and bacterial species are also our ancestors. Retroviruses left relics in our genomes, in the same way that both HHV-6A and B viruses can integrate into human chromosomes and may be vertically transmitted in the germ line (Arbuckle *et al*., 2010). *Trypanosoma cruzi* sequences were also integrated and identified into the genomes of patients (Hecht *et al*., 2010). Therefore, the definition of a common ancestor should be revised and instead of referring to a single ancestor, refer to viral ancestors, bacterial ancestors, eukaryotic ancestors and archaeal ancestors.

10. Conclusion

We think that the radical approach developed by the post-modern French philosophers is useful at this time, as technology has allowed for important discoveries. From this perspective, rickettsiologists, virologists and bacteriologists, all of whom have different points of view, can make a real contribution to their fields and to the study of the evolution of living organisms. Without the adoption of a non-traditional vision, a large proportion of living organisms, which are now within reach, will remain invisible because we will be trapped by the theories of the 20th century. Objects are constrained by their definitions. For example, giant viruses were missed by scientists and were not identified earlier because of the misleading definitions of virues that wanted them to be filterable and smaller than 200nm (Lwoff, 1957) If the definitions are false, like we demonstrated for the great denominations of microbiology, objects cannot be conceived in a reasonable way and the conclusions derived from the observations of the microorganisms will be biased by misleading believes and theories.

Conflict of interest statement

The authors declare that they have no competing interests.

Acknowledgements

We would like to thank Christelle Forzales for her help with Figure 5.

References

Acinas, S. G., Marcelino, L. A, Klepac–Ceraj, V., and Polz, M. F. (2004). Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. *J Bacteriol* 186**,** 2629–2635.

Adékambi, T., Drancourt, M., and Raoult, D. (2008). The *rpo*B gene as a tool for clinical microbiologists. *Trends Microbiol* 17, 37-45.

Andersson, S. G., Zomorodipour, A., Andersson, J. O., Sicheritz-Pontén, T., Alsmark, U. C., Podowski, R. M., Näslund, A. K., Eriksson, A. S., Winkler, H. H. and Kurland, C. G. (1998). The genome sequence *of Rickettsia prowazekii* and the origin of mitochondria. *Nature* 396, 133-140.

Andersson, J. O., and Andersson, S. G. (1999). Genome degradation is an ongoing process in *Rickettsia*. *Mol Biol Evol* 16, 1178-1191.

Andersson, J. O., and Kurland, C. G. (1998). Reductive evolution of resident genomes. *Trends Microbiol* 6, 263-268.

Aravalli, R. N., She, Q., and Garrett, R. A. (1998). Archea and the new age of microorganisms. *Trends Ecol Evol* 13, 190-194.

Arbucklea, J. H., Medveczkya, M. M., Lukab, J., Hadleya, S. H., Luegmayrc, A., Ablashid, D., Lunde, T. C., Tolare, J., Meirleirf, K. D., Montoyag, J. G., Komaroffh, A. L., Ambrosc, P.F., and Medveczkya, P.G. (2010). The latent human herpesvirus-6A genome specifically integrates in telomeres of human chromosomes *in vivo* and *in vitro*. *Proc Natl Acad Sci U S A* 107, 5563–5568.

Audic, S., Robert, C., Campagna, B., Parinello, H., Claverie, J. M., Raoult, D., and Drancourt, M. (2007). Genome analysis of *Minibacterium massiliensis* highlights the convergent evolution of water-living bacteria. *Plos Genet* 3, e138.

Bapteste, E., and Boucher, Y. (2008). Lateral gene transfer challenges principles of microbial systematics. *Trends Microbiol* 16, 200-207.

Bapteste, E., O'Malley, M. A., Beiko, R. G., Ereshefsky, M., Gogarten, J. P., Franklin-Hall, L., Lapointe, F. J., Dupré, J., Dagan, T., Boucher, Y., and Martin, W. (2009). Prokaryotic evolution and the tree of life are two different things. *Biol Direct* 4, 34.

Bechah, Y., El Karkouri, K., Mediannikov, O., Leroy, Q., Pelletier, N., Robert, C., Médigue, C., Mege, J. L., and Raoult, D. (2010). Genomic, proteomic and

transcriptomic analysis of virulent and avirulent *Rickettsia prowazekii* reveals its adaptive mutation capabilities. *Genome Res* doi:10.1101/gr.103564.109.

Begun, D. J., Lindfors, H.A., Thompson, M. E., and Holloway, A. K. (2006). Recently evolved genes identified from *Drosophila yakuba* and *D. erecta* accessory gland expressed sequence tags. *Genetics* 172, 1675-1681.

Belay, N., Johnson, R., Rajagopal, B. S., Conway de, M. E., and Daniels, L. (1988). Methanogenic bacteria from human dental plaque. *Appl Environ Microbiol* 54, 600-603.

Bell, P. J. (2001). Viral eukaryogenesis: was the ancestor of the nucleus a complex DNA virus? *J Mol Evol* 53, 251-256.

Blanc, G., Ogata, H., Robert, C., Audic, S., Suhre, K., Vestris, G., Claverie, J.,M., and Raoult, D. (2007*a*). Reductive genome evolution from the mother of *Rickettsia*. *Plos Genet* 3, e14.

Blanc, G., Ogata, H., Robert, C., Audic, S., Claverie, J. M., and Raoult, D. (2007*b*). Lateral gene transfer between obligate intracellular bacteria: evidence from the *Rickettsia massiliae* genome. *Genome Res* 17, 1657-1664.

Bokum, A. M. C., Movahedzadeh, F., Frita, R., Bancroft, G. J., and Stoker, N. G. (2008). The case for hypervirulence through gene deletion in *Mycobacterium tuberculosis*. *Cell* 16, 436-441.

Boyer, M., Madoui, M. A., Gimenez, G., La Scola, B., and Raoult, D. (2010*a*), Phylogenetic and phyletic studies of informational genes in genomes highlight existence of a 4 domain of life including giant viruses. *Plos One* 5, e15530.

Boyer, M., Gimenez, G., Suzan-Monti, M., and Raoult, D. (2010*b*). Classification and determination of possible origins of ORFans through analysis of nucleocytoplasmic large DNA viruses. *Intervirology* 53, 310-320.

Boyer, M., Yutin, N., Pagnier, I., Barrassi, L., Fournous, G., Espinosa, L., Robert, C., Azza, S., Sun, S., Rossmann, M. G., Suzan-Monti, M., La Scola, B., Koonin, E. V., and Raoult, D. (2009). Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proc Natl Acad Sci U S A* 106, 21848-21853.

Brochier-Armanet, C., Boussau, B., Gribaldo, S., and Forterre, P. (2008). Mesophilic *Crenarcheaota*: proposal for a third archeal phylum, the *Traumarcheaota*. *Nat Rev Microbiol* 6, 245-252.

Bromham, L., and Penny, D. (2003). The modern molecular clock. *Nat Rev Genet* 4, 216–224.

Cai, J., Zhao, R., Jiang, H., and Wang, W. (2008). *De novo* origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics* 179, 487-496.

Cavalier-Smith, T. (2004). Only six kingdoms of life. *Proc Biol Sci* 271, 1251-1262.

Cavalier-Smith, T. (2002). The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. *Int J Syst Evol Microbiol* 52, 7-76.

Cavicchioli, R. (2006). Cold-adapted archea. *Nat Rev Microbiol* 4, 331-343.

Chatton, E. (1925), *Pansporella perplexa*. Réflexions sur la biologie et la phylogènie des protozoaires. 10e serie. *Ann Sci Nat Zool* 7: 1-84.

Chen, S., Zhang, Y. E., and Long, M. (2010). New genes in *Drosophila* quickly become essential. *Science* 330, 1682-1685.

Colson, P., and Raoult, D. (2010). Gene repertoire of amoeba-associated giant viruses. *Intervirology* 53, 330-343.

Coscolla, M., Comas, I., and Gonzales-Candelas, F. (2011). Quantifying nonvertical inheritance in the evolution of *Legionella pneumophila*. *Mol Biol Evol* 28, 985-1001.

Cummings, C. A., Brinig, M. M., Lepp, P. W., van de Pas, S., and Relman, D. A. (2004). *Bordetella* species are distinguished by patterns of substantial gene loss and gene adaptation. *J Bacteriol* 186, 1484-1492.

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.

Dahllöf, I., Baillie, H., and Kjelleberg, S. (2000). *rpo*B-based microbial community analysis avoids limitations inherent in 16s rRNA genes intraspecies heterogeneity. *Appl Environ Microbiolol* 66, 3376-3380.

Darby, A. C., Cho, N. H., Fuxelius, H. H., Westberg, J., and Andersson, S. G. (2007). Intracellular pathogens go extreme: genome evolution in the *Rickettsiales. Trends Genet* 23, 511-520.

Daubin, V., and Ochman, H. (2004). Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. *Genome Res* 14, 1036-1042.

Deleuze, G., and Guattari, F. (1976). *Rhizome*. Paris: Les éditions de minuit.

DeLong, F. F. (1998). Everything in moderation: Archea as "non-extremophiles". *Curr Opin Genet Dev* 8, 649-654.

DiMarco, A. A., Bobik, T. A., and Wolfe, R. S. (1990). Unusual coenzymes of methanogenesis. *Annu Rev Biochem* 59, 355-394.

Dinsdale, E. A., Edwards, E. A., Hall, D., Angly, F., Breitbart, M., Brulc, J. M., Furlan, M., Desnues, C., Haynes, M., Li, L., McDaniel, L., Moran, M. A., Nelson, K. E., Nilsson, C., Olson, R., Paul, J., Brito, B. R., Ruan, Y., Swan, B. K., Stevens, R., Valentine, D. L., Thurber, R. V., Wegley, L., White, B. A., and Rohwer, F. (2008). Functionnal metagenomic profiling of nine biomes. *Nature* 452, 629-632.

Dridi, B., Henry, M., El Karkouri, K., Raoult, D., and Drancourt, M. (2009). High prevalence of *Methanobrevibacter smithii* and *Methanosphaera stadtmanae* detected in the human gut using an improved DNA detection protocol. *Plos One* 4, e7063.

Domazet-Loso, T., and Tautz, D. (2003). An evolutionary analysis of orphan genes in *Drosophila*. *Genome Res* 13, 2213-2219.

Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. *Science* 284, 2124-2128.

Doolittle, W. F. and Papke, R. T. (2006). Genomics and the bacterial species problem. *Genome Biol* 7, 116.

Drancourt, M., Berger, P., and Raoult, D. (2004). Systematic 16S rRNA gene sequencing of atypical clinical isolates identified 27 new bacterial species associated with humans. *J Clin Microbiol* 42, 2197-2202.

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.

Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S. R., Nelson, K. E., and Relman, D. A. (2005). Diversity of the human intestinal microbial flora. *Science* 308, 1635-1638.

Edgell, D. R., and Doolittle, W. F. (1997). Archea and the origin(s) of DNA replication proteins. *Cell* 89, 995-998.

Edwards, R. A., Rodriguez-Brito, B., Wegley, L., Haynes, M., Breitbart, M., Peterson, D. M., Saar, M. O., Alexander, S., Alexander, E. C. Jr., and Rohwer, F. (2006). Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* 7, 57.

Edwards, R. A., and Rohwer, F. (2005). Viral metagenomics. *Nat Rev Microbiol* 3, 504-510.

Faguy, D. M., and Doolittle, W. F. (1999). Lessons from the *Aeropyrum pernix* genome. *Curr Biol* 9, R883-R886.

Fields, B. S., Benson, R. F., and Besser, R. E. (2002). *Legionella* and Legionnaires' disease: 25 years of investigation. *Clin Microbiol Rev* 15, 506-526.

Fischer, D., and Eisenberg, D. (1999). Finding families for genomic ORFans. *Bioinformatics* 15, 759-762.

Forterre, P. (2006). The origin of viruses and their possible roles in major evolutionary transitions. *Virus Res* 117, 5-16.

Fournier, P. E., Belghazi, I., Robert, C., Elkarkouri, K., Richards, A. L., Greub, G., Collyn, F., Ogawa, M., Portillo, A., Oteo, J. A., Psaroulaki, A., Bitam, I., and Raoult, D. (2008). Variations of plasmid content in *Rickettsia felis*. *Plos One* 3, e2289.

Fournier, P. E., Elkarkouri, K., Leroy, Q., Robert, C., Guimelli, B., Renesto, P., Socolovschi, C., Parola, P., Audic, S., and Raoult, D. (2009). Analysis of the *Rickettsia africae* genome reveals that virulence acquisition in *Rickettsia* species may be explained by genome reduction. *BMC Genome* 10, 166-181.

Fournier, P. E., and Raoult, D. (2009). Current knowledge on phylogeny and taxonomy of *Rickettsia* spp. *Ann N Y Acad Sci* 1166, 1-11.

Fox, G. E., Wisotzkey, J. D., and Jurtshuk, P. (1992). How close is close – 16S ribosomal –RNA sequence identity may no be sufficient to guarantee species identity. *Int J Syst Bacteriol* 42, 166-170.

Frias-Lopez, J., Shi, Y., Tyson, G. W., Coleman, M. L., Schuster, S. C., Chisholm, S. W., and DeLong, E. F. (2008). Microbial community gene expression in ocean surface waters. *Proc Natl Acd Sci U S A* 105, 3805-3810.

Fuerst, J. A. (1995). *Planctomycetes*-a phylum of emerging interest for microbial evolution and ecology. *Microbiology* 141, 1493-1506.

Fuerst, J. A. (2005). Intracellular compartmentation in *Planctomycetes*. *Ann Rev Microbiol* 59, 299-328.

Fuerst, J. A. (2010). Beyond Prokaryotes and Eukaryotes: *Planctomycetes* and cell organization. *Nature Education* 3(9), 44.

Georgiades, K., and Raoult, D (2011*a*). Defining pathogenic bacterial species in the genomic era. *Front Microbio* 1, 151.

Georgiades, K., and Raoult, D. (2011*b*). Genomes of the most dangerous epidemic bacteria have a virulence repertoire characterized by fewer genes but more Toxin-Antitoxin modules. *Plos One* **6(3):** e17962.

Georgiades, K., Merhej, V., El Karkouri, K., Raoult, D., and Pontarotti, P. (2011). Gene gain and loss events in *Rickettsia* and *Orientia* species. *Biol Direct* 6, 6.-

Gill, S. R., Pop, M., Deboy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M., and Nelson, K. E. (2006). Metagenomic analysis of the human distal gut microbiome. *Science* 312, 1355-1359.

Glasndorff, N., Xu, Y., and Labedan, B. (2008). The last universal common ancestor: emergence, constitution, and genetic legacy of an elusive forerunner. *Biol Direct* 3, 29.

Goldberg, M. B., and Theriot, J. A. (1995). *Shigella flexneri* surface protein IcsA is sufficient to direct actin-based motility. *Proc Natl Acad Sci U S A* 92, 6572-6576.

Goldberg, M. B., Theriot, J. A., and Sansonetti, P. J. (1994). Regulation of surface presentation of IcsA, a *Shigella* protein essential to intracellular movement and spread, is growth phase dependent. *Infect Immun* 62, 5664-5668.

Golding, G. B., and Gupta, R. S. (1995). Protein-based phylogenies support a chimeric origin for the eukaryotic genome. *Mol Biol Evol* 12, 1-6.

Gribaldo, S., Poole, A. M., Daubin, V., Forterre, P., and Brochier-Armanet, C. (2010). The origin of eukaryotes and their relationship with the Archea: are we at a phylogenomic impasse? *Nat Rev Microbiol* 8, 743-752.

Groussin, M., and Gouy, M. (2011). Adaptation to environmental temperature is a major determinant of molecular evolutionary rates in Archaea. *Mol Biol Evol* doi:10.1093/molbev/msr098.

Gupta, R. S. (1998*a*). Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among alphaproteobacteria, Eubacteria and eukaryotes. *Microbiol Mol Biol Rev* 62, 1435-1491.

Gupta, R. S. (1998*b*). What are archaebacteria: life's third domain or monoderm prokaryotes related to gram-positive bacteria? A new proposal for the classification of prokaryotic organisms. *Mol Microbiol* 29, 695-707.

Gupta, R. S. (2000). The natural evolutionary relationships among prokaryotes. *Crit Rev Microbiol* 26, 111-131.

Hacker, J., and Kaper, J.B. (2000). Pathogenicity islands and the evolution of microbes. *Annu Rev Microbiol* 54, 641-679.

Hale, T. L., Sansonetti, P. J., Schad, P. A., Austin, S., and Formal, S. B. (1983). Characterization of virulence plasmids and plasmid-associated outer membrane proteins in *Shigella flexneri*. *Infect Immun* 40, 340-350.

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* 68, 669-685.

Hecht, M. M., Nitz, N., Araujo, P. F., Sousa, A. O., de Cássia Rosa, A., Gomes, D. A., Leonardecz, D., and Teixeira, A. R. L. (2010). Inheritance of DNA transferred from American trypanosomes to human hosts. *Plos One* 5, e9181.

Hirata, A., and Murakami, K. S. (2009). Archaeal RNA polymerase. *Curr Opin Struct Biol* 19, 724-731.

Horn, M., Collingro, A., Schmitz-Esser, S., Beier, C. L., Purkhold, U., Fartmann, B., Brandt, P., Nyakatura, G. J., Droege, M., Frishman, D., Rattei, T., Mewes, H. W., and Wagner, M. (2004). Illuminating the evolutionary history of chlamydiae. *Science* 304, 728-730.

Huber, R., Huber, H., and Setter, K. O. (2000). Towards the ecology of hyperthermophiles: biotopes, new isolation strategies and novel metabolic properties. *FEMS Microbiol Rev* 24, 615-623.

Hugenholtz, P., Goebel, B.M., and Pace, N. R. (1998). Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J Bacteriol* 180, 4765-4774.

Ishino, Y., Komori, K., Cann, I. K., and Koga, Y. J. (1998). A novel DNA polymerase family found in Archaea. *Bacteriol* 180, 2232-2236.

Iyer, L.M., Blaji, S., Koonin, E. V. and Aravind, L. (2006). Evolutionary genomics of nucleo-cytoplamsic large DNA viruses. *Virus Res* 117, 156-184.

Johnson, P. L., and Statkin, M. (2009). Inference of microbial recombination rates from metagenomic data. *Plos Genet* 5, e1000674.

Karaolis, D. K. R., Lan, R. T., and Reeves, P. R. (1994). Sequence variation in *Shigella sonnei* (Sonnei), a pathogenic clone of *Escherichia coli*, over 4 continents and 41 years. *J Clin Microbiol* 32, 796-802.

Karner, M. B., DeLong, E. F., and Karl, D. M. (2001). Archaeal dominance in the mesopelagic zone of the Pacific Ocean. *Nature* 409, 507-510.

Kates, M. (1993). "Membrane lipids of Archea," in *The biochemistry of Archea*. ed M. Kates, D. J. Kushnen, A. T. Matheson (Elsevier, Amsterdam) 261-295.

Kleba, B., Clark, T. R., Lutter, E. L., Ellson, D.W., and Hackstadt, T. (2010). Disruption of the *Rickettsia rickettsii* Sca2 autotransporter inhibits actin-based motility. *Infect Immun* 78, 2240-2247.

Koonin, E. V., Mushegian, A. R., Galperin, M. Y., and Walker, R. D. (1997). Comparison of archeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archea. *Mol Microbiol* 25, 619-637.

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

Kuhn, S. T. (1962) *The structure of scientific revolutions*. University of Chicago Press.

La Scola, B., Audic, S., Robert, C., Jungang, L., De, L., Drancourt, M., Birtles, R., Claverie, J. M., and Raoult, D. (2003). A giant virus in amoebae. *Science* 299, 2033.

La Scola, B., Desnues, C., Pagnier, I., Robert, C., Barrassi L., Fournous, G., Merchat, M., Suzan-Monti, M., Forterre, P., Koonin, E. V., and Raoult, D. (2008). The virophage as a unique parasite of the giant mimivirus. *Nature* 455, 100-104.

Lake, J. A. (1988). Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. *Nature* 331, 184-186.

Lake, J. A., and Rivera, M. C. (1994) Was the nucleus the first endosymbiont? *Proc. Natl. Acad. Sci. USA* 91, 2880–2881.

Lang, F. B., Gray, M. W., and Burger, G. (1999). Mitochondrial genome evolution and the origin of Eukaryotes. *Annu Rev Genet* 33, 351-397.

Lawrence, J. G. (1999). Gene transfer speciation and the evolution of bacterial genomes. *Curr Opin Microbiol* 2, 519-523.

Lawrence, J. G. (2005). Common themes in the genome strategies of pathogens. *Curr Opin Genet Dev* 15, 584-588.

Lawton, G. (2009). Why Darwin was wrong about the tree of life. *New Scientist*. Issue number 2692, pp34-39.

Lescot, M., Audic, S., Robert, C., Nguyen, T. T., Blanc, G., Cutler, S. J., Wincker, P., Couloux, A., Claverie, J. M., Raoult, D., and Drancourt, M. (2008). The genome of *Borellia recurrentis*, the agent of deadly louse-borne relapsing fever, is a degraded subset of tick-borne *Borrelia duttonii*. *Plos Genet* 4, e1000185.

Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A., and Begun, D. J. (2006). Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc Natl Acad Sci U S A* 103, 9935-9939.

Ley, R. E., Turnbaugh, P. J., Klein, S., and Gordon, J. I. (2006). Microbial ecology: human gut microbes associated with obesity. *Nature* 444, 1022-1023.

Lurie-Weinberger, M. N., Gomez-Valero, L., Merault, N., Glockner, G., Buchrieser, C., and Gophna, U. (2010). The origins of eukaryotic-like proteins in *Legionella pneumophila*. *Int J Med Microbiol* 7, 470-481.

Lwoff, A. (1957). The concept of virus. *J Gen Microbiol* 17, 239-253.

Lyotard, J. F. (1979) *La condition postmoderne-Rapport sur le savoir*. Paris: Les éditions de minuit.

Makarova, K. S., Aravind, L., Galperin, M. Y., Grishin, N. V., Tatuson, R. L., Wolf, Y. I., and Koonin, E. V. (1999). Comparative genomics of the Archeal (Euryarchaeota): evolution of conserved protein families, the stable one, and the variable shell. *Genome Res* 9, 608-628.

Mayr, E. (1957). *The species problem*. Washington, DC: American Association for the Advancement of Science.

Mayr, E. (1998) Two empires or three? *Proc Natl Acad Sci U S A* 95, 9720-9723.

Maurelli, A. T., Fernandez, R. E., Bloch, C. A., Rode, C. K., and Fasano, A. (1998). Black holes and bacterial pathogenicity: a large genomic deletion that enhances the virulence of *Shigella* spp. and enteroinvasive *Escherichia coli*. *Proc Natl Acad Sci U S A* 95, 3943-3948.

Merhej, V., Royer-Carenzi, M., Pontarotti, P., and Raoult, D. (2009*a*). Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. *Biol Direct* 4, 13.

Merhej, V., El Karkouri, K., and Raoult, D. (2009*b*). Whole genome-based phylogenetic analysis of *Rickettsiae*. *Clin Micrbiol Infect Dis* doi:10.111/j.1469-0691.2008.02265.x.

Merhej, V., Notredame, C., Royer-Carenzi, M., Pontarotti, P., and Raoult, D. (2011). The rhizome of life: the sympatric *Rickettsia felis* paradigm demonstrates random transfer of sequences. *Mol Biol Evol* 2011, in press, MBE -11 – 0060.R2.

Merhej, V., and Raoult, D. (2010) Rickettsial evolution in the light of comparative genomics. *Biol Rev Camp Philos Soc* doi:10.1111/j.1469-185X.2010.00151.x.

Merkeev, I. V., and Mironov, A. A. (2008) Orphan genes: function, evolution and composition. *Mol Biol* 42, 127-132.

Miller, T. L., and Wolin, M. J. (1982) Enumeration of *Methanobrevibacter smithii* in human feces. *Arch Microbiol* 131, 14-18.

Moliner, C., Fournier, P. E., and Raoult, D. (2010). Genome analysis of microorganisms living in amoebae reveals a melting pot for evolution. *FEMS Microbiol Rev* 34, 281-294.

Moran, N. A. (1996). Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc Natl Acad Sci U S A* 93, 2873-2878.

Moran, N. A. (2002). Microbial minimalism: genome reduction in bacterial pathogens. *Cell* 108, 583-586.

Moreira, D., and Brochier-Armanet, C. (2008). Giant viruses, giant chimeras: the multiple evolutionary histories of Mimivirus genes. *BMC Evol Biol* 8, 12.

Moreira, D., and Lopez-Garcia, P. (2009). Ten reasons to exclude viruses from the tree of life. *Nat Rev Microbiol* 7, 306-311.

Morgan, J. L., Darling, A. E., and Eisen, J. A. (2010). Metagenomic sequencing of an in vitro-simulated microbial community. *Plos One* 5, e10209.

Mounier, J., Ryter, A., Coquis-Rondon, M., and Sansonetti, P. J. (1990). Intracellular and cell-to-cell spread of *Listeria monocytogenes* involves interaction with F-actin in the enterocyte like cell line Caco-2. *Infect Immun* 58, 1048-1058.

Nierman, W. C., DeShazer, D., Kim, H. S., Tettelin, H., Nelson, K. E., Feldblyum, T., Ulrich, R. L., Ronning, C. M., Brinkac, L. M., Daugherty, S. C., Davidsen, T. D., Deboy, R. T., Dimitrov, G., Dodson, R. J., Durkin, A. S., Gwinn, M. L., Haft, D. H., Khouri, H., Kolonay, J. F., Madupu, R., Mohammoud, Y., Nelson, W. C., Radune, D., Romero, C. M., Sarria, S., Selengut, J., Shamblin, C., Sullivan, S. A., White, O., Yu, Y., Zafar, N., Zhou, L., Fraser, C. M. (2004). Structural flexibility in the *Burkholderia mallei* genome. *Proc Natl Acad Sci U S A* 101, 14146-14251.

Ochman, H., Elwyn, S., and Moran, N. A. (1999). Calibrating bacterial evolution. *Proc Natl Acad Sci U S A* 96, 12638-12643.

Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* 405, 299–304.

Ochman, H., Lerat, E., and Daubin, V. (2005). Examining bacterial species under the specter of gene transfer and exchange. *Proc Natl Acad Sci U S A* 102, 6595-6599.

Ogata, H., La Scola, B., Audic, S., Renesto, P., Blanc, G., Robert, C., Fournier, P. E., Claverie, J. M., and Raoult, D. (2006). Genome sequence of *Rickettsia bellii* illuminates the role of amoebae in gene exchanges between intracellular pathogens. *Plos Genet* 2, e76.

Ogata, H., Renesto-Audiffren, P., Audic, S., Robert, C., Blanc, G., Fournier, P. E., Parinello, H., Claverie, J. M., and Raoult, D. (2005). The genome sequence of *Rickettsia felis* identifies the first putative conjugative plasmid in an obligate intracellular parasite. *Plos Biol* 3, e248.

Ogata, H., Renesto-Audiffren, P., Fournier, P. E., Barbe, V., Samson, D., Roux, V., Cossart, P., Weissenbach, J., Claverie, J. M., and Raoult, D. (2001). Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. *Science* 293, 2093-2098.

Olsen, G. J., and Woese, C. R. (1997). Archaeal genomics: an overview. *Cell* 89, 991-994.

Olsen, G. J., Woese, C. R., and Overbeek, R. (1994). The winds of (evolutionary) change: breathing new life into microbiology. *J Bacteriol* 176, 1-6.

Pace, N. R. (2006). Concept time for a change. *Nature* 441, 289.

Paddock, C. D., Sumner, J. W., Comer, J. A., Zari, S. R., Goldsmith, C. S., Goddard, J., McLellan, S. L., Tamminga, C. L., and Ohl, C. A. (2004). *Rickettsia parkeri*: a newly recognized cause of spotted fever rickettsiosis in the United States. *Clin Infect Dis* 38, 805-811.

Penny, D. (2011). Darwin's theory of descent with modification, versus the biblical tree of life. *Plos Biol* 9, e1001096.

Piganeau, G., and Moreau, H. (2007). Screening the Sargasso Sea metagenome for data to investigate genome evolution in *Ostreococcus* (*Prasinophyceae*, Chlorophyta). *Gene* 406, 184-190.

Podani, J., Oltvai, Z. N., Jeong, H., Tombor, B., Barabási, A. L., and Szathmáry, E. (2001). Comparable system-level organization of Archaea and Eukaryotes. *Nat Genet* 29, 54-56.

Popper, K. (1959). *The logic of scientific discovery*. New Ed editions. London: Taylor & Francis Group.

Pupo, G. M., Lan, R. T., and Reeves, P. R. (2000). Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc Natl Acad Sci U S A* 97, 10567-10572.

Raoult, D. (2010*a*). Technology-driven research will dominate hypothesis-driven research: the future of microbiology. *Future Microbiol* 5(2), 135-137.

Raoult, D. (2010*b*). « L'homme, cette chimère: l'évolution selon Darwin relue et corrigée », in *Dépasser Darwin*, ed. Plon, France, 15-55.

Raoult, D. (2010*c*). The post-Darwinist rhizome of life. *The Lancet* 375, 104-105.

Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., Suzan, M., and Claverie, J. M. (2004). The 1.2-megabase genome sequence of Mimivirus. *Science* 306, 1344-1350.

Raoult, D., and Boyer, M. (2010). Amoebae as genitors and reservoirs of giant viruses. *Intervirology* 53, 321-329.

Raoult, D., and Forterre, P. (2008). Redefining viruses: lessons from Mimivirus. *Nat Rev Microbiol* 6, 315-317.

Raoult, D., La Scola, B., and Birtles, R. (2007). The discovery and characterization of Mimivirus, the largest known virus and putative pneumonia agent. *Clin Infect Dis* 45, 95-102.

Renvoisé, A., Merhej, V., Georgiades, K., and Raoult, D. (2011). Intracellular *Rickettsiales*: insights into manipulators of eukaryotic cells. *Trends Mol Med*, doi:10.1016/j.molmed.2011.05.009.

Riesenfeld, C. S., Schloss, P. D., and Handelsman, J. (2004). Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet* 38, 525-552.

Rocha, M. T. G., Cooper, J. E., Smith, N. H., and Feil, E. J. (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J Theor Biol* 239, 226-235.

Romano, A. H., and Conway, T. (1996). Evolution of carbohydrate metabolic pathways. *Res Microbiol* 147, 448-455.

Rosello-Mora, R., and Amann, R. (2001). The species concept for prokaryotes. *FEMS Microbiol Rev* 25, 39–67.

Rothschild, L. J., and Mancinelli, R. L. (2001). Life in extreme environments. *Nature* 409, 1092-1101.

Roux, V., Drancourt, M., Stein, A., Riegel, P., Raoult, D., and La Scola, B. (2004). *Corynebacterium* species isolated from bone and joint infection identified by 16 rRNA gene sequence analysis. *J Clin Microbiol* 42, 2231–2233.

Roux, V., and Raoult, D. (1997). Rickettsioses as paradigms of new emerging infectious diseases. *Clin Microbiol Rev* 10, 694-719.
Rowbothan, T. J. (1980). Preliminary report on the pathogenicity of *Legionella pneumophila* for freshwater and soil amoebae. *J Clin Pathol* 33, 1179-1183.

Sanogo, Y. O., Zeaiter, Z., Caruso, G., Merola, F., Shpynov, S., Brouqui, P., and Raoult, D. (2003). *Bartonella henselae* in *Ixodes ricinus* ticks (Acari: Ixodida) removed from humans, Bellulo Province, Italy. *Emerg Infect Dis* 9, 329.332.

Sapp, J. (2005). The prokaryotic-eukaryotic dichotomy: meanings and mythology. *Microbiol Mol Biol Rev* 69, 292-305.

Sato, T., and Atomi, H. (2011). Novel metabolic pathways in Archaea. *Curr Opin Microbiol*
14, 307-14.

Schleper, C., Jurgens, G., and Jonuscheit, M. (2005). Genomic studies of uncultivated Archaea. *Nat Rev Microbiol* 3, 479-488.

Schmitz-Esser, S., Tischler, P., Arnold, R., Montanaro, J., Wagner, M., Rattei, T., and Horn, M. (2010). The genome of the amoeba symbionts '*Candidatus Amoebophilus asiaticus*' reveals common mechanisms for host cell interaction among amoeba-associated bacteria. *J Bacteriol* 192, 1045-1057.

Schirawski, J., Mannhaupt, G., and Münch, K. (2010). Pathogenicity determinants in smut fungi revealed by genome comparison. *Science* 330, 1546-1548.

Siew, N., and Fischer, D. (2003). Analysis of singleton ORFans in fully sequenced microbial genomes. *Proteins* 53, 241-251.

Simon, C., Wiezer, A., Strittmatter, A. W., and Daniel, R. (2009). Phylogenetic diversity and metabolic potential revealed in a glacier ice metagenome. *Appl Environ Microbiol* 75, 7519-7526.

Skophammer, R. G., Herbold, C. W., Rivera, M. C., Servin, J. A., and Lake, J. E. (2006). Evidence that the root of the tree of life is not within the Archea. *Mol Biol Evol* 23, 1648-1651.

Spang, A., Hatzenpichler, R., Brochier-Armanet, C., Rattei, T., Tischler, P., Spieck, E., Streit, W., Stahl, D. A., Wagner, M., and Schleper, C. (2010). Distinct gene set in two different lineages of ammonia-oxidizing Archea supports the phylum *Thaumarchaeota*. *Trends Microbiol* 18, 331-340.

Stackebrandt, E., and Ebers, J. (2006). Taxonomic parameters revisited: tarnished gold standards. *Microbiol Today* 33, 152–155.
Takemura, M. (2001). Poxviruses and the origin of the eukaryotic nucleus. *J Mol Evol* 52, 419-425.

Todone, F., Brick, P., Werner, F., Weinzierl, R. O., and Onesti, S. (2011). Structure of an archaeal homolog of the eukaryotic RNA polymerase II RPB4/RPB7 complex. *Mol Cell* 8, 1137-1143.

Stanier, R. Y., and van Niel, C. B. (1962). The concept of a bacterium. *Arch Mikrobiol* 42, 17-35.

Sun, F. J. and Caetano-Anolles, G. (2008). Evolutionary patterns in the sequence and structure of transfer RNA: early origins of archaea and Viruses. *Plos Comput Biol* 4, e1000018.

Taylor, F. J. R. (1976). Autogenous theories for the origin of eukaryotes. *Taxon* 25, 377-390.

Teysseire, N., Chiche-Portiche, C., and Raoult, D. (1992). Intracellular movements in *Rickettsia conorii* and *R. typhi* based on actin polymerization. *Res Microbiol* 143, 821-829.

Tilney, L. G., and Portnoy, D. A. (1989). Actin filaments and the growth, movement and spread of the intracellular bacterial parasite *Listeria monocytogenes*. *J Cell Biol* 109, 1597-1608.

Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X., and Alba, M. M. (2009). Origin of primate orphan genes: a comparative genomics approach. *Mol Biol Evol* 26, 603-612.

Trevors, J. T., AND Abel D. L. (2004). Chance and necessity do not explain the origin of life. *Cell Biol Int* 28, 729-739.

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., and Gordon, J. I. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444, 1027-1031.

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richerdson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., and Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428, 37-43.

Vallery-Radot, P. (1885). *Lois Pasteur: Life and Labours, by his son in-law. Tr, from the French by Lady Claud Hamilton*. London,, Longmans, Green and Co.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y. H., and Smith, H. O. (2004). Environmental genomes shotgun sequencing of the Sargasso Sea. *Science* 304, 66-74.

Viezens, J., and Arvand, M. (2008). Simultaneous presence of two different copies of the 16S rRNA gene in *Bartonella henselae*. *Microbiology* 154, 2881-2886.

Villarreal, L. P., and De Filippis, V. R. (2000). A hypothesis for DNA viruses as the origin of eukaryotic replication proteins. *J Virol* 74, 7079-7084.

Ward, N., Rainey, F., Hedlund, B., Staley, J., Ludwig, W., and Stackebrandt, E. (2000). Comparative phylogenetic analyses of members of the order *Planctomycetales* and the division *Verrucomicrobia*: 23S rRNA gene sequence analysis supports the 16S rRNA gene sequence-derived phylogeny. *Int J Syst Evol Microbiol* 50, 1965–1972.

Werner, F., and Grohmann, D. (2011). Evolution of multisubunit RNA polymerases in the three domains of life. *Nat Rev Microbiol* 9, 85-98.

Wicks, R. (2003). *Modern French philosophy: from existentialism to postmodernism*. Onewolrd Publications, Oxford, England.

Williams, J. (1998). *Lyotard: Towars a postmodern philosophy*. Polity Press: Cambridge, UK.

Willner, D., Furlan, M., Haynes, M., Schmieder, R., Angly, F.,E., Silva, J., Tammadoni, S., Nosrat, B., Conrad, D., and Rohwer, F. (2009). Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *Plos One* 4, e7370.

Wirth, T., Hildebrand, F., Allix-Béguec, Wölbeling, F., Kubica, T., Kremer, K., van Soolingen, D., Rüsch-Gerdes, S., Locht, C., Brisse, S., Meyer, A., Supply, P., Niemann, S. (2008). Origin, spread and demography of the *Mycobacterium tuberculosis* complex. *PLoS Pathog* 4, e1000160.

Woese, C. R. (1994). There must be a prokaryote somewhere-microbiology search for itself. *Microbiol Rev* 58, 1-9.

Woese, C. R. (1998). The universal ancestor. *Proc Natl Acad Sci U S A* 95, 6854-6859.

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.

Woese, C. R., Fox, G. E., Zablen, L., Uchida, T., Bonen, L., Pechman, K., Lewis, K. B. J, and Stahl, D. (1975). Conservation of primary structure in 16 ribosomal-RNA. *Nature* 254, 83-86.

Woese, C. R., Kandler, O., and Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domain Archea, Bacteria and Eucarya. *Proc Natl Acad Sci U S A* 87, 4576-4579.

Wolf, Y. I, Aravind, L., and Koonin, E. V. (1999). *Rickettsiae* and *Chlamydiae*: evidence of horizontal gene transfer and gene exchange. *Trends Genet* 15, 173-175.

Yutin, N., Makarova, K. S., Mekhedov, S. L., Wolf, Y. I., and Koonin, E. V. (2008). The deep archaeal roots of eukaryotes. *Mol Biol Evol* 25:,1619-1630.

Zengler, K., Walcher, M., Clark, G., Haller, I., Toledo, G., Holland, T., Mathur, E. J., Woodnutt, G., Short, J. M., and Keller, M. (2005). High-throughput cultivation of microorganisms using microcapsules. *Methods Enzymol* 397, 124.130.

Zhang, G., Wang, H., Shi, J., Wang, X., Zheng, H., Wong, G. K., Clark, T., Wang, W., Wang, J., and Kang, L. (2007). Identification and characterization of insect-specific proteins by genome protein data analysis. *BMC Genomics* 8, 93.

Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., Li, X., Ding, Y., Yang, S., and Wang, W. (2008). On the origin of new genes in *Drosophila*. *Genome Res* 18, 1446-1455.

Figure Legends

Figure 1: Compartmentation in prokaryotes and eukaryotes. Compartmentation of *Gemmata obscuriglobus* (**A**) and a eukaryotic cell (**B**) is comparable.

Figure 2: Time scale of eukaryogenesis and nucleogenesis. Eukaryotes are not the only species with compartmentation. First eukaryotes emerged from an endosymbiotic event. The first nucleus appeared approximately three billion years ago in Planctomycetes and *Chlamydia*. These numbers are approximations (Bromham and Penny, 2003; Cavalier-Smith, 2004; Trevors and Abel, 2004).

Figure 3: Motility is not necessarily a virulence factor. (**A**) *Rickettsia conorii* is mobile in the cytoplasm and in the nucleus and moves quickly. Actin-based motility is associated with the Sca2 and RickA proteins (represented in red and blue circles respectively). Yellow stars are to demonstrate that the bacterium moves fast in the cytoplasm; dotted lines are to show that the bacterium can be found anywhere in the cytoplasm and even in the nucleus. (**B**) *Rickettsia typhi* is also mobile in the cytoplasm, but it moves less quickly than *R. conorii*. Its mobility is associated with the Sca2 protein. (**C**) *R. prowazekii* is completely immobile in the cytoplasm. The Sca2 and RickA proteins are absent.

Figure 4: Phylogenomic tree based on whole gene content (present/absent genes) in pathogenic and non-pathogenic *E. coli* species and *S. dysenteriae*. Two clusters are formed: one for pathogenic species (in red), and one for non-pathogenic species (in blue). Pathogenic strains are divided into 5 groups: enteropathogenic (EPEC); enterohaemorrhagic (EHEC); uropathogenic (UPEC); enterotoxigenic (ETEC); enteroaggregative (EAEC).

Figure 5: Each one of the four domains of life, (**A**) Eukaryotes (in yellow), (**B**) Archaea (in blue), (**C**) Viruses (in pink) and (**D**) Bacteria (in green), is represented as mosaics containing genes from all four domains. Purple squares represent ORFan genes.

Annexe 2

Phylogenomic analysis of *Odyssella thessalonicensis* fortifies the common origin of *Rickettsiales*, *Pelagibacter ubique* and *Reclimonas americana* mitochondrion

Kalliopi Georgiades[1], Mohammed-Amine Madoui[1], Phuong Le[1, 2], Catherine Robert[1] and Didier Raoult[1*]

Accepted in *Plos One*

[1] Unité des Rickettsies, URMITE-CNRS UMR 6236 IRD 198 IFR 48, Faculté de Médecine, Université de la Méditerranée, 27, Bd Jean Moulin, 13005, Marseille, France

[2] Evolutionary biology and Modeling, LATP UMR CNRS 6632 FR 3098 IFR 48, University of Provence, Marseilles, France

[*] Corresponding author: Email: didier.raoult@gmail.com

Abstract

Background: The evolution of the *Alphaproteobacteria* and origin of the mitochondria are topics of considerable debate. Most studies have placed the mitochondria ancestor within the *Rickettsiales* order. Ten years ago, the bacterium *Odyssella thessalonicensis* was isolated from *Acanthamoeba spp*., and the 16S rDNA phylogeny placed it within the *Rickettsiales*. Recently, the whole genome of *O. thessalonicensis* has been sequenced, and 16S rDNA phylogeny and more robust and accurate phylogenomic analyses have been performed with 65 highly conserved proteins.

Methodology/Principal findings: The results suggested that the *O. thessalonicensis* emerged between the *Rickettsiales* and other *Alphaproteobacteria*. The mitochondrial proteins of the *Reclinomonas americana* have been used to locate the phylogenetic position of the mitochondrion ancestor within the *Alphaproteobacteria* tree. Using the K tree score method, nine mitochondrion-encoded proteins, whose phylogenies were congruent with the *Alphaproteobacteria* phylogenomic tree, have been selected and concatenated for Bayesian and Maximum Likelihood phylogenies. The *Reclinomonas americana* mitochondrion is a sister taxon to the free-living bacteria *Candidatus Pelagibacter ubique*, and together, they form a clade that is deeply rooted in the *Rickettsiales* clade.

Conclusions/Significance: The *Reclinomonas americana* mitochondrion phylogenomic study confirmed that mitochondria emerged deeply in the *Rickettsiales* clade and that they are closely related to *Candidatus Pelagibacter ubique*.

Introduction

Proteobacteria are one of the best-studied phyla within bacteria. According to the 16S rDNA phylogeny, *Proteobacteria* are subdivided into five classes: α, β, γ, δ and ε [1]. *Alphaproteobacteria* biodiversity and evolution has been well studied through phylogenetic analyses [2]. Current phylogenomic analysis allows the subdivision of the *Alphaproteobacteria* into six major orders: *Rhodospirillales, Caulobacterales, Sphingomonadales, Rickettsiales, Rhodobacterales* and *Rhizobiales*. Among them, *Sphingomonadales, Rhodobacterales* and *Rhizobiales* have a strong record of free-living organisms and are widespread in aquatic and terrestrial habitats; these organisms also have intracellular lifestyles as plant mutualists or pathogens and animal pathogens [3]. Unlike the three previous orders, *Rickettsiales* members are mostly obligate intracellular bacteria, and either parasitic, for *Rickettsia and Orientia,* or symbiotic for *Wolbachia*. Gene losses often occurred during the evolution of the intracellular species, which explains the small genome sizes of intracellular versus free-living *Alphaproteobacteria* [4]. *Wolbachia* is a special case study that lives in symbiosis with arthropods and annelids. This species shows evidence of genome reduction, but it also experienced several gene integration events from the symbiont genome to the host nuclear genome [5]. It is thought that mitochondria originated through an endosymbiotic event that occurred between the proto-*Rickettsiales* and a pro-

eukaryotic cell [6, 7]. Based on biological arguments [8], the endosymbiotic event occurred during the early stages of eukaryotic evolution approximately one billion years ago. Phylogenetic analyses have attempted to reveal the nature of the engulfed bacterium, but this remains a subject of debate [9]. Molecular phylogenomic analyses of whole mitochondrial proteins rooted the mitochondrion among the *Alphaproteobacteria* [10-12] but revealed that the heterogeneous origin of mitochondrial genes did not clearly locate the position of the mitochondrion ancestor within the *Alphaproteobacteria* tree. Studies of mitochondrial proteins that are congruent with the *Alphaproteobacteria* evolution place the mitochondrion at the root of the *Rickettsiales* order [13]. *Candidatus Pelagibacter ubique* is a marine free-living bacterium, member of the SAR11 clade, with a small genome and an AT rich genome [14] that was included in the *Rickettsiales* clade since 2007 [2] although there is still discussion on whether a free-living bacterium could be part of a clade including obligate intracellular species [15]. However, phylogenomic studies including *Candidatus Pelagibacter ubique* located the mitochondrion ancestor within the *Rickettsiales* order [2]. Furthermore, its very small and AT rich genome constitute two features that are typical of mitochondria and related obligate intracellular parasites such as the *Rickettsiales* [15]. More recent studies, on the mitochondria of *Chlamydomonas reinhardtii* [10] and *Saccharomyces cerevisiae* [16] find *Rhizobiales* and *Rhodobacterales* as sister

taxa of the mitochondria more often that *Rickettsiales*. Therefore, because of limitations in phylogenomic methods and data availability, the origin of the mitochondrial ancestor remains unclear.

Ten years ago, the intra-amoebal gram-negative bacteria, *Odyssella thessalonicensis,* was isolated from *Acanthamoeba* spp. [17]; the 16S rDNA was sequenced and phylogenetic analysis was performed. The resulting tree placed *O. thessalonicensis* in the same clade as *Paraholospora* and in a sister clade to *Rickettsiales*. It was suggested that *Holosporaceae* comprised *O. thessalonicensis, Holospora obtusa*, *NHP Bacterium* and *Caedibacter caryophilus,* and that it was within the *Rickettsiales* order.

Whole genome shotgun sequencing of *O. thessalonicensis* recently yielded genomic data on a new intracellular *Alphaproteobacteria*. In this study, we have used the sequenced *O. thessalonicensis* genome and the available alphaproteobacterial genomes to reanalyze the phylogenetic position of *O. thessalonicensis* and the evolutionary relationship between the *Alphaproteobacteria* and the *Reclimonas americana* mitochondrion which resembles the most the ancestral proto-mitochondrial genome than any other mitochondrial DNA investigated to date [18].

Results

Alphaproteobacteria 16S rDNA phylogeny

Phylogenies to recover the position of *Candidatus Pelagibacter ubique* and *O. thessalonicensis*, were built with 53 sequences of 16S rDNA, including 49 *Alphaproteobacteria* 16S rDNA sequences, comprising *O. thessalonicensis*, one *Gammaproteobacteria* (*Escherichia coli* K-12), one *Betaproteobacteria* (*Bordetella holmesii*), one *Epsilonproteobacteria* (*Sulfurimonas autotrophica*) and one *Deltaproteobacteria* (*Desulfobacterium indolicum*). Three phylogenetic methods were used: Maximum Likelihood (ML), Maximum Parsimony (MP) and Neighbor Joining (NJ). All three methods yielded the same topology, although branches were better supported by ML and MP methods. The monophyly of *Rhodospirillales* was not supported **(Figure 1A)**. Instead, this order was split into two clades corresponding to the *Acetobacteraceae* and *Rhodospirillaceae* families. It appears that *O. thessalonicensis* is a sister taxon to the clade formed by *Caedibacter caryophilus* and the four *Acetobacteraceae* species (Bootstrap (BP) = 100). Phylogenies placed *Paraholospora* deep in the *Rickettsiales* clade (BP = 99), while *Candidatus Pelagibacter ubique* is a sister taxon to *Paraholospora*. *Candidatus Pelagibacter ubique* was grouped with *Rickettsiales.* These results were also consistent with those obtained on the phylogenetic tree realized without the *O. thessalonicensis* 16S rDNA sequence **(Figure 1B)**, however, the topology of *Candidatus Pelagibacter ubique* branching right outside the *Rickettsiales* was not well supported (BP = 76). Both ribosomal DNA

phylogenies (with or without *O. thessalonicensis*) also showed that the *Magnetococcus sp*. was the first diverging *Alphaproteobacteria*.

Alphaproteobacteria phylogenomic tree

Because the 16S rDNA does not guarantee an accurate delineation of bacterial species [4, 19, 20], we performed a phylogenomic analysis involving highly conserved proteins among 42 *Alphaproteobacteria*. We selected non-duplicated proteins in the *Alphaproteobacteria* proteomes and performed an all-against-all BLAST analysis. Proteins present in all *Alphaproteobacteria* with high-scoring segment pair lengths of more than 150 amino acids and 20% identity were selected; only 65 proteins matched these criteria. The 65 corresponding alignments were performed, conserved blocks were selected, and the resulting cured alignments were concatenated in a single 12,437 amino acid alignment and used for phylogeny construction. The ML and MP methods showed similar topologies with high branch supports, while the NJ method gave very low bootstrap values. The *O. thessalonicensis* clustered together with *Alphaproteobacteria* other than the *Rickettsiales* clades, with high support values (BP=85), even though the absence of *Holosporaceae* from the dataset does not allow a strong confirmation of this topology **(Figure 2)**. *Candidatus Pelagibacter ubique* topology as sister taxon to *Rickettsiales* however, was confirmed, as it formed a deep branch alongside *Rickettsiales* also with high support values (BP=87), and there was an early divergence

between the intracellular *Rickettsiales* and the free-living *Pelagibacter* **(Figure 2)**. The phylogenomic tree suggested that all *Alphaproteobacteria* have evolved from an ancestor located between the *Rickettsiales* clade and the other *Alphaproteobacteria*.

Mitochondrion and Alphaproteobacteria relationship

Mitochondrion-encoded proteins whose phylogeny is closest to the previous *Alphaproteobacteria* phylogenomic tree were selected according to the K tree score method **(Table 1)** and used to place the mitochondrion within the *Alphaproteobacteria* tree. The nine best protein alignments were concatenated, and ML and Bayesian phylogenies were inferred **(Figure 3A)**. The Bayesian tree had the same topology as the ML tree, although the Bayesian tree branches were better supported. *O. thessalonicensis* was located alongside the group formed by the *Rhodospirillaceae* and the *Acetobacteraceaea* (Posterior Probability (PP) =1, BP=95), and its branch appeared early in the *Alphaproteobacteria* evolution. ML and Bayesian methods showed that *R. americana* mitochondrion grouped with the free-living *Candidatus Pelagibacter ubique* (PP= 0.98, BP= 96), and that together they branched deeply alongside *Rickettsiales*. The mitochondrion phylogenomic tree also suggested an early divergence between *Candidatus Pelagibacter ubique* and the mitochondrion as shown by the length of the branches and the position of the node. Mitochondrion trees without *O.*

thessalonicensis presented the same topology for the *Candidatus Pelagibacter ubique* and the mitochondrion, but the branches were not well supported (BP= 63) **(Figure 3B)**.

Discussion

Ten years ago, 16S rDNA phylogeny studies described *O. thessalonicensis* as belonging to the *Rickettsiales* order [17]. Here, except from the 16S rDNA tree, we constructed a phylogenomic analysis, more accurate for species delineation [4], and we used the three classic inference methods (ML, MP, NJ), as well as the Bayesian approach, not used in the original paper [17]. The phylogenetic positioning of *O. thessalonicensis* within the *Rickettsiales* clade was only in part verified by the 16S rDNA phylogenetic tree that grouped *O. thessalonicensis* with the *Acetobacteraceae* and as sister taxon of the *Holosporaceae* member, *Caedibacter caryophilus*. In the original paper, the 16S rDNA datasets used were not the same, as the study [17] did not include *Acetobacteraceaea*, *Candidatus Paraholospora*, or *Candidatus Pelagibacter ubique*. Phylogenomic analyses of *Alphaproteobacteria* allowed more robust trees to be built, which help to establish a more reliable position of *O. thessalonicensis* in the *Alphaproteobacteria* family. However, we recognize that the unavailability of the *Holosporaceae* genomes may raise questions on the accuracy of the topology of *O. thessalonicensis*. Nevertheless, the phylogeny of the 65 proteins undoubtedly supports that *Candidatus*

Pelagibacter ubique emerged deeply alongside *Rickettsiales*, while the *O. thessalonicensis* branch was well supported between *Rickettsiales* and other *Alphaproteobacteria*. As previously described [2, 13], the selection of the mitochondrion proteins whose phylogeny was closest to the *Alphaproteobacteria* phylogenomic tree was a powerful approach for locating the ancestor of mitochondria. Most studies argued that mitochondria are closely related to the *Rickettsiales* order. However, recent studies on the mitochondrion of the green algae, *C. reinhardtii,* have proposed that most of its mitochondrial protein sister taxon were members of the *Rhizobiales* and the *Rhodobacterales* [11] more often than the *Rickettsiales* order, while a study on *S. cerevisiae* mitochondrion proposes that its sister taxa are more often members of the *Rhizobiales* [16]. The latter study argues the possibility that mitochondrial genomes have a mosaic structure [16], so maybe their origin and evolution is dictated by different elements according to the organism they belong to. Mitochondria are heterogenous and their genomes structure suggests possible genome fusions, addition of different elements and recombination. Different analyses using mitochondria of different organisms would give different results and comparison would not be possible (data not shown). Therefore, for this study, we decided to focus only on the origins of *R. americana* mitochondrion that resembles the most to the proto-

mitochondrion ancestor, avoiding to add noise to the phylogenies with the addition of many mitochondria.

It has been proven that adding characters while constructing phylogenetic trees increases the probability that the topology of the obtained tree is correct. The more signals are tested, the more the branches are well supported [21]. *O. thessalonicensis* is, most probably, a non-*Rickettsiales* species the closest related to *Rickettsiales*, as raised by our phylogenomic analyses, it is therefore legitimate to account the *O. thessalonicensis* genome when studying the origin of mitochondria. The addition of new data from *O. thessalonicensis* whole-genome sequencing and the mitochondrial protein selection method using the K tree score partially confirmed the results found by Williams three years ago [2], as well as the ones found by other studies supporting the grouping of the mitochondria with *Alphaproteobacteria* [12], and more specifically with *Rickettsiales* [13] and reinforced the topology presenting *Pelagibacter ubique* in the *Rickettsiales* order [2]. Further and more surprisingly, in our study, the mitochondrion branch emerged as a sister taxon of *Candidatus Pelagibacter ubique*, a result strongly supported by the chosen approach. In the studies mentioned above the mitochondrion branch does not emerge as a sister taxon of *Candidatus Pelagibacter ubique*. Differences are probably due to the different datasets used by each study that do not allow a fair comparison. Moreover, mitochondria seem to have chimeric and heterogenous structures

[15] that vary from one organism to another, introducing different results according on which mitochondrion is used in every study. There still is an important debate on whether the free-living organism *Candidatus Pelagibacter ubique* is a member of the *Rickettsiales* order or not, because it is not an intracellular species [15]. Our results link, for the first time undoubtedly, *Candidatus Pelagibacter ubique* to the *Rickettsiales* order and furthermore, with the *Reclimonas americana* mitochondrial ancestor. Trees with or without *O. thessalonicensis* present the same topology, but the positioning of *Candidatus Pelagibacter ubique* as a sister taxon of *Reclinomonas americana* mitochondrion is better supported when *O. thessalonicensis* is used for the tree reconstruction. The node presenting *Candidatus Pelagibacter ubique* and the mitochondrion as sister taxa are better supported when *O. thessalonicensis* is used **(Figure 3A,B)**. *O. thessalonicensis* data reinforce previous results.

From the *R. americana* mitochondrion phylogenomic tree, we can suggest two hypotheses. In the first, the *Rickettsiales* (also including *Candidatus Pelagibacter ubique* and the proto-mitochondrion ancestor) had a free-living common ancestor with a rather small genome. There were two endosymbiotic events, one for the *Rickettsiaceae* and one for mitochondria. In the second and more parsimonious hypothesis, there was a single facultative intracellular *Rickettsiales* common ancestor with two clades evolving into a strict

intracellular species contemporary to the emergence of eukaryotes and to the creation of proto-mitochondria. In contrast, *Candidatus Pelagibacter ubique* later evolved into a free-living form due to environmental changes that facilitated its adjustment to a relatively stable extracellular environment. *Candidatus Pelagibacter ubique* is the smallest free-living bacterium. Such a massive genome reduction can only be explained by extreme specialisation [22]. However, *Pelagibacter* ubique is a bacterium found in water everywhere in the world. Its small size may therefore witness its previous lifestyle. If its ancestor was a facultative intracellular species the genome reduction took place during its association with a proto-eukaryote **(Figure 4)**. One single endosymbiotic event is by itself complicated enough and absolutely more parsimonious than two simultaneous endosymbiotic events, so we believe that the most plausible hypothesis is the latter. Moreover, a scenario suggesting that the *Rickettsiales* ancestor became intracellular after diverging from *Candidatus Pelagibacter ubique* would not explain this species extremely small genome. Many hypotheses were described on *Rickettsiales* and mitochondria, in our study we argument on our hypothesis that was never explored before and which was well-supported by our results and by the use of data used for the first time in a study on mitochondrial origins.

The use of new data from *O. thessalonicensis* whole-genome sequencing in the reconstruction of *Alphaproteobacteria* phylogenies, strongly

confirmed the emergence of the *R. americana* mitochondrion branch between *Candidatus Pelagibacter ubique* and the other *Rickettsiales* genera. Indeed, the topology of the trees built without *O. thessalonicensis* that presented the *Candidatus Pelagibacter ubique* in the *Rickettsiales* clade was not well supported; therefore, proper determination of its taxonomy was not possible. By adding *O. thessalonicensis*, the classification of *Candidatus Pelagibacter ubique* as member of *Rickettsiales* is strongly sustained by all topologies obtained by classic phylogenetic methods, such as ML, and by the Bayesian method. Finally, positioning the *R. americana* mitochondrion ancestor within *Alphaproteobacteria* has proven that the monophyly of *Rickettsiales* and the *R. americana* mitochondrion, and the evolution of *Candidatus Pelagibacter ubique*, emerged from an intracellular to a free-living organism. Currently, the most numerous and successful extracellular bacterial species on Earth, *Candidatus Pelagibacter ubique*, which is highly dominant in both salt and fresh water worldwide [23], is actually a member of the intracellular *Rickettsiales* order.

Materials and Methods

Sequence data

O. thessalonicensis [Genome Project: ID63085] was isolated from *Acanthamoeba spp.* as described by Britles *et al*. [17]. Genomic DNA was extracted and sequenced using the same method as for *Legionella drancourtii*

[24]. The first genome assembly was performed using Newbler software (454 Life Sciences, Roche) and produced 106 contigs (20 scaffolds). Contigs were sent to the RAST platform [25] for rapid gene prediction and annotation. The 16S rDNA sequences from the 49 *Alphaproteobacteria* were extracted from the Ribosomal Database Project [26]. The *Alphaproteobacteria*, *Escherichia coli* K-12 substr. MG1655 [27] and the *Reclinomonas americana* mitochondrion proteomes [28] were downloaded from the NCBI database.

Alphaproteobacteria 16 rDNA phylogeny

The 53 16S rDNA sequences were aligned using MUSCLE [29], and conserved blocks were selected using Gblocks [30]. The curated alignments were realized and used for phylogeny construction. Phylogeny inference was constructed using three different methods, ML, MP, NJ, and a four-category gamma distribution was fit for among-site rate variation. One hundred bootstrap replicates were completed, and the resulting trees were summarized using the majority-rule consensus method. Bootstrap values were considered high when they were higher that 85. Trees were displayed using MEGA4 [31].

Alphaproteobacteria phylogenomics

We used a stringent method in order to establish a protein list that would be representative of all the *Alphaproteobacteria* used in the study. Duplicated genes were discarded from *Alphaproteobacteria* proteomes using the BLASTClust program [32] with a minimum overlap of 70% and a minimum

identity of 30%. Proteins considered as non-paralogous were then gathered and used for the cluster of orthologous group (COG) searches. An all-against-all NCBI-BLASTp search was performed on the 42 *Alphaproteobacteria* dataset. All of the proteins present in all species with an identity of 20% and a high-scoring segment pair (HSP) length over 150 amino acids were considered orthologous. Through this method, 65 clusters were identified. Corresponding proteins were aligned with MUSCLE, and conserved blocks were selected with Gblocks. The 65-curated alignments were concatenated and used for phylogeny construction. Phylogenies were constructed using three different methods, ML, MP, NJ, and 100 bootstrap replicates were sampled. *Holosporaceae* were not included because of unavailability of their proteomes. Bootstrap values were considered high when they were higher that 85. The 65 protein sequences from the *Odyssella* were submitted to the GenBank database **(File S1)**.

Mitochondrion phylogeny relationships

The 67 proteins coded by the mitochondrial DNA of *Reclinomonas americana* were compared to the *Alphaproteobacteria* proteomes using NCBI4 BLASTp. Mitochondrial proteins with the best blast hits (BBHs) and an e-value under e^{-20} were selected; 59 proteins matched these criteria. For each of the 59 successful proteins, corresponding BBHs were aligned, and an ML tree was built using PhyML [33]. Trees with 42 leaves were compared to the

Alphaproteobacteria multiprotein tree using the K tree score. Only 43 trees had 42 leaves. The nine best trees were determined according to the K tree score [34]. Mitochondrion-encoded proteins were added to each of the nine successful alignments. The conserved blocks were concatenated in a single 728-amino acid alignment, and the mitochondrion phylogeny was inferred by ML and Bayesian inference methods. For the Bayesian approach, phylogeny was performed using MrBayes [35]; the GTR matrix was used, and model parameters (gamma shape and proportion of invariant) were allowed to vary through the Markov Chain Monte Carlo Chain (MCMC). Four MCMC chains were run for one million generations and sampled every 100th generation. The first 100,000 trees were discarded, and the "sumt" command of MrBayes was used to compute the clade posterior probabilities. *Holosporaceae* were not included because of unavailability of their proteomes. Bootstrap values were considered high when they were higher that 85 and PP higher that 0.85.The trees were rendered with MEGA4.

Acknowledgements

We thank Ghislain Fournous for technical support and Christelle Forzale for her help with the Figure 4. We would also like to thank the reviewers for their strong comments.

References

1. Woese CR (1987) Bacterial evolution. Microbiol Rev 51: 221-271.

2. Williams KP, Sobral BW, Dickerman AW (2007) A robust species tree for the *Alphaproteobacteria*. J Bacteriol 189: 4578-4586.

3. Batut J, Andersson SG, O'Callaghan D (2004) The evolution of chronic infection strategies in the *alpha-proteobacteria*. Nat Rev Microbiol 2: 933-945.

4. Merhej V, El Karkouri K, Raoult D (2009) Whole genome-based phylogenetic analysis of *Rickettsiae*. Clin Microbiol Infect 15: 336-337.

5. Nikoh N, Tanaka K, Shibata F, Kondo N, Hizume M *et al*. (2008) *Wolbachia* genome integrated in an insect chromosome: evolution and fate of laterally transferred endosymbiont genes. Genome Res 18: 272-280.

6. Cavalier-Smith T (2002) The phagotrophic origin of eukaryotes and phylogenetic classification of protozoa. Int J Syst Evol Microbiol 52: 297-354.

7. Emelyanov VV (2001) *Rickettsiaceae, Rickettsia*-like endosymbionts and the origin of mitochondria. Biosc Rep 21**:** 1-17.

8. Cavalier-Smith T (2004) Only six kingdoms of life. Proc Biol Sci 271: 1251-1262.

9. Wu M, Sun LV, Vamathevan J, Riegler M, Deboy R *et al*. (2004) Phylogenomics of the reproductive parasite *Wolbachia pipientis wMel*: a streamlined genome overrun by mobile genetic elements. Plos Biol 2: e69.

10. Esser C, Ahmadinejad N, Wiegand C, Rotte C, Sebastiani F *et al*. (2004) A genome phylogeny for mitochondria among *alpha-proteobacteria* and a predominantly eubacterial ancestry of yeast nuclear genes. Mol Biol Evol 21: 1643-1660.

11. Atteia A, Adrait A, Brugiere S, Tardif M, van Lis R *et al*. (2009) A proteomic survey of *Chlamydomonas reinhardtii* mitochondria sheds new light on the metabolic plasticity of the organelle and on the nature of the alpha-proteobacterial mitochondrial ancestor. Mol Biol Evol 26: 1533-1548.

12. Esser C, Martin W (2007) Supertrees and symbiosis in eukaryote genome evolution. Trends Microbiol 15: 435-437.

13. Fitzpatrick DA, Creevey CJ, McInerney JO (2006) Genome phylogenies indicate a meaningful alpha-proteobacterial phylogeny and support a grouping of the mitochondria with the *Rickettsiales*. Mol Biol Evol 23: 74-85.

14. Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL *et al*. (2005) Genome streamlining in a cosmopolitan oceanic bacterium. Science 309: 1242-1245.

15. Rodriguez-Ezpeleta I, Embley TM (2011) Data from: The closest *Alphaproteobacteria* relative to mitochondria: insights from *Candidatus Pelagibacter ubique*. Mol Biol Evol doi:10.5061/dryad.6477p.

16. Abhishek A, Bavishi A, Bavishi A, Choudhary M (2011) Bacterial genome chimaerism and the origin of mitochondria. Can J Microbiol 57: 49-61.

17. Birtles RJ, Rowbotham TJ, Michel R, Pitcher DG, La Scola B *et al*. (2000) '*Candidatus Odyssella thessalonicensis*' gen. nov., sp. nov., an obligate intracellular parasite of *Acanthamoeba* species. Int J Syst Evol Microbiol 50: 63-72.

18. Lang BF, Burger G, O'Kelly CJ, Cedergren R, Golding GB *et al*. (1997) An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. Nature 387: 493-497.

19. Fox GE, Wisotzkey JD, Jurtshuk P (1992) How close is close- 16S ribosomal RNA sequence identity may not be sufficient to guarantee species identity. Int J Syst Bacteriol 42: 166-170.

20. Rossello-Mora R, Amann R (2001) The species concept for prokaryotes. FEMS Microbiol Rev 25: 39-67.

21. Soltis SP, Soltis ED (2003) Applying the bootstrap in phylogeny reconstruction. Statist Sci 18: 256-267.

22. Merhej V, Royer-Carenzi M, Pontarotti P, Raoult D (2009) Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. Biol Direct 4: 13.

23. Morris MB, Rappé SM, Connon SA, Vergin LK, Siebold WA *et al*. (2002) SAR11 dominates ocean surface bacterioplankton communities. Nature 420: 806-810.

24. Moliner C, Raoult D, Fournier PE (2009) Evidence that the intra-amoebal *Legionella drancourtii* acquired a sterol reductase gene from eukaryotes. BMC Res Notes 2: 51.

25. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T *et al.* (2008) The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9: 75.

26. Larsen N, Olsen GJ, Maidak BL, McCaughey MJ, Overbeek R *et al*. (1993) The ribosomal database project. Nucleic Acids Res 21: 3021-3023.

27. Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V *et al* (1997) The complete genome sequence of *Escherichia coli K-12*. Science 277: 1453-1462.

28. Lang BF, Burger G, O'Kelly CJ, Cedergren R, Golding GB *et al*. (1997) An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. Nature 387: 493-497.

29. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113.

30. Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56: 564-577.

31. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596-1599.

32. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.

33. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696-704.

34. Soria-Carrasco V, Talavera G, Igea J, Castresana J (2007) The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees. Bioinformatics 23: 2954-2956.

35. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572-1574.

36. Ogata H, Audic S, Renesto-Audiffren P, Fournier PE, Barbe V *et al*. (2001) Mechanisms of evolution in *Rickettsia conorii* and *Rickettsia prowazekii*. Science 293: 2093-2098.

Authors' contributions

Designed and supervised the study: DR. Performed the analyses: PL KG MAM. Sequenced the *O. thessalonicensis* genomic DNA: CR. Wrote the paper: PL KG MAM DR. Attributed equally to the study: KG MAM.

Competing interests

The authors declare that they have no competing interests.

Figure Legends

Figure 1. *Alphaproteobacteria* 16S rDNA phylogeny. A. A ML phylogenetic tree of 49 *Alphaproteobacteria* ribosomal DNA sequences is rooted with a non-*Alphaproteobacteria* as outgroup. B. *Alphaproteobacteria* 16S rDNA phylogeny without *Odyssella thessalonicensis*. A ML phylogenetic tree of *Alphaproteobacteria* ribosomal DNA sequences is rooted with a non-*Alphaproteobacteria* as outgroup. Bootstrap values are indicated near branches as a percentage. Different colors correspond to different orders. *Candidatus Pelagibacter ubique* and *Odyssella thessalonicensis* are shown in red.

Figure 2. Phylogenomic tree of *Alphaproteobacteria.* Phylogenomic tree of 65 concatenated highly conserved proteins representing the evolution of 42 *Alphaproteobacteria* species. Important bootstrap values are indicated near branches as a percentage. The tree is rooted on *Escherichia coli*. Different orders of *Alphaproteobacteria* are labeled by different colors. *Candidatus Pelagibacter ubique* and *Odyssella thessalonicensis* are shown in red. Some of the *Rickettsiales* species are collapsed.

Figure 3. Phylogenomic tree of *Alphaproteobacteria* and the *Reclinomonas americana* mitochondrion. A. ML and Bayesian tree of nine concatenated proteins whose phylogeny is closest to the *Alphaproteobacteria* phylogenomic tree. The *Rhodobacterales* are collapsed. B. Phylogenomic tree of 42 *Alphaproteobacteria* and the *Reclinomonas americana* mitochondrion without *Odyssella thessalonicensis*. The *Rhizobiales* and *Rhodobacterales* are collapsed. Both trees are rooted on *Escherichia coli*. Values near nodes are Bayesian posterior probabilities and ML bootstraps, respectively. Different orders of *Alphaproteobacteria* are labeled by different colors. *Candidatus Pelagibacter ubique* and *Odyssella thessalonicensis* are shown in red.

Figure 4. The hypothesis for mitochondrion and free-living *Candidatus Pelagibacter ubique* emergence. From a single facultative intracellular *Rickettsiales* common ancestor, two clades evolved into strict, intracellular species contemporary to the emergence of eukaryotes. *Candidatus Pelagibacter ubique* later evolved into a free-living form. Arrows on the left represent the 16S rRNA percentage divergence scale and the time scale in million of years. A 16S rRNA percentage divergence of 1-2% corresponds to about 50 million years [36]. The arrows on the right represent the emergence events, divergence events and endosymbiotic events.

Supporting Information

File S1: 65 *Odyssella thessalonicensis* proteins

AF1.txt, 26K

http://www.biomedcentral.com/imedia/1660338084382525/supp1.txt

Tables

Table 1. Selection of the nine mitochondrion proteins whose phylogeny is closest to the *Alphaproteobacteria* phylogenomic tree

K-score	Function
0.48375	LSU ribosomal protein L2p (L8e)
0.51416	NADH-ubiquinone oxidoreductase chain G
0.64953	SSU ribosomal protein S4p (S9e)
0.6564	Cyytochrome c-type biosynthesis protein CcmC
0.68246	LSU tribosomal protein L6p (L9e)
0.68879	NADH-ubiquinone oxidoreductase chain I
0.71266	LSU ribosomal protein L5p (L11e)
071299	Succinate deshydrogenase iron-sulfur protein
0.72572	SSU ribosomal protein S3p (S3e)

Remerciements

Je tenais à remercier

Monsieur le **Professeur Didier Raoult** pour m'avoir accueillie au sein de son laboratoire. Votre enthousiasme et votre passion pour la recherche sont une source d'inspiration qui constitue un moteur de progression et de perfectionnement de mes travaux. En me proposant des sujets aussi variés et novateurs vous m'avez permis d'élargir et d'approfondir mes connaissances, non seulement en biologie et en évolution, mais également en philosophie. Vous m'avez appris à porter un autre regard sur tout ce qu'on qualifie de dogmes.

Monsieur le **Docteur Pierre Pontarotti** pour m'avoir initiée au monde des phylogénies et de l'évolution. Pendant ces quatre dernières années vous m'avez appris les bases du travail de recherche et vous m'avez toujours aidée dans les situations difficiles, qu'elles soient d'ordre professionnel ou personnel.

Monsieur le **Professeur Jean-Louis Mège** pour m'avoir fait l'honneur d'accepter la présidence de ce jury de thèse.

Monsieur le **Professeur Xavier Nassif** et Messieurs les **Docteurs Gilbert Greub** et **Etienne Danchin** pour avoir accepté de faire partie du jury et d'évaluer mon travail de thèse.

Merci à **Vicky**, **Khalid, Adil**, **Ghislain**, **Greg**, **Phuong** et **Isa** pour leur aide précieuse et leur générosité.

Merci à **Olivier**, sans qui je n'aurais jamais eu le module de programmation, ni des arbres par milliers.Merci à tous ceux avec qui, pendant ces trois dernières années, nous avons partagé des rires, des larmes, des déceptions, du bonheur, des cigarettes, des bouteilles d'alcool, des rêves… **Manolis,** avec sa musique et ses

lactobacilles, **Nada,** et nos discussions philosophiques arrosées, **Fabrice,** qui a toujours répondu à mes questions embêtantes, **Quentin,** même si il pose toujours des questions embêtantes.

Xin, qui à force d'entendre du grec comprend toutes nos discussions avec Manolis.

Merci à **Niki** pour tous les samedis de shopping et de rigolade à la chypriote.

Merci à ma **« (grande) sœur » Elina** de supporter mes délires avec une remarquable patience depuis six ans maintenant, et pour ses conseils, même si je ne les ai pas toujours suivis, je le reconnais.

Merci à **ma famille Aixoise** pour leur soutien, leur compréhension et leur prévenance, me donnant ainsi le sentiment d'être chez moi.
Merci particulièrement à **« Elle Woods »** et **« Clark Wallace »** pour avoir si gentiment corrigé mon français.

Merci à **Frank,** l'homme qui m'a montré que même en France il y a le soleil. Tu es resté à mes côtés quand j'ai douté, quand j'ai faibli. Tu m'as appris à avoir confiance en moi… Merci de partager ma vie et de la rendre plus belle.

Merci à mon frère **Andreas**, à **Marina**, à **Antonis** et à **Zoe** d'avoir toujours cru en moi et d'avoir toujours été là, malgré la distance qui nous sépare.

Enfin, le plus grand Merci à **mes parents** grâce auxquels mon rêve de devenir « Docteur » s'est transformé en réalité.

Merci à **mon père,** qui m'a toujours demandé pas moins que le meilleur…

… **Papa** tu avais encore raison.

Merci à **ma mère** que je faisais pleurer à chaque séparation…

… **Maman** ça valait la peine.

A Evelthon,

"Keep Ithaca always in your mind.
Arriving there is what you are destined for.
But don't hurry the journey at all."