Étude de l'évolution réductive des génomes bactériens par expériences d'évolution in silico et analyses bioinformatiques

Bérénice Batut

Numéro d'ordre 2014-ISAL-0108
Année 2014

Étude de l'évolution réductive des génomes bactériens par expériences d'évolution *in silico* et analyses bioinformatiques

Thèse présentée par
Bérénice Batut

Devant
L'Institut National des Sciences Appliquées de Lyon

Pour obtenir
Le grade de Docteur

Formation doctorale
Mathématiques et Informatique (InfoMaths)

Spécialité
Informatique

Soutenance prévue le 21 Novembre 2014 devant le jury composé de :

Guillaume Achaz	Maître de conférence HDR, UMPC Paris , rapporteur
Guillaume Beslon	Professeur, INSA de Lyon, directeur de thèse
Michael Blum	Chargé de recherche HDR, CNRS, rapporteur
Vincent Daubin	Directeur de recherche, CNRS, membre invité
Carole Knibbe	Maître de conférences, Université Lyon 1, directrice de thèse
Gabriel Marais	Directeur de recherche, CNRS, directeur de thèse
Frédéric Partensky	Directeur de recherche, CNRS, examinateur
Olivier Tenaillon	Directeur de recherche, INSERM, examinateur

SIGLE	ECOLE DOCTORALE	NOM ET COORDONNEES DU RESPONSABLE
CHIMIE	**CHIMIE DE LYON** **http://www.edchimie-lyon.fr** Sec : Renée EL MELHEM Bat Blaise Pascal 3e etage 04 72 43 80 46 Insa : R. GOURDON	M. Jean Marc LANCELIN Université de Lyon – Collège Doctoral Bât ESCPE 43 bd du 11 novembre 1918 69622 VILLEURBANNE Cedex Tél : 04.72.43 13 95 directeur@edchimie-lyon.fr
E.E.A.	**ELECTRONIQUE, ELECTROTECHNIQUE, AUTOMATIQUE** **http://edeea.ec-lyon.fr** Sec : M.C. HAVGOUDOUKIAN eea@ec-lyon.fr	M. Gérard SCORLETTI Ecole Centrale de Lyon 36 avenue Guy de Collongue 69134 ECULLY Tél : 04.72.18 60.97 Fax : 04 78 43 37 17 Gerard.scorletti@ec-lyon.fr
E2M2	**EVOLUTION, ECOSYSTEME, MICROBIOLOGIE, MODELISATION** **http://e2m2.universite-lyon.fr** Sec : Safia AIT CHALAL Bat Darwin - UCB Lyon 1 04.72.43.28.91 Insa : H. CHARLES	Mme Gudrun BORNETTE CNRS UMR 5023 LEHNA Université Claude Bernard Lyon 1 Bât Forel 43 bd du 11 novembre 1918 69622 VILLEURBANNE Cédex Tél : 06.07.53.89.13 e2m2@ univ-lyon1.fr
EDISS	**INTERDISCIPLINAIRE SCIENCES-SANTE** **http://www.ediss-lyon.fr** Sec : Safia AIT CHALAL Hôpital Louis Pradel - Bron 04 72 68 49 09 Insa : M. LAGARDE Safia.ait-chalal@univ-lyon1.fr	Mme Emmanuelle CANET-SOULAS INSERM U1060, CarMeN lab, Univ. Lyon 1 Bâtiment IMBL 11 avenue Jean Capelle INSA de Lyon 696621 Villeurbanne Tél : 04.72.68.49.09 Fax :04 72 68 49 16 Emmanuelle.canet@univ-lyon1.fr
INFOMATHS	**INFORMATIQUE ET MATHEMATIQUES** **http://infomaths.univ-lyon1.fr** Sec :Renée EL MELHEM Bat Blaise Pascal 3e etage infomaths@univ-lyon1.fr	Mme Sylvie CALABRETTO LIRIS – INSA de Lyon Bat Blaise Pascal 7 avenue Jean Capelle 69622 VILLEURBANNE Cedex Tél : 04.72. 43. 80. 46 Fax 04 72 43 16 87 Sylvie.calabretto@insa-lyon.fr
Matériaux	**MATERIAUX DE LYON** **http://ed34.universite-lyon.fr** Sec : M. LABOUNE PM : 71.70 –Fax : 87.12 Bat. Saint Exupéry Ed.materiaux@insa-lyon.fr	M. Jean-Yves BUFFIERE INSA de Lyon MATEIS Bâtiment Saint Exupéry 7 avenue Jean Capelle 69621 VILLEURBANNE Cedex Tél : 04.72.43 83 18 Fax 04 72 43 85 28 Jean-yves.buffiere@insa-lyon.fr
MEGA	**MECANIQUE, ENERGETIQUE, GENIE CIVIL, ACOUSTIQUE** **http://mega.universite-lyon.fr** Sec : M. LABOUNE PM : 71.70 –Fax : 87.12 Bat. Saint Exupéry mega@insa-lyon.fr	M. Philippe BOISSE INSA de Lyon Laboratoire LAMCOS Bâtiment Jacquard 25 bis avenue Jean Capelle 69621 VILLEURBANNE Cedex Tél :04.72 .43.71.70 Fax : 04 72 43 72 37 Philippe.boisse@insa-lyon.fr
ScSo	**ScSo*** **http://recherche.univ-lyon2.fr/scso/** Sec : Viviane POLSINELLI Brigitte DUBOIS Insa : J.Y. TOUSSAINT	Mme Isabelle VON BUELTZINGLOEWEN Université Lyon 2 86 rue Pasteur 69365 LYON Cedex 07 Tél : 04.78.77.23.86 Fax : 04.37.28.04.48 viviane.polsinelli@univ-lyon2.fr

*ScSo : Histoire, Géographie, Aménagement, Urbanisme, Archéologie, Science politique, Sociologie, Anthropologie

Remerciements

Je tiens à remercier toutes les personnes qui ont participé de près ou de loin à cette thèse et m'ont soutenu durant cette période, en particulier...

... Michael Blum et Guillaume Achaz pour le temps passé à la relecture de ce manuscrit de thèse et leurs retours positifs

... Olivier Tenaillon et Frédéric Partensky pour avoir accepté de participer à mon jury

... mes directeurs de thèse qui m'ont permis de réaliser ce travail

- ... Vincent et Gabriel, dont l'émulsion intellectuelle n'est pas toujours facile à suivre mais toujours passionnante, source de nouvelles connaissances, m'ayant permis de découvrir et aimer une communauté et des méthodologies que je connaissais peu (et que j'avais rejetées durant mes études)
- ... Guillaume et Carole qui ont cru en moi depuis mon stage de Master et ont tout fait pour que je puisse continuer en thèse, qui m'ont permis de comprendre, par leur complémentarité, que le tout est plus que la somme des parties, même si leur contribution individuelle reste forte, Guillaume par son importance dans la gestion de la thèse, les conseils et la prise de recul, Carole par son implication (en particulier, quotidienne) beaucoup plus importante que ne peuvent laisser penser les documents officiels, toujours présente dans les moments de doutes surtout pour me permettre d'apprécier mon travail

... Hubert Charles sans qui je n'aurais pas eu l'allocation doctorale

... Atilla Baskurt et Dominique Mouchiroud de m'avoir accueillie dans leurs laboratoires

... le personnel administratif et informatique du LBBE, en particulier Nathalie, pour son aide

... Yann, Mathieu, Florent, Marie, Magali, Fanny, Rémi et toutes les autres personnes du 3^{e} étage et du LBBE en général pour l'accueil et les différents feuilletons

... Clothilde, Aline, Erika, Jos et tous les compagnons d'infortune des repas du midi, n'ayant pas accès à Domus

... Thomas, Héloïse, Eugénie, Murray et les différents occupants du bureau, anciennement nommé (à tort) "bureau Bigot", pour l'accueil, la défense de "ma place", les conseils et tous les bons moments

... Éric pour m'avoir donné des billes de compréhension et m'avoir suggéré certains outils ou méthodes

... les aventuriers du Beagle et plus généralement de l'Antenne, passés ou actuels, pour leur accueil chaleureux, les discussions toujours animées autour d'un café (ou d'un thé)

... Caroline pour être la maman de l'Antenne mais surtout pour les discussions sur des sujets divers et variés

... Fabien et Hugues pour les conseils toujours bienvenus

... Hédi pour les cours d'informatique en BIM et la découverte de la biologie computationnelle

... Maurizio et Sotiris dont les discussions sur les subtilités du français et de ses expressions me manquent

... Gael pour la gentillesse et la présence quand ça ne va pas

... David pour les blagues très subtiles, les superbes tee-shirts, les remontrances constructives sur l'utilisation d'*aevol* et du SVN mais surtout le soutien et l'aide à tout moment

... Marine pour le soutien, la présence au jour le jour et le fait d'être toujours à l'écoute

... Magali pour la sensibilisation écologique et humaine, les partages de déjeuner permettant de découvrir de nouvelles saveurs et surtout les moments de soutien

... mes différents cobureaux qui ont réussi à me supporter et ont rendu la thèse agréable au quotidien

  ... Mathilde, pour ces deux mois de stage agrémentés de théières pleines

  ... Jonathan, qui, même si on a partagé très peu de temps le même bureau, est toujours à l'écoute, prêt à aider et pleins de bons conseils surtout quand j'en ai eu besoin ces deux derniers mois

  ... Ilya pour avoir réussi à supporter la pression de prendre la place de Stephan

  ... Baptiste pour son stage et ce qu'il a apporté à mes analyses

  ... Yoram pour avoir accepté de faire des simulations pour moi

  ... Jules pour l'apprentissage aux belles figures et au choix des couleurs, le soutien au quotidien dans les différents phases de la thèse

  ... Stephan pour nos discussions qui m'ont permis de prendre du recul vis-à-vis des choses (elles me manquent terriblement), le thé et son rituel

... celles et ceux qui m'ont hébergé régulièrement durant cette dernière année

... celles et ceux qui ont joué le jeu des ateliers de décoration en rouleaux de papier toilette ou boites à thé en recyclage, des essais de plats sucrés-salés et de mes diverses lubies

... Nicolas pour les années BIM, les discussions de soutien, la persévérance dans le refus de sortir et le post-doc

... Élise avec qui j'ai pu partager les différentes phases de la thèse

... Alice pour les séances d'escalade-discussion et l'organisation des week-end BIM, toujours bienvenus

... Camille dont l'amitié depuis notre première année de fac est importante pour moi

... Véronique, Pascal, Romain, Marie-Lise, Anne-Claire et Amélie, la super belle famille

... Virginie dont les conseils d'une vieille thésarde ont été importants pour la gestion du quotidien des derniers mois

... Papa discret mais présent et qui m'a donné les moyens de réussir ces longues études

... Madeleine, Aloïs, Cyprien et Marie-Astrid, mes frères et sœurs géniaux (même s'ils n'en sont pas toujours conscients) dont le soutien et les taquineries durant toutes ses années mais aussi le rôle de grande sœur m'ont permis de me forger telle que je suis actuellement

... Maman, qui a eu le courage de relire et d'essayer de comprendre ce manuscrit, mais surtout qui nous a tous soutenu dans tout ce qu'on a pu entreprendre et nous a fait comprendre que tout est réalisable tant qu'on s'en donne les moyens

... Xavier qui a réussi à supporter les moments difficiles, qui a cru en moi quand je doutais, qui m'a forcé à toujours me dépasser (surtout en vélo, mais ça m'a quand même aidé pour le reste), qui a mis les vélos dans le bureau pour que je me sente moins seule dans la journée et sans qui ces années auraient été plus difficiles...

Résumé

Selon une vision populaire, l'évolution serait un processus de "progrès" qui s'accompagnerait d'un accroissement de la complexité moléculaire des êtres vivants. Cependant, les programmes de séquençage des génomes ont révélé l'existence d'espèces dont les lignées ont, au contraire, subi une réduction massive de leur génome. Ainsi, chez les cyanobactéries *Prochlorococcus* et *Pelagibacter ubique*, certaines lignées ont subi une réduction de 30% de leur génome. Une telle évolution "à rebours", dite évolution réductive, avait déjà été observée pour des bactéries endosymbiotiques, pour lesquelles la sélection naturelle n'est pas assez efficace pour éliminer les mutations délétères comme les pertes de gènes, notamment car les bactéries endosymbiotiques subissent, à chaque reproduction de leur hôte, une réduction drastique de leur taille de population. Cette explication semble peu plausible pour des cyanobactéries marines comme *Prochlorococcus* et *Pelagibacter*, qui ont un mode de vie libre et qui font partie des bactéries les plus abondantes des océans. D'autres hypothèses ont ainsi été proposées pour expliquer l'évolution réductive, comme l'adaptation à un environnement stable et pauvre en nutriments ou des forts taux de mutation, mais aucune de ces hypothèses ne semble capable d'expliquer toutes les caractéristiques génomiques observées.

Dans cette thèse, nous nous intéressons au cas de l'évolution réductive chez *Prochlorococcus*, moins étudiée que celle chez les endosymbiotes, mais pour laquelle de nombreuses séquences et des données sont disponibles. Deux approches sont utilisées pour cette étude : une approche théorique de simulation où sont testés différents scénarios évolutifs pouvant conduire à une évolution réductive et une analyse des génomes de *Prochlorococcus* dans un cadre phylogénétique où sont analysées certaines caractéristiques des changements chez *Prochlorococcus* pour déterminer les causes et les caractéristiques de l'évolution réductive.

L'évolution réductive de *Prochlorococcus* avait été étudiée principalement dans un cadre écologique, par des analyses de génomique comparative. Cependant, aucun consensus n'en est ressorti et certaines des hypothèses proposées font des prédictions différentes sur l'évolution de caractéristiques génomiques.

Comme les données génomiques résultent d'une combinaison de plusieurs mécanismes, il est difficile de comprendre l'effet isolé de l'un d'entre eux. La simulation permet d'isoler et de comprendre les différentes pressions évolutives en ayant accès à toutes les séquences y compris les séquences ancestrales. Ainsi, à l'aide d'expériences d'évolution *in silico* réalisées avec la plateforme *aevol*, nous testons dans la première partie de ce travail un

certain nombre d'hypothèses émises pour l'évolution réductive chez *Prochlorococcus* en les décomposant en onze scénarios.

Dans la seconde partie de ce travail, nous réexaminons, dans un cadre phylogénétique, l'évolution du contenu en gènes et de la longueur des gènes, mais aussi les changements d'usage des codons et des acides aminés accompagnant l'enrichissement en bases AT et surtout les pressions de sélection afin de proposer des causes possibles de l'évolution réductive chez *Prochlorococcus*.

La combinaison de ces deux approches permet finalement de proposer une histoire évolutive plausible pour expliquer l'évolution réductive chez *Prochlorococcus*.

Liste des publications personnelles

Articles parus

- **B. Batut**, C. Knibbe, G. Marais et V. Daubin (2014). Reductive genome evolution at both ends of bacterial population size spectrum. *Nature Review Microbiology*, 12, 841-850.
- **B. Batut**, D.P. Parsons, S. Fischer, G. Beslon, et C. Knibbe (2013). *In silico* experimental evolution : a tool to test evolutionary scenarios. *BMC Bioinformatics*, 14 (Suppl 15) : S11
- G. Beslon, **B. Batut**, D.P. Parsons, D. Schneider et C. Knibbe (2013). An alife game to teach evolution of antibiotic resistance. *Proceedings of the Twelfth European Conference on the Synthesis and Simulation of Living Systems (ECAL 12)*, Taormina, Italie

Résumés des conférences

- **B. Batut**, D.P. Parsons, S. Fischer, G. Beslon, et C. Knibbe (2013). In silico experimental evolution : a tool to test evolutionary scenarios. 11th Recomb - Comparative Genomics, Villeurbanne, France. Communication orale
- **B. Batut**, M. Dumond, G. Marais, G. Beslon, C. Knibbe (2012). Simulating evolutionary scenarios to test whether they can induce reductive evolution. Annual Conference of Society for Molecular Biology and Evolution, Dublin, Irlande. Poster.

Table des matières

Table des figures

Avant-propos

L'évolution des êtres vivants repose sur trois mécanismes : (i) leur capacité à se reproduire en transmettant leur information génétique, occasionnellement modifiée par des mutations, (ii) les différences fortuites de succès reproductif entre les individus, ou dérive génétique, et (iii) la sélection naturelle des lignées qui ont plus systématiquement un meilleur succès reproductif que les autres, étant donné le contexte écologique de la compétition pour la survie et la reproduction. Selon une vision populaire de l'évolution, celle-ci produirait des êtres vivants de plus en plus complexes, avec certainement, à l'échelle moléculaire, de plus en plus de gènes. Cette vision est cependant erronée, à double titre. Premièrement, tous les êtres vivants actuels ont le même temps d'évolution derrière eux, qu'ils nous semblent simples et "primitifs" comme des bactéries ou des vers de terre, ou complexes et "évolués" comme des primates. Deuxièmement, les programmes de séquençage ont montré que ni le nombre de gènes ni la taille totale des génomes (incluant les gènes mais aussi l'ADN non codant) ne corrèlent avec la complexité apparente des organismes. Comings résume ainsi ce paradoxe dit de la "valeur C"[1] (Comings, 1972) :

> Being a little chauvinistic toward our own species, we like to think that man is surely one of the most complicated species on earth and thus needs just about the maximum number of genes. However, the lowly liverwort has 18 times as much DNA as we, and the slimy, dull salamander known as *Amphiuma* has 26 times our complement of DNA. To further add to the insult, the unicellular *Euglena* has almost as much DNA as man.

L'évolution n'est donc pas un processus linéaire au cours duquel le nombre de gènes augmente, mais un processus buissonnant dans lequel chaque branche a son propre nombre de gènes, non corrélé à la complexité apparente de l'organisme. Plus déconcertant encore, il arrive que lors de la formation d'une nouvelle branche, celle-ci se mette à perdre des gènes que l'évolution avait mis des milliers, voire des millions d'années à construire. Une telle évolution "à rebours" est appelée évolution réductive. Elle s'observe par exemple chez les endosymbiotes, bactéries vivant au sein des cellules eucaryotes. Cet environnement particulier, où les bactéries utilisent les ressources fournies par l'hôte et ont moins de tâches à accomplir, semble être la cause de la réduction des génomes. Cependant, la

[1] En utilisant le poids de l'ADN, nommé valeur C, comme estimateur de sa taille, Thomas (1971) a montré que la valeur C n'est pas corrélée de manière évidente à la complexité des organismes.

réduction des génomes s'observe aussi dans des lignées de cyanobactéries, comme *Prochlorococcus marinus* ou *Pelagibacter ubique*, ayant des modes de vie libre dans les océans. Pourquoi les génomes au sein de ces lignées se sont-ils réduits ? Serait-ce lié à des changements d'environnement, à des évènements géologiques ou à des causes intrinsèques aux individus ?

L'objectif de cette thèse est d'analyser l'évolution réductive de génomes bactériens en se focalisant sur le cas de la cyanobactérie *Prochlorococcus*. Ce travail repose sur l'utilisation de deux méthodes complémentaires : des expériences d'évolution *in silico* à l'aide d'un modèle d'évolution de populations bactériennes pour simuler l'impact de certains changements sur les structures génomiques ; l'analyse par génomique comparative des génomes actuellement disponibles de *Prochlorococcus* afin de définir les caractéristiques de l'évolution réductive au sein des lignées de cette bactérie dans un contexte évolutif.

Ce manuscrit commence par un chapitre présentant l'évolution réductive dans les lignées bactériennes et les différentes hypothèses proposées dans la littérature pour expliquer ce phénomène, en particulier chez *Prochlorococcus*. Le manuscrit se divise ensuite en deux parties. Dans la première, nous décrivons les simulations effectuées sur des génomes virtuels pour tester les différentes hypothèses proposées dans la littérature pour expliquer l'évolution réductive. Nous expliquons pourquoi l'évolution expérimentale *in silico* est l'approche de simulation la plus appropriée pour notre problématique, puis nous détaillons le modèle utilisé et la méthodologie spécialement développée pour ces expériences. Les résultats obtenus montrent que tous les scénarios ne conduisent pas à une évolution réductive, et que ceux qui le font ne le font pas toujours avec la même dynamique et les mêmes effets sur le génome. Dans la seconde partie du manuscrit, nous présentons les différentes analyses effectuées sur les génomes réels de *Prochlorococcus*, avec l'étude du contenu en gènes, l'évolution de la longueur des gènes, les changements d'usage des codons et l'évolution des séquences et des pressions de sélection. Nous concluons en proposant une histoire évolutive plausible pour expliquer l'évolution réductive de *Prochlorococcus*.

Chapitre I

Introduction : Évolution réductive aux deux extrémités du spectre des tailles de populations bactériennes

Les génomes des organismes actuels sont le résultat d'une longue histoire évolutive. Pour comprendre les processus dirigeant cette évolution des génomes, les traits et modifications des génomes doivent être mis en relation avec l'écologie, au sens large, des organismes, ce qui est difficile et parfois contre-intuitif. Par exemple, une caractéristique intéressante mais controversée, est la taille du génome et ses déterminants. La taille des génomes des organismes unicellulaires se répartit sur plusieurs ordres de grandeur, de moins de 10^6 paires de bases pour certaines bactéries ayant établi une symbiose obligatoire avec un hôte eucaryote, jusqu'à 10^{11} paires de bases pour l'amibe *Amoebia dubia* au mode de vie libre (McGrath et Katz, 2004). Avec le développement de la génomique, il est devenu évident que la relation entre la taille des génomes et la complexité des organismes est compliquée.

Ainsi, les génomes eucaryotes couvrent une grande partie du spectre des tailles de génomes, et les plus grands génomes sont principalement composés d'ADN non codant, peut-être non fonctionnel. La taille des génomes ne corrèle donc pas en général avec le nombre de gènes ou toute autre mesure de la complexité des organismes. Des espèces proches, avec peu de différences morphologiques, peuvent ainsi avoir des tailles de génomes différant de plusieurs ordres de grandeur (Doolittle, 2013). Il est généralement supposé que les espèces n'ont pas acquis de grands génomes par nécessité, mais plutôt comme le résultat d'une invasion d'ADN non fonctionnel.

Si l'on écarte les eucaryotes pour se focaliser uniquement sur les bactéries et les archées, la situation est clairement différente. La taille du génome est systématiquement petite et fortement liée au nombre de gènes avec une fraction remarquablement uniforme d'ADN non codant (Giovannoni *et al.*, 2005). Les bactéries aux grands génomes ont des modes de vie polyvalents, et sont probablement adaptées à des environnements variables, alors que

les bactéries aux petits génomes semblent associées à des habitats plus stables. Parmi ces dernières, une variété de lignées bactériennes vie en coopération intime avec un hôte eucaryote : les symbiotes ou endosymbiotes. Les insectes, par exemple, sont souvent associés à des bactéries endosymbiotiques. Ces bactéries fournissent des ressources vitales à leur hôte, typiquement par la synthèse d'acides aminés essentiels, "en échange" de l'habitat sécurisé du cytoplasme des cellules eucaryotes. Dans un tel environnement protégé, ces bactéries peuvent perdre de nombreux gènes qui étaient nécessaires chez leurs ancêtres libres. Ces pertes de gènes peuvent expliquer dans une certaine mesure la tendance à la réduction de la taille des génomes observée chez ces bactéries. Cependant, de nombreuses caractéristiques suggèrent que cette réduction n'est pas seulement le résultat de l'adaptation à un environnement stable. En général, les génomes des bactéries endosymbiotiques montrent des traces de dégénérescence. Par exemple, de nombreux gènes impliqués dans des fonctions cellulaires fondamentales comme la réparation de l'ADN, qui sont les plus conservées chez les bactéries et considérées comme cruciales voire essentielles à la vie cellulaires, sont absents de ces génomes. De plus, les séquences des gènes conservés changent rapidement. Le taux d'évolution des séquences a donc fortement augmenté et la composition en nucléotides du génome entier s'est enrichi en bases A et T au détriment des bases G et C. La composition en acides aminés est alors touchée et la stabilité des protéines dégradée (van Ham *et al.*, 2003). Le style de vie de ces bactéries semble ainsi avoir d'importantes conséquences sur l'efficacité de la sélection naturelle dans ces génomes.

I.1 Taille efficace de population : un paramètre clé pour les dynamiques des génomes

Afin de comprendre comment les traits d'histoire de vie affectent la taille des génomes, un paramètre écologique est crucial : la taille efficace de population d'une espèce ou N_e, qui décrit la taille d'une population théorique ayant des caractéristiques idéales[1] et un niveau de diversité similaire à la population naturelle étudiée (Charlesworth, 2009). N_e indique le niveau de dérive génétique que la population naturelle subit, c'est-à-dire les changements stochastiques des fréquences alléliques[2] lorsque la population est finie. En effet, même si certains individus ont une plus grande probabilité de survie et de reproduction que d'autres car ils sont mieux adaptés à l'environnement, cela n'est qu'une probabilité. Le nombre effectif de descendants peut différer du nombre théorique attendu, car une population naturelle n'est pas de taille infinie. Il n'est pas possible, par exemple, d'avoir exactement 2.4 descendants : en pratique, ce sera 2 ou 3, selon le fruit du hasard. Ainsi, les allèles des descendants d'une population sont un échantillon de ceux des parents

[1]Une population idéalisée est généralement une population de Wright-Fisher (Wright, 1931; Fisher, 1922, 1930). C'est une population d'individus diploïdes, où les rencontres sont aléatoires et les générations de reproduction sont discrètes et non chevauchantes. Les nouveaux individus sont formés à chaque génération par l'échantillonnage aléatoire, avec remise, des gamètes produites par les parents, qui meurent immédiatement après la reproduction. Chaque parent a une probabilité équivalente de contribuer à un individu de la génération suivante.

[2]Un allèle est l'une des multiples version d'un même gène ou d'un même locus génétique.

et le hasard joue un rôle important dans la détermination de cet échantillon. Plus la taille de la population est petite, plus l'effet de l'échantillonnage aléatoire est important, entraînant la disparition possible de variants et une réduction de la diversité génétique. La dérive génétique augmente aussi la probabilité de fixation dans la population de mutations par hasard, en particulier les mutations délétères. En effet, la probabilité de fixation d'un variant dépend du produit $N_e s$ avec s le coefficient de sélection[1] du variant. Quand $|N_e s| < 1$, le variant est quasiment neutre et sa fixation dépend plus de la dérive génétique que de la sélection (Ohta, 1972). Dans les espèces avec un faible N_e, $N_e s$ aura naturellement tendance à être faible, la dérive est alors forte et la sélection peu efficace.

Michael Lynch et ses collègues suggèrent que N_e a un impact décisif sur la taille des génomes. En effet, l'efficacité de la sélection pour limiter la propagation des segments d'ADN égoïstes (ou de toute séquence d'ADN qui peut à terme interférer avec la fitness de l'organisme), augmente avec la taille efficace de population (Lynch et Conery, 2003; Lynch, 2007, 2006). Les espèces avec de petites tailles efficaces de populations auraient ainsi tendance à accumuler de l'ADN non fonctionnel dans leur génome, alors que les espèces aux grandes tailles efficaces de population peuvent maintenir des petits génomes denses en séquences fonctionnelles et dépourvues de sources inutiles de mutations souvent délétères. Selon cette hypothèse, les eucaryotes multicellulaires ont évolué vers des génomes complexes, avec une forte proportion d'ADN non fonctionnel dans les régions intergéniques et introniques principalement, car leur petit N_e les empêche de maintenir des génomes compacts. Par contraste, les espèces bactériennes et les archées les plus étudiées sont supposées avoir maintenu de larges populations et ont ainsi de petits génomes principalement dépourvus d'ADN "poubelle".

Cette hypothèse, appelée parfois "hypothèse du péril mutationnel" (*mutational hazard hypothesis* ou MH) (Lynch, 2011), a eu un grand impact en génomique évolutive par son élégance et sa capacité à expliquer de nombreuses caractéristiques évolutives des génomes eucaryotes (Lynch et Conery, 2003; Lynch, 2007, 2012; Lynch et Abegg, 2010; Koonin, 2004). Elle a l'avantage de faire des prédictions claires sur l'évolution des génomes. Une de ces prédictions clés est la relation négative entre N_e, ou tout estimateur de l'efficacité de la sélection, et la taille des génomes. Bien que certaines questions liées aux taux de changements des populations et des tailles de génomes dans l'histoire évolutive puissent compliquer les motifs attendus de co-variation (Lynch, 2011; Whitney et Garland, 2010; Whitney *et al.*, 2011), la relation reste généralement acceptée pour les eucaryotes (Koonin, 2004; Boussau *et al.*, 2011). L'idée qu'un génome va grossir à cause de la dérive génétique a cependant été contestée chez les bactéries peu après que l'hypothèse MH ait été proposée (Daubin et Moran, 2004).

Certaines espèces bactériennes ont établi des relations tellement intimes avec un hôte eucaryote que leur survie et leur propagation dépendent seulement de la descendance de l'hôte. C'est le cas des bactéries endosymbiotiques comme *Buchnera aphidicola*, qui vivent à l'intérieur d'un compartiment cellulaire particulier de l'insecte hôte. *Buchnera* est alors transmise exclusivement aux descendants des insectes, sans aucune chance de contact avec

[1]Le coefficient de sélection est une mesure de la fitness relative d'un phénotype. Il compare la fitness, généralement estimée par le taux de reproduction, du phénotype par rappport à un phénotype privilégié.

d'autres lignées bactériennes. Cette transmission verticale impose des goulets d'étranglements qui affectent drastiquement la structure des populations de *Buchnera* (Mira et Moran, 2002). En effet, lors de goulets d'étranglement, seule une partie de la population est conservée et peut se reproduire. La diversité génétique est réduite, augmentant ainsi la dérive génétique et diminuant N_e. L'impact dépend du nombre d'hôtes, du nombre de cellules infectées et de l'espace disponible pour la croissance (Toft et Andersson, 2010). Plusieurs lignées bactériennes ont suivi le même chemin que *Buchnera* et sont associées de façon stable avec leur hôte, parfois depuis des centaines de millions d'années (Moran *et al.*, 2008). De façon intéressante, les génomes des endosymbiotes bactériens ont tous subi le même "syndrôme de résidence" (Andersson et Kurland, 1998), c'est-à-dire une réduction drastique de la taille du génome et du nombre de gènes, accompagnée de nombreuses traces de dégénérescence des génomes comme de forts taux d'évolutions et un contenu en bases AT extrêmement fort[1]. Le premier génome séquencé de *Buchnera* contient seulement 618 gènes, correspondant ainsi à une réduction de 80% par rapport à une bactérie proche libre comme *Escherichia coli*, et son contenu en GC est seulement de 26% par rapport au 50% de *E. coli* (Shigenobu *et al.*, 2000). Ainsi, en opposition avec l'hypothèse MH, certaines bactéries avec N_e faible et aucune opportunité de recombinaison avec d'autres lignées, une situation où l'efficacité de sélection est considérée être à son niveau le plus bas, réduisent drastiquement leur génome.

Bien que de nombreuses caractéristiques génomiques soient affectées par une balance entre dérive et sélection et donc par N_e (Lynch, 2007), N_e est souvent difficile à mesurer. En effet, la figure I.1 montre que deux estimateurs de N_e, le niveau de polymorphisme neutre et le ratio ${}^{K_s}/_{K_a}$[2], ont des tendances opposées dans leur relation à la taille des génomes (Lynch et Conery, 2003; Lynch, 2007; Kuo *et al.*, 2009). Ces deux motifs suggèrent qu'il existe des problèmes associés à la mesure de N_e, car ces estimateurs retournent des estimations composites. Le niveau de polymorphisme neutre est une méthode répandue pour mesurer N_e. Elle repose sur l'estimation de la variation génétique neutre, c'est-à-dire ne modifiant pas les protéines, ayant lieu au sein des populations. Cette mesure fournit un paramètre composite $N_e\mu$ avec μ est le taux de mutation, et malheureusement pas N_e directement (Charlesworth, 2009). Quelques difficultés résident dans cette méthode de mesure de N_e : (i) une estimation fiable du taux de mutation n'est pas facile à obtenir ; (ii) définir le périmètre d'une espèce ou d'une population (requis pour les études du polymorphisme) n'est pas aisé pour certains organismes, en particulier chez les bactéries[3] ; (iii) l'estimation de N_e avec des données de polymorphisme donne des informations seulement sur l'évolution récente car le polymorphisme observable remonte au plus à $2N_e$ générations en moyenne (Charlesworth, 2009). Ceci est un problème pour comparer N_e pour des organismes parents distants et étudier des temps évolutifs longs.

[1]Les organelles comme les mitochondries représentent un cas extrême d'une telle voie évolutive (McCutcheon et Moran, 2012; Andersson et Andersson, 1999; Andersson *et al.*, 2003; Lynch *et al.*, 2006).

[2]Le ratio ${}^{K_a}/_{K_s}$ (noté aussi ${}^{d_N}/_{d_S}$) est le ratio entre le taux de substitutions non synonymes, changeant la séquence d'acides aminés, et le taux de substitutions synonymes, ne modifiant pas la séquence d'acides aminés.

[3]Par exemple, le genre *Prochlorococcus* est composé de différents écotypes, qui pourraient ressembler à différentes populations, mais les séquences sont fortement divergentes entre ces écotypes et il n'est ainsi pas facile de dire quelles sont les populations, les sous-espèces, les espèces dans ce cas (Thompson *et al.*, 2013).

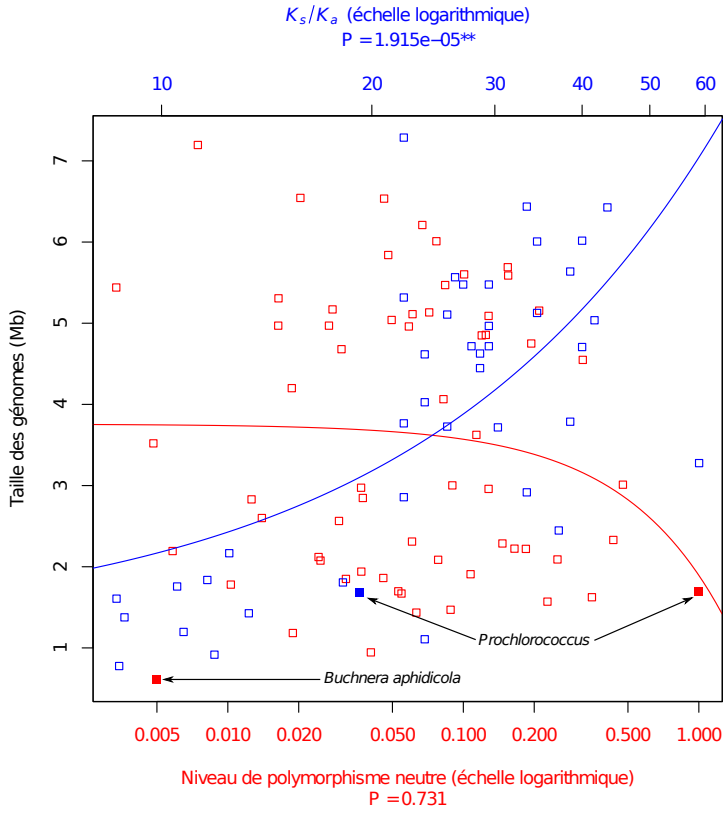


Figure I.1 – Relation entre la taille du génome et deux mesures de dérive génétique, le niveau de polymorphisme neutre et le ratio $^{K_s}/_{K_a}$ (Lynch et Conery, 2003; Lynch, 2007; Kuo *et al.*, 2009). Les données sont adaptées de Kuo *et al.* (2009) ($^{K_s}/_{K_a}$, en bleu) et de Lynch (2007) (niveau de polymorphisme neutre, en rouge).
Kuo *et al.* (2009) utilisent $^{K_a}/_{K_s}$, mais celui-ci étant lié inversement à N_e, nous avons représenté ici le ratio inverse $^{K_s}/_{K_a}$. Ainsi les deux indicateurs sont supposés croître avec N_e.
L'estimation de $^{K_s}/_{K_a}$ chez *Buchnera* n'est pas présente dans le jeu de données de Kuo *et al.* (2009) et ne peut être calculée à cause d'une saturation de K_s. Comme dans Kuo *et al.* (2009), l'estimation de $^{K_s}/_{K_a}$ de la paire de cyanobactéries de l'ordre de Nostocales est exclue du jeu de données.
Les souches de *Prochlorococcus* AS9601 et MIT9211 sont utilisées pour l'estimation $^{K_s}/_{K_a}$ (Kuo *et al.*, 2009). Comme les données de Lynch (2007) n'incluent pas *Prochlorococcus*, une précédente estimation (Lynch et Conery, 2003), basée sur les souches de *Prochlorococcus* SS120 et MIT9303, est utilisée.
Les carrés pleins représentent *Prochlorococcus* et *Buchnera* et les carrés vides les autres bactéries.
Une régression linéaire est estimée sur les données ($^{K_s}/_{K_a}$ vs taille de génome en bleu, niveau de polymorphisme vs taille de génome en rouge). Les axes des abscisses sont présentés en échelle logarithmique et les droites de régression n'apparaissent ainsi pas linéaires. Les valeurs des p-values des corrélations de Kendall sont affichées sous le nom des axes.

Une solution consiste à utiliser le ratio $^{K_a}/_{K_s}$. Cet estimateur peut être obtenu à partir de l'alignement multiple de gènes orthologues, c'est-à-dire partageant une histoire évolutive commune, de plusieurs espèces dans un contexte phylogénétique. Basé sur l'hypothèse que la sélection sur les sites synonymes est négligeable car tout changement ne modifie pas la séquence protéique, $^{K_a}/_{K_s}$ reflète l'intensité de la sélection agissant sur une protéine. Plus précisément, sous sélection négative (purificatrice), c'est-à-dire une sélection pour l'élimination rapide des variants nuisibles, $^{K_a}/_{K_s}$ décroit avec $N_e s$, où s est le coefficient de sélection d'un gène, et devrait refléter N_e d'une espèce à long terme. Mais $^{K_a}/_{K_s}$ a aussi ses difficultés : (i) $^{K_a}/_{K_s}$ donne aussi une estimation de paramètres composites ($N_e s$). La fonction d'un gène et donc son coefficient de sélection, peut changer au cours de l'évolution, rendant difficiles les comparaisons de $^{K_a}/_{K_s}$ entre des lignées ; (ii) les sites synonymes, changements de codons sans changements d'acides aminés, peuvent aussi être sous sélection pour l'utilisation de codons synonymes optimaux, et ainsi ne pas être neutres. C'est le cas dans de nombreuses espèces bactériennes. Des changements de la force de sélection sur l'usage des codons entre des lignées peuvent brouiller complètement les changements de N_e. Plus généralement, tout changement dans la façon dont les sites synonymes évoluent (incluant les changements de contenu en GC) peuvent fausser l'estimation $^{K_a}/_{K_s}$. Des modèles non homogènes incluant la possibilité de tels changements sont seulement en train d'émerger pour des estimations fiables de $^{K_a}/_{K_s}$ (Dutheil et Boussau, 2008; Guéguen *et al.*, 2013; Nielsen *et al.*, 2007; Yang et Nielsen, 2008; Pouyet *et al.*, 2013).

Malgré ces problèmes liés à l'estimation de N_e, l'idée que la dérive génétique favorise la réduction des génomes chez les bactéries, comme proposé par Kuo *et al.* (2009), idée basée sur $^{K_s}/_{K_a}$, semble plus ou moins acceptée (Lynch, 2011), au contraire de l'hypothèse MH. En effet, la relation positive entre $^{K_s}/_{K_a}$ et taille des génomes n'est pas seulement une conséquence de la présence des endosymbiotes, elle s'applique aussi aux organismes au mode de vie libre (Kuo *et al.*, 2009).

Ainsi, bien que N_e puisse être invoqué comme un facteur déterminant de l'évolution de la taille et de la complexité des génomes, ce facteur semble agir dans deux directions opposées chez les eucaryotes et les bactéries. Une différence des biais mutationnels entre eucaryotes et bactéries a été mise en avant pour expliquer ces tendances opposées (Lynch, 2011; Whitney et Garland, 2010; Kuo *et al.*, 2009). Les bactéries montrent en effet un biais vers les petites délétions (Mira *et al.*, 2001), qui pousserait les génomes à évoluer vers la réduction, alors qu'en absence d'un tel biais, les génomes eucaryotes ont une tendance à la croissance par l'invasion d'éléments égoïstes.

Il existe cependant certaines bactéries au mode de vie libre qui brouillent la vision selon laquelle les génomes bactériens se réduisent quand N_e décroit. *Pelagibacter ubique*, une bactérie océanique hétérotrophe, a un petit génome de 1.3 Mb avec 1 354 gènes (Giovannoni *et al.*, 2005). *Pelagibacter* est une des bactéries les plus abondantes sur Terre et la dérive génétique est supposée être très faible dans des espèces avec de telles tailles de population. De façon similaire, chez *Prochlorococcus*, une cyanobactérie marine photosynthétique aussi considérée comme une des bactéries les plus abondantes, l'évolution du génome apparait réminiscente d'un syndrome de résidence. Certains écotypes de *Prochlorococcus* très abondants, maintenant considérés comme des espèces différentes (Thompson

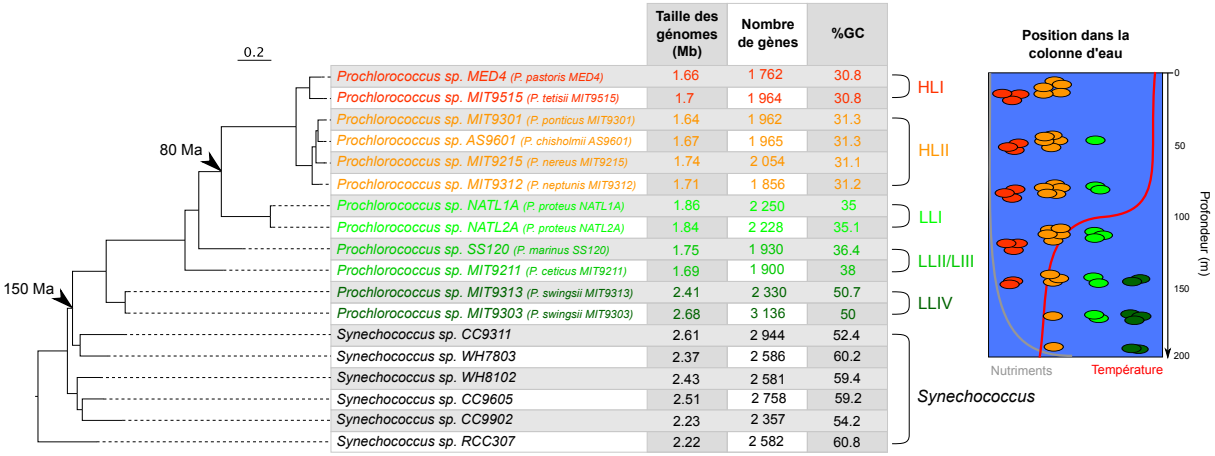


Figure I.2 – Phylogénie, statistiques des génomes et préférences écologiques des écotypes de *Prochlorococcus*
L'arbre phylogénétique a été construit en utilisant PhyML (Guindon *et al.*, 2010) sur un ensemble de gènes orthologues 1 à 1 (HOGENOM v6 (Penel *et al.*, 2009)) alignés avec Prank (Löytynoja et Goldman, 2005) (Annexe C). Les temps de divergence sont issus de Dufresne *et al.* (2005).
Les données de taille de génomes, nombre de gènes et contenu en bases GC ont été obtenues à partir des bases de données du NCBI.
La figure des données écologiques de *Prochlorococcus* est une reproduction de la figure 3 de Partensky et Garczarek (2010). Le nombre de cellules symbolise l'abondance des écotypes aux différentes profondeurs de la colonne d'eau dans les océans. Les couleurs des cellules et les caractéristiques génomiques sont liées et représentent les écotypes où des souches de *Prochlorococcus* sont trouvées : rouge "high-light" I (HLI), orange "high-light" II (HLII), vert clair "low-light" I (LLI), vert "low-light" II et "low-light" III (LLII/LLIII), vert foncé "low-light" IV (LLIV), noir *Synechococcus*. Des génomes réduits sont trouvés dans toutes les lignées des souches de *Prochlorococcus* sauf LLIV.
Les noms entre parenthèse sont les noms d'espèces définis par Thompson *et al.* (2013)

et al., 2013), ont des tailles de génomes réduits avec de nombreuses pertes de gènes, une réduction drastique du contenu en bases GC et une accélération du taux d'évolution des séquences (Partensky et Garczarek, 2010) (Figure I.2). Des travaux récents suggèrent que les génomes réduits pourraient être répandus chez les bactéries marines (Giovannoni *et al.*, 2014).

Tandis qu'il existe une littérature abondante sur la réduction des génomes chez les endosymbiotes et peu de contestation du rôle clé de la dérive génétique dans leur évolution réductive (Moran *et al.*, 2008; McCutcheon et Moran, 2012), la réduction des génomes dans les bactéries libres est moins comprise et est probablement dirigée par d'autres forces évolutives que la dérive génétique. Nous nous focalisons sur *Prochlorococcus*, la bactérie marine libre où les réductions de génomes sont les plus étudiées (par rapport à *Pelagibacter*), et dans la suite de ce chapitre, nous comparons ses caractéristiques à celles des génomes réduits des endosymbiotes, principalement *Buchnera*.

I.2 Patrons de l'évolution réductive

Nous avons vu que plusieurs espèces bactériennes ont subi des évolutions réductives. Parmi celles-ci se trouvent les endosymbiotes mais aussi plusieurs bactéries marines comme *Prochlorococcus*. Nous allons voir dans cette section que ces deux types d'évolution réductive partagent des caractéristiques communes, mais qu'elles se distinguent aussi par des patrons particuliers.

I.2.1 Patrons communs entre les endosymbiotes et les micro-organismes libres

Les génomes des endosymbiotes sont généralement petits et riches en bases AT. Une comparaison avec leurs proches en terme évolutif suggère qu'il s'agit d'un motif récurrent dans les lignées endosymbiotiques, et que ces caractéristiques sont toujours associées à une augmentation des taux de changements évolutifs au niveau moléculaire (Moran *et al.*, 2009; Moran, 1996). Le cas de *Buchnera aphidicola* est particulièrement emblématique à cet égard. Son génome inclut sept fois moins de gènes qu'*Escherichia coli*, un de ses plus proches parents libres, et le processus de perte de gènes semble toujours en cours dans plusieurs lignées de *Buchnera* (Moran *et al.*, 2009). De manière frappante, une comparaison similaire des écotypes de *Prochlorococcus* avec des génomes réduits et non réduits montre des pertes de gènes jusqu'à 43%. Étant donné le faible ratio d'ADN non codant à la fois chez les endosymbiotes (Mira *et al.*, 2001) et chez *Prochlorococcus*, les gènes perdus par pseudogénéisation sont probablement ensuite éliminés du fait d'un biais spontané vers les délétions (Mira *et al.*, 2001; Kuo et Ochman, 2009). La réduction totale de la taille du génome est de 86% pour *Buchnera* et jusqu'à 38% pour les souches réduites de *Prochlorococcus*. Endosymbiotes et *Prochlorococcus* ont de faibles contenus en bases GC, avec 26% pour *Buchnera* et 30.8-38% pour les souches réduites de *Prochlorococcus*. Une réduction extrême du contenu en GC est connue pour fortement influencer le contenu en acides aminés des protéines. Chez *Buchnera*, le biais dans la composition nucléotidique a eu un impact négatif fort sur la stabilité des protéines, un handicap qui semble en partie compensé par une surexpression constitutive des protéines chaperonnes (van Ham *et al.*, 2003), dont la fonction est d'assister d'autres protéines en leur assurant un repliement tridimensionnel adéquat. Les souches réduites de *Prochlorococcus* ont subi des changements majeurs de la constitution des protéines, supposée liée à une optimisation balancée entre la stabilité et la flexibilité des protéines (Paul *et al.*, 2010). Le biais dans la composition nucléotidique semble ainsi moins nocif chez *Prochlorococcus*.

La réduction des génomes, les pertes de gènes, l'enrichissement en bases AT et l'évolution rapide des séquences sont les motifs communs aux endosymbiotes et *Prochlorococcus*. Cependant, les génomes réduits des taxons *Prochlorococcus* ont aussi des caractéristiques particulières qui les distinguent des génomes des endosymbiotes.

I.2.2 Patrons différenciant les micro-organismes libres des endosymbiotes

Chez les endosymbiotes, l'évolution réductive est supposée avoir lieu en plusieurs étapes (McCutcheon et Moran, 2012). Les symbiotes récents comme *Serratia symbiotica* présentent de nombreux pseudogènes, des éléments ADN mobiles, des grandes et de petites délétions et des réarrangements chromosomiques (McCutcheon et Moran, 2012). Les endosymbiotes les plus anciens comme *Buchnera* ont vraisemblablement atteint un certain équilibre dans l'évolution de l'architecture de leur génome. Étant donné le style de vie restreint à l'hôte et la perte de gènes codant les protéines de recombinaison (Moran *et al.*, 2008), très peu de transferts horizontaux de gènes sont attendus, et observés, dans ces endosymbiotes (van Ham *et al.*, 2003; Toft et Andersson, 2010; Tamas *et al.*, 2002). L'évolution des génomes dans les écotypes de *Prochlorococcus* réduits ressemble à celle des endosymbiotes récents (Coleman *et al.*, 2006). Comme la plupart des autres organismes marins libres, *Prochlorococcus* subit beaucoup plus d'échanges génétiques que les endosymbiotes restreints à leur hôte (Coleman *et al.*, 2006; Sullivan *et al.*, 2003, 2005). Leurs génomes incluent quelques portions du chromosome, appelées ilôts génomiques, où de nombreux échanges génétiques et de transferts de gènes peuvent avoir lieu (Coleman *et al.*, 2006; Kettler *et al.*, 2007; Luo *et al.*, 2011) ainsi que des gains et des pertes de gènes. Les catégories fonctionnelles des gènes transférés ne sont pas aléatoires (Kettler *et al.*, 2007; Luo *et al.*, 2011), suggérant que la sélection dirige les changements de répertoires géniques chez *Prochlorococcus*. Ceci contraste avec les endosymbiotes où les répertoires de gènes sont supposés décroitre principalement par dérive génétique, touchant toutes les catégories fonctionnelles.

Une autre divergence apparente entre les génomes de *Prochlorococcus* et des endosymbiotes est le ratio $^{K_a}/_{K_s}$, qui indique l'intensité de la sélection affectant les gènes codant les protéines dans une lignée. Tandis que $^{K_a}/_{K_s}$ chez les endosymbiotes est plus fort que chez *E. coli* (Clark *et al.*, 1999), les souches réduites de *Prochlorococcus* ont des $^{K_a}/_{K_s}$ plus faibles que le genre *Synechococcus*, proche phylogénétiquement (Hu et Blanchard, 2009; Sun et Blanchard, 2014; Luo *et al.*, 2011). Comme détaillé précédemment, il n'est pas cependant pas évident d'interpréter ces motifs, comme les ratio $^{K_a}/_{K_s}$ peuvent être difficiles à estimer dans certaines circonstances.

I.3 Hypothèses pour la réduction des génomes

Le syndrome de résidence chez les endosymbiotes est généralement expliqué par le cliquet de Muller (van Ham *et al.*, 2003; Moran, 1996; Wernegreen, 2002; Wernegreen et Moran, 1999). Lors de la reproduction asexuée d'un organisme, son génome entier est copié et transmis à sa descendance, y compris les éventuelles mutations délétères. La qualité génétique de la lignée se détériore, comme un cliquet dont les dents n'autorisent le mouvement que dans une seule direction (Muller, 1964). Ce processus dégénératif affecte principalement les populations non recombinantes à faible N_e (Felsenstein, 2005; Muller, 1964). Le

cliquet clique quand les génomes les moins chargés (avec $n-1$ mutations délétères) sont perdus par dérive, et que tout génome de la population possède au moins n mutations délétères. Cette accumulation de mutations délétères est rapide dans les petites populations et irréversible en l'absence de recombinaison (Muller, 1964), car seules des mutations de reversion improbables peuvent restaurer le type sauvage. Les endosymbiotes sont supposés être fortement affectés par le cliquet de Muller car leurs petites populations sont impactées par des goulets d'étranglements fréquents (Mira et Moran, 2002) et manquent souvent de la machinerie mais surtout d'opportunités de recombinaison avec d'autres bactéries. Avec le cliquet de Muller, la sélection naturelle est dépassée par la dérive, principalement dans les gènes non essentiels, qui deviennent des pseudogènes et sont ensuite éliminés via le biais spontané vers les délétions, typique des bactéries (Mira *et al.*, 2001; Kuo et Ochman, 2009). Ces gènes perdus ne peuvent être regagnés par manque de recombinaison. Même les gènes essentiels, pour lesquels la sélection est assez forte pour éviter une dégénérescence complète, peuvent toujours accumuler quelques substitutions non-synonymes délétères. Ainsi, sous l'effet du cliquet de Muller, les séquences d'un génome évoluent rapidement et les motifs mutationnels dominent la sélection, comme observé chez les endosymbiotes (van Ham *et al.*, 2003; Moran, 1996; Tamas *et al.*, 2002; Wernegreen, 2002; Pérez-Brocal *et al.*, 2006; Degnan *et al.*, 2011).

Par la suite, nous discutons si le cliquet de Muller et d'autres hypothèses peuvent expliquer l'évolution des génomes chez *Prochlorococcus*.

I.3.1 Cliquet de Muller

Les dynamiques de populations de *Prochlorococcus* sont probablement drastiquement différentes de celles des endosymbiotes bactériens. L'abondance globale moyenne annuelle de *Prochlorococcus* est estimée à $2.9 \cdot 10^{27}$ cellules (Flombaum *et al.*, 2013) et la taille efficace de population a été estimée à 10^{11} (Baumdicker *et al.*, 2012). Cette estimation est nettement supérieure à celle de $5 \cdot 10^7$ d'*E. coli* (Charlesworth et Eyre-Walker, 2006). Bien que l'estimation de N_e chez *Prochlorococcus* puisse être remise en cause à cause des hypothèses simplificatrices sur lesquelles elle repose (tailles de populations constantes le long de la phylogénie, taux constants de gains et pertes de gènes,...), il semble difficile d'affirmer que les populations de*Prochlorococcus* ont de petits N_e. De plus, les souches de *Prochlorococcus* avec les génomes réduits sont supposées être plus abondantes et avoir un plus fort taux de croissance que les souches non réduites (Johnson *et al.*, 2006; Malmstrom *et al.*, 2010), ce qui semble en contradiction avec une hypothèse où une taille réduite de population aurait entrainé une réduction des génomes. En utilisant une approche standard d'estimation du niveau de polymorphisme neutre au niveau des sites synonymes, N_e d'un écotype réduit (MIT9312) a été récemment estimé à environ 10^9 (Kashtan *et al.*, 2014).

Prochlorococcus n'est donc pas un très bon candidat pour le cliquet de Muller. En outre, contrairement à de nombreux endosymbiotes, *Prochlorococcus* a conservé la plupart de sa machinerie de recombinaison. Des évènements de transferts horizontaux de gènes ont été documentés dans les souches de *Prochlorococcus* adaptées aux fortes lumières (Kettler

et al., 2007), mais dans une proportion plus faible que ceux trouvés dans les environnements pauvres en lumière (Zhaxybayeva *et al.*, 2009). Il est donc peu probable que *Prochlorococcus* puisse être sujet au cliquet de Muller, mais des tentatives plus approfondies de mesures des paramètres clés de génétique des populations comme N_e et le taux de recombinaison des différents écotypes, à de grandes échelles évolutives[1] sont nécessaires pour éliminer définitivement cette hypothèse comme explication de la réduction des génomes chez *Prochlorococcus*.

I.3.2 Adaptation à l'environnement pauvre en nutriments

La plupart des écotypes de *Prochlorococcus* réduits se trouvent dans les eaux de surface des mers tropicales et subtropicales, qui sont connues pour être pauvres en nutriments tout au long de l'année. Peu après la découverte de ces cyanobactéries à petits génomes, il a été proposé que leur taille de génome soit une adaptation à la vie dans un environnement pauvre en nutriments (Rocap *et al.*, 2003; Dufresne *et al.*, 2005). Un plus petit génome signifie moins d'ADN dans la cellule et donc moins de besoins en azote et phosphore, deux éléments très rares dans les eaux de surface. De plus, avoir un petit génome permet un petit volume cellulaire, qui augmente le ratio surface-volume et améliore l'assimilation des nutriments (Giovannoni *et al.*, 2005; Dufresne *et al.*, 2005). En outre, l'environnement de *Prochlorococcus*, en plus d'être pauvre en nutriments, est très stable toute l'année. Les conditions dans les eaux tropicales et subtropicales ne changent pas significativement, au contraire des eaux tempérées, et les concentrations en nutriments dans les eaux de surface sont constamment faibles. Dans un tel environnement stable, une machinerie de régulation sophistiquée n'est pas indispensable. Ainsi, un certain nombre de gènes de régulation ont été perdus (Kettler *et al.*, 2007). Généralement, tout gène non essentiel peut être perdu tant que le bénéfice de sa perte est plus fort que son coût. Les gènes perdus chez *Prochlorococcus* ont tendance à avoir un $^{K_a}/_{K_s}$ plus faible, c'est-à-dire une efficacité de sélection plus faible, que ceux conservés chez *Synechococcus* ou les souches non réduites de *Prochlorococcus* (Sun et Blanchard, 2014). Cependant, d'autres hypothèses comme le cliquet de Muller et l'hypothèse de forts taux de mutation (expliquée dans la suite) prédisent aussi que les gènes non essentiels sont préférentiellement perdus.

Cependant, un certain nombre de pertes de gènes ne coïncident pas bien avec le schéma d'adaptation à un environnement pauvre en nutriments. En particulier, pour les souches de *Prochlorococcus* avec des génomes très réduits, la perte de nombreux gènes de réparation est déroutante car le bénéfice adaptatif de telles pertes n'est pas évident (Marais *et al.*, 2008). Curieusement, cette tendance est partagée avec *Pelagibacter ubique* où de nombreux gènes de réparation de l'ADN ont été perdus par rapport à d'autres alpha-protéobactéries (Viklund *et al.*, 2012). L'ATP étant le nucléotide le moins coûteux à produire (Rocha et Danchin, 2002), il a été suggéré que la perte de gènes réparant les mutations GC vers AT pourrait être favorisée par la sélection pour un génome moins coûteux dans un environnement pauvre en nutriments (Giovannoni *et al.*, 2005, 2014). Ceci pourrait

[1] La plupart de la réduction des génomes semble avoir eu lieu dans l'évolution primitive de *Prochlorococcus*

expliquer la perte des gènes impliqués dans la réparation des mutations GC vers AT dans certains génomes de *Prochlorococcus* (Partensky et Garczarek, 2010). Cependant, chez *P. ubique*, un des seuls gènes de réparations retenus est impliqué dans la réparation des mésappariements G :U, et plus généralement chez les alpha-protéobactéries, la corrélation est faible entre les gènes réparant les mutations GC vers AT et le contenu en GC génomique (Viklund *et al.*, 2012).

L'hypothèse adaptative pour la réduction des génomes prédit aussi une corrélation entre la réduction des génomes et les niches écologiques, c'est-à-dire, dans le cas de *Prochlorococcus*, la profondeur à laquelle les bactéries croissent. *Prochlorococcus* se trouve dans deux écotypes principaux (Figure I.2) : un écotype riche en lumière (*high-light* ou HL) principalement situé dans la partie haute de la colonne d'eau (0 à -100 mètres) où les nutriments sont rares, et un écotype pauvre en lumière (*low-light* ou LL) plus profond (-100 à -200 mètres) où les nutriments sont plus abondants. La plupart des sous-écotypes correspondent à l'hypothèse adaptative : les génomes non réduits sont des écotypes LL et les génomes réduits principalement HL (Figure I.2). Cependant, les lignées de *Prochlorococcus* LLII/LLIII ont des génomes réduits alors qu'ils se trouvent dans un environnement pauvre en lumière, riche en nutriments à la même profondeur que LLIV. En fait, la position phylogénétique des lignées LLII/LLIII suggère que le processus de réduction des génomes a démarré dans un environnement pauvre en lumière et antidate la colonisation de l'environnement riche en lumière. Dans ce cadre, l'hypothèse adaptative peut seulement expliquer une amplification de la réduction des génomes, mais les facteurs ayant initié le processus restent mystérieux.

L'écologie de cette cyanobactérie est toujours à l'étude et de nouvelles données pourraient modifier la vision actuelle. De meilleures annotations des génomes et des gènes gagnés et perdus pourraient renforcer l'hypothèse adaptative. Il y aussi d'autres indices soutenant l'adaptation généralisée dans les génomes de *Prochlorococcus*. Comme discuté précédemment, certains gains de gènes ont été rapportés dans les souches réduites de *Prochlorococcus* par des duplications de gènes et des transferts horizontaux de gènes depuis d'autres cyanobactéries, évènements médiés par des phages (Kettler *et al.*, 2007; Rocap *et al.*, 2003; Coleman et Chisholm, 2010). De plus, de nombreux changements dans le protéome de *Prochlorococcus* ne se reflètent pas dans ceux de *Buchnera* ou d'autres endosymbiotes. La décroissance du contenu en GC n'explique pas tous les remplacements d'acides aminés, et les propriétés physico-chimiques des protéines dans les génomes réduits de *Prochlorococcus* suggèrent qu'elles sont plus stables et flexibles que celles des génomes non réduits de *Prochlorococcus* et non maladaptées (Paul *et al.*, 2010). La surexpression constitutive des protéines chaperonnes, qui sert probablement chez les endosymbiotes de mécanisme de compensation global contre des protéines généralement instables, n'a pas été observée chez *Prochlorococcus* (Mary *et al.*, 2004). Cependant, la perte de certains gènes, comme les gènes codant pour des protéines impliquées dans la réparation de l'ADN, reste inexpliquée par cette hypothèse.

I.3.3 Hypothèse de la Reine Noire

L'hypothèse de la Reine Noire est une hypothèse récente proposée pour expliquer la réduction des génomes (Morris *et al.*, 2012). Si des tâches effectuées par certaines bactéries bénéficient à la communauté bactérienne entière, la plupart des bactéries vont perdre la capacité à effectuer ces tâches. Une minorité de bactéries est alors piégée et continue d'effectuer ces tâches, payant le prix de les faire seules. Par exemple, la production de protéines catalase-peroxidase (KatG) est le principal mécanisme de défense contre le peroxide d'hydrogène (HOOH) externe chez les cyanobactéries (Morris *et al.*, 2012). KatG est présent chez certaines souches de *Synechococcus*, mais absent de toutes les souches de *Prochlorococcus* séquencées jusqu'à présent. Il a ainsi été suggéré que dans un environnement où plusieurs espèces co-existent, *Prochlorococcus* peut bénéficier d'une protection contre les dommages oxydatifs fournie par d'autres espèces comme *Synechococcus* (Morris *et al.*, 2012). Selon l'hypothèse de la Reine Noire, la perte de KatG pourrait être adaptative au niveau individuel tant que la production du bien commun est suffisant pour toute la communauté. Ce mécanisme pourrait expliquer pourquoi certaines bactéries marines sont si difficiles à cultiver en laboratoire, à cause de la forte connectivité nutritionnelle et donc de la dépendance entre les bactéries marines (Giovannoni *et al.*, 2014) et en l'absence des autres bactéries produisant le bien commun. Il est cependant difficile de savoir combien de gènes pourraient être potentiellement sujets à l'hypothèse de la Reine Noire et dans quelle proportion cette hypothèse peut contribuer à la réduction des génomes de *Prochlorococcus* et de *Pelagibacter* (Giovannoni *et al.*, 2014). Initialement, l'hypothèse de la Reine Noire a été proposée pour expliquer des situations comme celle de *Prochlorococcus* (Morris *et al.*, 2012). Cependant, elle pourrait aussi jouer un rôle dans l'évolution de consortiums d'endosymbiotes où plusieurs endosymbiotes co-évoluent dans un hôte (McCutcheon et Moran, 2012).

I.3.4 Fort taux de mutation

L'adaptation des bactéries aux changements environnementaux peut être accélérée par une augmentation transitoire du taux de mutation (Taddei *et al.*, 1997; Tenaillon *et al.*, 1999). Des isolats bactériens avec des taux élevés de mutation sont ainsi couramment observés dans la nature. Quand un changement d'environnement a lieu, ces mutateurs peuvent acquérir plus rapidement des mutations bénéfiques dans ces nouvelles conditions et se répandre si le bénéfice de ces mutations dépasse le coût des mutations délétères engendrées. Dans ce cas, et en l'absence de recombinaison, le gène mutateur va se fixer dans la population en même temps que la mutation bénéfique.

Curieusement, les organismes mutateurs ont souvent perdu des gènes de réparation de l'ADN et évoluent rapidement, deux caractéristiques présentes dans les souches réduites de *Prochlorococcus*. Marais *et al.* (2008) ont proposé que la réduction des génomes chez *Prochlorococcus* pourrait résulter d'une augmentation des taux de mutation. En utilisant un modèle de "seuil d'erreur" où la fréquence d'équilibre d'un gène dans la population dépend de u/s avec u le taux de mutation et s le coefficient de sélection du gène, la sélection

naturelle semble incapable de maintenir les gènes non essentiels quand le taux de mutation est augmenté. La perte des gènes de réparation de l'ADN pourrait ainsi expliquer la perte importante de gènes, l'enrichissement en bases AT et l'évolution rapide, c'est-à-dire des caractéristiques qui ne sont pas clairement adaptatives. Cependant, ce scénario seul ne peut pas expliquer l'évolution réductive chez *Prochlorococcus*, car il n'explique pas l'augmentation initiale des taux de mutation. Certains auteurs ont donc combiné l'hypothèse de simplification adaptative à l'hypothèse mutatrice pour expliquer la réduction des génomes chez *Prochlorococcus* (Partensky et Garczarek, 2010).

Une question reste cependant non résolue. Dans la théorie sur les lignées mutatrices, un taux de mutation réduit sera sélectionné après un évènement adaptatif s'il n'y a pas de nouvelle mutation bénéfique à fixer. Dans ce cas, le coût des mutations délétères sélectionne clairement un taux de mutation réduit (Denamur *et al.*, 2000). Pourquoi les gènes de réparation de l'ADN n'ont-ils pas été regagnés chez *Prochlorococcus* après l'évènement adaptatif ? La forte proportion de pseudogènes dans les génomes réduits (Paul *et al.*, 2010) suggère que le processus de réduction du génome est toujours en cours. Ceci suggère que les génomes réduits ne sont pas à l'équilibre et que les taux de mutation pourraient être toujours trop élevés pour maintenir une taille de génome stable. Des données supplémentaires sur la fréquence des transferts de gènes vers les génomes réduits de *Prochlorococcus* sont probablement nécessaires pour évaluer la probabilité de réacquisition par transfert des gènes de réparation. Le taux spontané de mutation a été mesuré dans plusieurs souches de *Prochlorococcus* (Osburne *et al.*, 2011), et aucune augmentation n'a été trouvée dans les souches réduites de *Prochlorococcus*. Cependant, il reste donc à comprendre comment des bactéries différant si drastiquement dans leur répertoire de gènes de réparation peuvent avoir des taux de mutation similaires. Notons cependant que ces taux de mutation ont été estimés en observant des mutants résistant aux antibiotiques dans des cultures liquides de souches de *Prochlorococcus* préalablement non exposées à la sélection antibiotique (Osburne *et al.*, 2011), ce qui n'est pas idéal pour estimer les taux de mutation (Kissling *et al.*, 2013; Lynch *et al.*, 2011). Des méthodes moins biaisées pour mesurer les mutations (comme des expériences d'accumulation de mutations) sont ainsi nécessaires pour confirmer ce résultat.

I.4 Discussion

Les endosymbiotes et les bactéries marines libres *Prochlorococcus* et *Pelagibacter ubique* montrent des signes d'évolution réductive de leur génome, incluant un fort taux de pertes de gènes, un fort contenu en bases AT et une évolution rapide des séquences géniques (Dufresne *et al.*, 2005; Viklund *et al.*, 2012). Pourtant, ces bactéries ont des styles de vie, des écologies et des génétiques des populations extrêmement différents. Ce simple constat questionne l'hypothèse classique selon laquelle l'évolution réductive est induite par des mécanismes de dérive génétique. Les mécanismes de réduction sont-ils alors, au moins partiellement, similaires ? Or, certains motifs d'évolution sont différents pour *Prochlorococcus* : la perte de gènes n'est pas aléatoire, les protéines ne montrent pas d'indices de

maladaptation et les dynamiques des gènes de réparation de l'ADN et de recombinaison ne correspondent pas à un simple scénario de perte par dérive. Le tableau I.1 résume ainsi les points de convergence et de divergence entre les caractéristiques génomiques des endosymbiotes et de *Prochlorococcus*. Parmi les différentes hypothèses proposées dans la littérature, l'adaptation à un environnement pauvre en nutriments serait celle correspondant le mieux aux observations, mais certains motifs évolutifs restent inexpliqués avec cette hypothèse (Tableau I.1)

Bien que ce soit difficile, il parait nécessaire, pour discriminer les différentes hypothèses, d'estimer les paramètres clés décrivant la génétique des populations de *Prochlorococcus* et de *Pelagibacter* (en particulier, les taux de recombinaison et de mutation, et N_e). Par leur mode de vie restreint à l'hôte, les endosymbiotes ont des populations de petite taille, sont génétiquement isolés et ne recombinent pas. Ils ont un faible niveau de polymorphisme nucléotidique et un petit N_e (Abbot et Moran, 2002). Des analyses antérieures (Zhao et Qin, 2007; Scanlan *et al.*, 1996; Jameson *et al.*, 2008) chez *Prochlorococcus* suggèrent un fort niveau de polymorphisme nucléotidique, ce qui est attendu pour des bactéries si abondantes (Partensky *et al.*, 1999). Des mesures de la densité de cellules par litre d'eau suggèrent que ces bactéries ont des populations gigantesques (Partensky et Garczarek, 2010; Flombaum *et al.*, 2013)[1]. Mais, il est bien connu en génétique des populations que la taille de population et N_e peuvent être très différentes (Charlesworth, 2009). Il y aurait ainsi une différence de 16 ordres de grandeur entre la taille de population (Flombaum *et al.*, 2013) et l'estimation de N_e (Baumdicker *et al.*, 2012) chez *Prochlorococcus*. Cependant, ces estimations pourraient ne pas refléter l'évolution des structures des populations chez *Prochlorococcus*, qui sont supposées varier significativement entre les écotypes. Il a été récemment suggéré que le genre *Prochlorococcus*, comprenant douze souches et six écotypes principaux, pourrait être divisé en 10 espèces différentes (Thompson *et al.*, 2013). Des estimations du polymorphisme nucléotidique et N_e sont clairement nécessaires pour les différents écotypes de *Prochlorococcus*. Un tel effort a été effectué récemment par l'étude de centaines de cellules de l'écotype MIT9312 conduisant à une estimation de N_e à au moins $1.5 \cdot 10^9$ et des conclusions très intéressantes sur la diversité actuelle de cet écotype (Kashtan *et al.*, 2014). D'autres études de cette sorte sont nécessaires.

Plusieurs aspects de l'évolution des génomes, comme les motifs de gains/pertes de gènes, la taille des gènes, l'usage des codons, le contenu en GC et $^{K_a}/_{K_s}$ nécessitent des investigations plus poussées chez *Prochlorococcus* (Tableau I.1). Pour $^{K_a}/_{K_s}$, les écotypes de *Prochlorococcus* réduits ont seulement été comparés à *Synechococcus* (Hu et Blanchard, 2009; Luo *et al.*, 2011). *Synechococcus* et *Prochlorococcus* sont assez distants (plus de 150 millions d'années (Dufresne *et al.*, 2005)) et la différence de $^{K_a}/_{K_s}$ pourrait être indépendante de la réduction des génomes. De plus, les souches réduites de *Prochlorococcus* ont subi des changements majeurs dans la constitution des protéines liées à une optimisation entre la stabilité et la flexibilité (Paul *et al.*, 2010) qui ne correspondent pas bien à la forte conservation des protéines suggérée par le faible $^{K_a}/_{K_s}$. Cependant, ce faible ratio pourrait être principalement dû à une augmentation de K_s, associée aux changements de contenu GC, plutôt qu'à une décroissance de K_a. Des souches de *Prochlorococcus* réduites

[1]En utilisant la concentration de chlorophylle, les populations de *Prochlorococcus* peuvent être vues depuis l'espace.

		Motifs		Hypothèses			
		Buchnera vs *E. coli*	*Prochlorococcus* réduites vs non réduites	CM	AE	FTM	HRN
Caractéristiques globables des génomes	Taille du génome	Réduction jusqu'à 80% (Toft et Andersson, 2010; Tamas *et al.*, 2002; Wernegreen, 2002)	Réduction jusqu'à 38% (Rocap *et al.*, 2003; Dufresne *et al.*, 2005, 2003)	+	+	+	+
	ADN codant	**Proportion stable** (Moran *et al.*, 2008)	**Proportion plus forte** (Rocap *et al.*, 2003)	-	+	?	=
	%GC	Réduction jusqu'à 26% (van Ham *et al.*, 2003; Moran *et al.*, 2008; Moran, 1996; Pérez-Brocal *et al.*, 2006)	Réduction jusqu'à 30.8-38% (Rocap *et al.*, 2003; Dufresne *et al.*, 2005, 2003)	+	?	+	=
Répertoires de gènes	Nombre de gènes	Réduction jusqu'à 80% (Tamas *et al.*, 2002; Pérez-Brocal *et al.*, 2006)	Réduction jusqu'à 43% (Rocap *et al.*, 2003; Dufresne *et al.*, 2005; Luo *et al.*, 2011)	+	+	+	+
	Familles de gènes	Plus petites	Plus petites (Luo *et al.*, 2011)	+	+	+	-
	Pseudogènes	Proportion plus forte (van Ham *et al.*, 2003; Tamas *et al.*, 2002)	Proportion potentiellement plus forte (Paul *et al.*, 2010)	+	-	+	+
	Gènes de recombinaison	**Pertes** (van Ham *et al.*, 2003; Moran *et al.*, 2008)	**Quelques pertes** (Partensky et Garczarek, 2010)	+	-	+	-
	Gènes de réparation et réplication	**Pertes** (van Ham *et al.*, 2003; Moran *et al.*, 2008; Tamas *et al.*, 2002; Pérez-Brocal *et al.*, 2006)	**Pertes** (Dufresne *et al.*, 2005, 2003; Partensky et Garczarek, 2010; Kettler *et al.*, 2007) **et Gains** (Kettler *et al.*, 2007)	+ -	- +	+ +	- -
	Gènes de régulation	Pertes (van Ham *et al.*, 2003)	Pertes (Rocap *et al.*, 2003; Dufresne *et al.*, 2003; García-Fernández *et al.*, 2004)	+	+	+	+
	Gènes métaboliques	**Pertes** (van Ham *et al.*, 2003; Pérez-Brocal *et al.*, 2006)	**Pertes** (Dufresne *et al.*, 2003; Kettler *et al.*, 2007; García-Fernández *et al.*, 2004) **et Gains** (Kettler *et al.*, 2007)	+ -	+ +	+ -	+ -
Evolution des séquences	Evolution des séquences	Plus rapide (Moran, 1996; Wernegreen et Moran, 1999; Pérez-Brocal *et al.*, 2006; Itoh *et al.*, 2002)	Plus rapide (Dufresne *et al.*, 2005; Hu et Blanchard, 2009)	+	-	+	-
	Ka/Ks	N_e **plus faible** (Moran, 1996; Tamas *et al.*, 2002; Clark *et al.*, 1999)	N_e **plus large ?** (Hu et Blanchard, 2009)	-	+	+	?
	Polymorphisme	N_e plus faible (Abbot et Moran, 2002)	Pas clair				
	Changement dans la constitution en acides aminés	**Changements délétères** (Moran *et al.*, 2008; Wernegreen et Moran, 1999; Itoh *et al.*, 2002)	**Nombreux changements probablement adaptatifs** (Paul *et al.*, 2010) **ou neutres** (Dufresne *et al.*, 2005)	-	+	?	-
Architecture des génomes	Architecture des génomes	**Statique** (van Ham *et al.*, 2003; Toft et Andersson, 2010; Tamas *et al.*, 2002)	**Non statique** (Coleman *et al.*, 2006)	+	+	+	=
	HGT, ilôts génomiques et bactériophages	**Non** (van Ham *et al.*, 2003; Toft et Andersson, 2010; Tamas *et al.*, 2002)	**Oui** (Luo *et al.*, 2011; Kettler *et al.*, 2007; Coleman *et al.*, 2006; Avrani *et al.*, 2011)	-	+	=	=
Usage des codons	Codons optimaux	Préférences plus faibles (Wernegreen et Moran, 1999)	Préférences plus faibles (Yu *et al.*, 2012)	+	?	+	=
	Motifs de gènes ARN_t	Dégénéré (Hansen et Moran, 2012)	Pas d'information				

Table I.1 – Motifs et hypothèses pour l'évolution réductive chez *Prochlorococcus*
CM : Cliquet de Muller, EA : adaptation à l'environnement pauvre en nutriments, FTM : fort taux de mutation, HRN : hypothèse de la reine noire.
Les motifs différents entre endosymbiotes et *Prochlorococcus* réduits sont en gras.
"+" et "-" respectivement indiquent les observations qui confirment et contredisent une hypothèse donnée pour l'évolution réductive chez *Prochlorococcus*. "=" symbolise une hypothèse ne faisant aucune prédiction pour un motif donné. "?" indique que des travaux théoriques supplémentaires sont nécessaires pour étudier la prédiction d'une hypothèse donnée. La bibliographie liée à chaque motif est indiquée.

et non réduites plus proches doivent être comparées en utilisant des méthodes de calcul de $^{K_a}/_{K_s}$ qui ne sont pas affectées par les fortes différences de contenu en bases GC entre les organismes. Avoir une estimation fiable de $^{K_a}/_{K_s}$ chez *Prochlorococcus* serait une clé pour déterminer si la bactérie correspond à la vision globale fournie par Kuo *et al.* (2009).

Jusqu'à présent, seules des études descriptives du contenu en bases GC et du biais d'usage des codons ont été faites chez *Prochlorococcus* (Rocap *et al.*, 2003; Dufresne *et al.*, 2005; Yu *et al.*, 2012; Dufresne *et al.*, 2003). Chez les endosymbiotes, le nombre de codons optimaux et le biais d'usage des codons sont réduits. La transcription et la traduction ont perdu en efficacité et en fidélité : les codons optimaux ne reflètent plus la composition de leur groupe d'ARN$_t$ génomique respectif et le groupe d'ARN$_t$ devient dégénéré (Hansen et Moran, 2012). Explorer plus en détail le contenu GC et l'évolution de l'usage des codons chez *Prochlorococcus* et en particulier identifier les forces évolutives en action permettrait de mieux comprendre qui de la sélection ou de la dérive dirige principalement la réduction des génomes chez *Prochlorococcus*.

La reconstruction des gains/pertes de gènes le long de la phylogénie de *Prochlorococcus* permettrait d'identifier les branches où des changements évolutifs importants ont eu lieu. Ainsi, la reconstruction faite par Kettler *et al.* (2007) suggère la possibilité que les petits génomes trouvés pour certaines souches de *Prochlorococcus* seraient dus en partie aux gains de gènes chez *Synechococcus* plutôt qu'à la perte de gènes chez *Prochlorococcus*. Une autre étude suggère que les pertes de gènes auraient principalement eu lieu chez *Prochlorococcus* peu après la divergence entre *Prochlorococcus* et *Synechococcus*, principalement à cause d'une forte sélection chez *Prochlorococcus* entrainant la perte de gènes ayant des petits effets de fitness (Sun et Blanchard, 2014). La réduction des génomes serait ainsi découplée de la diversification des souches de *Prochlorococcus*. Une troisième étude s'est intéressée à l'impact à la fois des gains et des pertes de familles de gènes, mais aussi des gains et des pertes de paralogues[1] sur l'évolution des tailles de génomes chez *Prochlorococcus* (Luo *et al.*, 2011). Dans ces trois études, la reconstruction des gains et pertes de gènes le long de l'arbre phylogénétique est basée sur une approche de maximum de parcimonie. Malgré des approches similaires pour ces trois études, les résultats obtenus et les conclusions tirées diffèrent. La principale différence est le nombre de branches de l'arbre phylogénétique où le nombre de pertes excède le nombre de gains. Ainsi pour Kettler *et al.* (2007), 3 branches sur les 27 renseignées ont un excès de pertes alors que pour Luo *et al.* (2011), ce sont 14 branches sur les 23 renseignées et 2 sur les 7 renseignées pour Sun et Blanchard (2014). De plus, les positions de ces branches dans l'arbre diffèrent. Alors que les branches avec excès de pertes sont principalement situées à la base de l'arbre des souches de *Prochlorococcus* pour Sun et Blanchard (2014), ces branches sont plus réparties le long de l'arbre pour l'étude de Luo *et al.* (2011). Les conclusions de ces trois études sont donc naturellement différentes. Avec les données de Kettler *et al.* (2007), l'évolution réductive de certaines souches de *Prochlorococcus* ne serait pas vraiment une réduction du nombre de gènes mais une augmentation du nombre de gènes pour les souches aux génomes non réduits. Sun et Blanchard (2014) concluent de leurs données que l'évolution réductive a eu lieu juste après la divergence entre *Prochlorococcus* et *Synechococcus* et ce serait arrêtée avant la

[1]Deux gènes sont dits paralogues s'ils ont acquis leur indépendance évolutive après un évènement de duplication. Des gènes paralogues font généralement parti d'une même famille de gènes.

diversification des souches de *Prochlorococcus* dans les différents écotypes. Au contraire, l'étude de Luo *et al.* (2011) montre une évolution réductive continue le long de l'arbre phylogénétique des souches réduites de *Prochlorococcus*. La reconstruction des gains et pertes de gènes avec une méthodologie différente, plus rigoureuse statistiquement comme la vraisemblance, pourrait permettre de conclure à quel moment les pertes de gènes liées à l'évolution réductive ont eu lieu.

La réduction des génomes se fait principalement par la perte de gènes, mais peut aussi affecter d'autres compartiments génomiques. Ainsi, d'après Wang *et al.* (2011), lorsque le nombre de gènes codants pour des protéines décroit, la taille des protéines décroit aussi. Ainsi, chez les endosymbiotes, il est souvent supposé que les gènes auraient eux-même subi une réduction. D'après l'hypothèse du cliquet de Muller, les gènes de *Buchnera* subiraient des pressions de sélection moins fortes que ceux d'*E. coli*. Ils pourraient ainsi refléter les biais mutationnels, comme le biais vers la délétion (Mira *et al.*, 2001; Kuo et Ochman, 2009). Les gènes des endosymbiotes seraient ainsi plus petits que ceux des bactéries libres. Cette hypothèse est confirmée par une étude sur 85 gènes de *Buchnera* (Charles *et al.*, 1999). Cependant, cette étude, datant de 1999, repose sur un faible nombre de gènes. Avec un plus grand nombre de gènes, Kenyon et Sabree (2014) ont montré que la taille des protéines des endosymbiotes ne semble pas se réduire de façon uniforme avec la réduction du génome, mais présente une grande variabilité. L'étude ne regarde cependant pas les changements de longueurs dans un contexte phylogénétique et suppute que les changements observés ont eu lieu chez les endosymbiotes et non chez les bactéries libres utilisées comme comparaison. Pour *Prochlorococcus*, la longueur des gènes a fait l'objet de deux analyses montrant l'absence de différence de longueur des gènes entre *Synechococcus* et *Prochlorococcus* (Sun et Blanchard, 2014) et entre les souches réduites et non réduites (Marais *et al.*, 2008). Cependant, ces analyses sont sommaires. Des analyses plus détaillées sont ainsi nécessaires pour étudier l'évolution de la taille des gènes le long de la phylogénie de *Prochlorococcus*.

Il semble que *Pelagibacter* et *Prochlorococcus* ont évolué vers des caractéristiques similaires de façon indépendante, faisant de ces bactéries un modèle idéal pour comprendre les forces conduisant à une évolution convergente dans ces deux lignées. Il est remarquable qu'elles partagent certaines adaptations, comme les gènes photolyase (impliqués dans la réparation des dommages ADN induits par les UV), qui ont apparemment été transférés entre *Prochlorococcus* et *Pelagibacter* (Viklund *et al.*, 2012).

Un travail théorique est aussi nécessaire pour étudier certaines hypothèses actuellement seulement verbales. Des modèles mathématiques et informatiques ont été conçus pour étudier l'évolution de la taille des génomes en général, mais peu de connexions ont été faites pour le cas spécifique de l'évolution réductive chez les endosymbiotes et les cyanobactéries libres, qui sont presque exclusivement étudiées par génomique comparative et phylogénie moléculaire. Combiner ces approches pourrait aider à résoudre le cas mystérieux de l'évolution réductive chez les bactéries libres, et, au-delà, pourrait aider au développement d'une théorie de l'évolution de la taille des génomes.

⋆⋆⋆

Ainsi, dans ce travail de thèse, nous nous attachons à revisiter à la fois les hypothèses et les patrons génomiques qui été décrits dans la littérature et résumés dans le présent chapitre. Dans la première partie du manuscrit, nous utilisons une approche de simulation, appelée évolution expérimentale *in silico*, pour revisiter les différents hypothèses proposées dans la littérature pour expliquer les phénomènes d'évolution réductive. Nous mettrons par là en évidence les limites de raisonnements seulement verbaux, qui peuvent manquer des effets secondaires ou indirects parfois importants, allant parfois jusqu'à inverser la tendance prédite. Dans la seconde partie du manuscrit, nous revisitons les analyses des génomes réels de *Prochlorococcus*, afin de dégager les caractéristiques manquantes ou incomplètes relevées ci-dessus : les gains et pertes de gènes, l'évolution de l'usage des codons, l'évolution de la longueur des gènes et la vitesse d'évolution des séquences protéiques.

Première partie

Expériences d'évolution *in silico*

Chapitre II

Comment tester les hypothèses d'évolution réductive ?

Alors que les causes de l'évolution réductive chez les endosymbiotes semblent bien caractérisées (cliquet de Muller induit par le mode de vie intracellulaire), ce n'est pas le cas pour l'évolution réductive des bactéries libres comme *Prochlorococcus*. Ainsi, de nombreuses hypothèses sont proposées dans la littérature (adaptation à un nouvel environnement, augmentation des taux de mutation, ...) mais aucune ne semble satisfaire toutes les caractéristiques génomiques des lignées réduites de *Prochlorococcus* (Tableau I.1).

Les méthodes de génomique comparative permettent d'analyser des données disponibles sur des temps évolutifs longs. Comme les données génomiques résultent d'une combinaison de plusieurs mécanismes, il est difficile de comprendre l'effet isolé de l'un d'entre eux et d'être sûr que les observations faites sont la conséquence de tel ou tel mécanisme. Afin de maîtriser et de comprendre les différentes pressions évolutives et de les étudier de manière dynamique, d'autres méthodes peuvent être utilisées comme les expériences d'évolution *in vivo* mais aussi la simulation en phylogénie moléculaire, la simulation en génétique des populations ou la simulation d'expériences d'évolution *in silico*. Ces approches permettent d'étudier des mécanismes différents sur des temps évolutifs différents (de quelques milliers de générations pour les expériences d'évolution *in vivo* à des milliards de générations pour les simulations de phylogénie moléculaire) pour répondre à des questions différentes. Nous abordons dans ce chapitre les différentes approches évoquées précédemment pour les confronter au problème qui nous intéresse : comment tester les différentes mécanismes proposés pour l'évolution réductive. Nous verrons que la simulation d'expériences d'évolution semble l'approche la plus appropriée pour répondre à cette question.

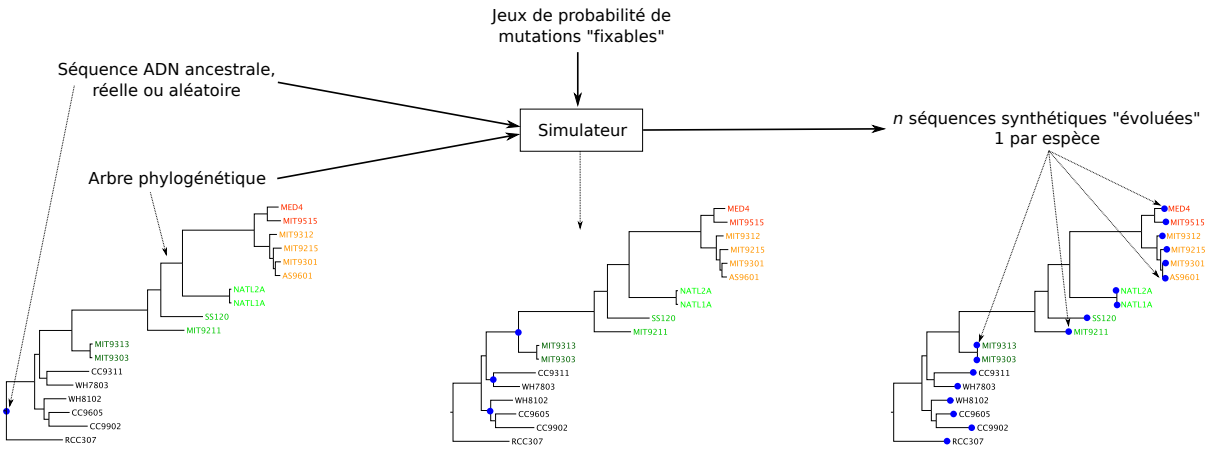


Figure II.1 – Principe des simulateurs en phylogénie moléculaire, appliqué au cas de *Prochlorococcus* et *Synechococcus*
Les simulateurs sont initialisés avec un arbre phylogénétique, comme celui de *Prochlorococcus* et *Synechococcus* et une séquence ADN ancestrale, représentée par le point bleu à la base de l'arbre phylogénétique à gauche.
Au cours des simulations, les séquences évoluent selon des jeux de probabilité de mutations "fixables" et des évènements définis en suivant l'arbre phylogénétique fourni (celui de *Prochlorococcus* et *Synechococcus*, dans cet exemple). Dans l'arbre correspondant à un instant aléatoire de la simulation, les séquences de certains nœuds (en bleu sur l'arbre au centre) sont simulées.
En fin de simulations, les séquences de toutes les feuilles de l'arbre (en bleu sur l'arbre de droite) ayant évolué à partir de la séquence ancestrale via le simulateur sont accessibles.

II.1 Simulateurs utilisés en phylogénie moléculaire

La simulation de séquences est un outil important pour tester des hypothèses biologiques mais aussi valider les méthodes de phylogénie utilisées pour reconstruire les relations entre génomes. En effet, pour des organismes réels, ces dernières sont impossibles à connaître avec certitude. Des programmes simulant l'évolution des séquences d'ADN sont ainsi utiles pour évaluer la précision des méthodes et des outils de phylogénie, en générant des jeux de tests standardisés pour les méthodes de reconstruction (*evolver* de *paml* (Yang, 2007), *seq-gen* (Rambaut et Grass, 1997)). Ces programmes peuvent aussi permettre l'investigation des processus évolutifs et des hypothèses émises quant à l'évolution de séquences lorsque l'histoire réelle des séquences est inconnue.

Dans ces programmes de simulation, seule une séquence supposée représentative de toute l'espèce est simulée, et non tous les individus, le long d'un arbre (Figure II.1). La sélection est implicitement intégrée au processus mutationnel puisque seules des mutations neutres ou favorables fixées sont prises en compte. Les vitesses d'évolution par site sont prédéfinies. Les simulateurs sont ainsi contraints par les limites de la compréhension et de l'implémentation des processus évolutifs.

Depuis *evolver* de *paml* (Yang, 1997) et *seq-gen* (Rambaut et Grass, 1997), les programmes de simulation en phylogénie moléculaire ont évolué. Ainsi, contrairement aux

programmes initiaux, *evolveAGene* (Hall, 2008), *MySSP* (Rosenberg, 2007), *Hetero* (Jermiin *et al.*, 2003), *rose* (Stoye *et al.*, 1998) simulent les insertions et les délétions en plus des mutations ponctuelles et les taux peuvent varier dans le temps. Avec *simprot* (Pang *et al.*, 2005), *dawg* (Cartwright, 2005), *indelible* (Fletcher et Yang, 2009), la formation et la distribution des insertions et des délétions deviennent plus complexes, avec même des substitutions multiples (*phylosim* (Sipos *et al.*, 2011)). Le simulateur *evolsimulator* (Beiko et Charlebois, 2007) va au-delà des simples simulations de séquences en incorporant des évènements génomiques tels les duplications et les transferts de gènes. Ce simulateur a été utilisé pour la génération de modèles nuls dans l'étude des transferts horizontaux chez les cyanobactéries dont *Prochlorococcus* (Zhaxybayeva *et al.*, 2006).

Le simulateur *ALF* (Dalquen *et al.*, 2012) a été développé dans le but de simuler la gamme complète des forces évolutives agissant sur les génomes, avec trois niveaux d'évolution. Au niveau site, les sites évoluent selon plusieurs modèles de substitutions (ADN, codons ou acides aminés), avec des changements de taux de GC mais aussi des indels. Au niveau gène, les gènes peuvent être perdus, gagnés et les tailles des familles de gènes peuvent évoluer. Au niveau génome, les génomes peuvent acquérir des gènes par transfert mais aussi subir des réarrangements. Ce simulateur pourrait être utilisé pour tester les hypothèses proposées dans la littérature, au moins pour les changements des pressions mutationnelles.

Cependant, dans *ALF* comme dans tous les simulateurs utilisés en phylogénie moléculaire, la sélection n'est pas explicite. En effet, la sélection est intégrée aux probabilités de mutations car seules les mutations fixées, et non les mutations spontanées, sont simulées. Ainsi, seules les mutations supposées *a priori* neutres ou favorables sont simulées. Par exemple, ALF interdit les mutations qui transformeraient un codon d'un gène en un codon stop, parce qu'on fait l'hypothèse qu'une telle mutation serait délétère et purgée immédiatement par la sélection naturelle. La vitesse d'évolution des différents gènes ou domaines de gènes est aussi fixée *a priori* par l'utilisateur. Dans *evolsimulator*, des probabilités spécifiques de duplication et de perte sont assignées à chaque gène par l'utilisateur. Les simulateurs en phylogénie moléculaire ne semblent donc pas adaptés à notre problématique car ils fusionnent les processus de variation et de sélection, ce qui revient à définir *a priori* une liste de mutations fixables et donc à décider à l'avance comment la sélection naturelle va agir sur le répertoire génique et les séquences des gènes. Notons également que cette fusion des processus de variation et de sélection empêche ces simulateurs de simuler correctement un scénario d'augmentation des taux spontanés de mutation. On peut augmenter le taux de mutation fixée, mais cela peut tout autant simuler une perte de gènes de réparation qu'une baisse de la pression de sélection contre les mutations.

II.2 Simulateurs utilisés en génétique des populations

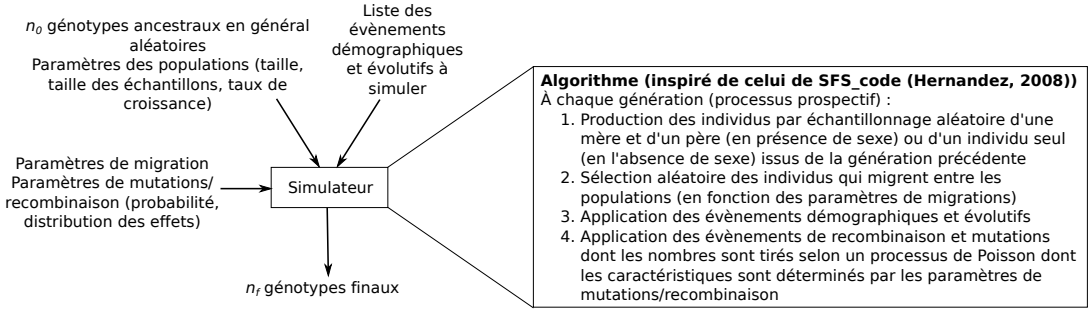
Des modèles de génétique des populations conçus pour les endosymbiotes ont montré que la pérennité d'un gène dans le génome dépend non seulement du bénéfice sélectif conféré à la bactérie, mais aussi de sa contribution à la fitness de l'hôte, de la taille de population

de l'hôte, du nombre de symbiotes transmis à la progéniture de l'hôte et du taux de mutation (Rispe et Moran, 2000; Pettersson et Berg, 2007; O'Fallon, 2008). Le cliquet de Muller, c'est-à-dire la fixation de mutations délétères comme peuvent l'être les pertes de gènes, clique plus rapidement car l'association avec l'hôte entraîne une diminution de la taille efficace de la population des bactéries. Les modèles utilisés sont très simples et des simulations à l'aide des simulateurs en génétique des populations développés ces dernières années pourraient apporter des informations supplémentaires et être aussi appliqués au cas de *Prochlorococcus*.

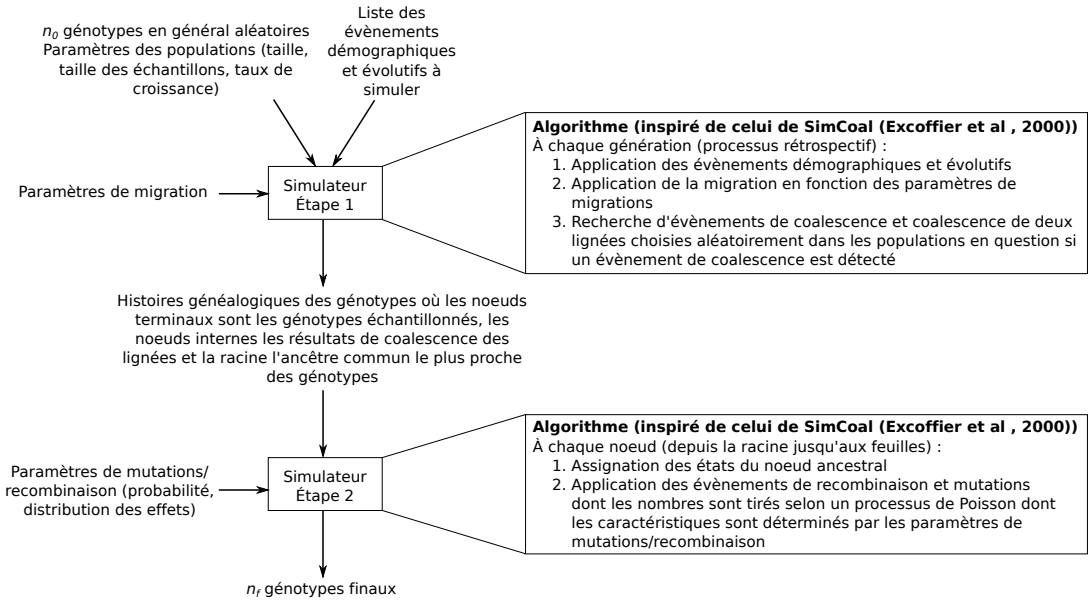
En génétique des populations, les simulateurs sont utilisés pour tester l'effet de scénarios sur la diversité moléculaire, en prenant en compte les changements au sein et entre des populations d'individus qui apparaissent, avec des modèles de sexe, de recombinaison et/ou de conversion de gènes (Hoban *et al.*, 2012). Dans ces simulateurs, tous les individus de la population sont simulés explicitement et les mutations peuvent être délétères, neutres ou favorables, mais la distribution de l'effet des mutations est définie *a priori*. Par exemple, dans un cas simple, le simulateur sera paramétré pour que, dans un gène, 60% des mutations soient délétères avec un coefficient de sélection $s = 0.005$, 38% soient neutres avec $s = 0$ et 2% soient avantageuses avec $s = 0.005$. Cette distribution peut être inférée par des expériences de mutagénèse et d'accumulation de mutations (Eyre-Walker et Keightley, 2007). Mais, c'est l'utilisateur qui détermine à l'avance la force de la sélection qui s'appliquera sur les différents gènes ou domaines de gènes. Deux approches sont implémentées dans les simulateurs : rétrospective et prospective (Figure II.2).

Les approches rétrospectives sont basées sur la théorie de la coalescence. Dans cette théorie (Kingman, 1982), sont décrites les probabilités des différentes histoires généalogiques d'un ensemble des gènes échantillonnés dans une population théorique suivant le modèle neutre de Wright-Fisher, dans lequel la taille de population est constante, les générations non chevauchantes et la rencontre entre les individus aléatoire. Dans les simulateurs basés sur une approche rétrospective, un échantillon d'allèle est suivi dans le temps jusqu'à l'ancêtre commun, et des mutations sont placées aléatoirement sur les branches de cet arbre de coalescence (Figure II.2b). Des simulateurs comme *CodonRecSim* (Anisimova *et al.*, 2003), *recodon* (Arenas et Posada, 2007) et *netcodon* (Arenas et Posada, 2010) peuvent être utilisés pour simuler des séquences codantes. Cette approche est computationnellement efficace (toute la population n'est pas requise et donc simulée) mais les scénarios simulables sont limités, principalement des changements de taille de population, des structures de population et de la migration ou de la sélection (Arenas, 2012).

Les simulateurs basés sur des approches prospectives sont centrés sur les individus (Figure II.2a). Chaque individu d'une population a un cycle de vie. Les changements démographiques et génétiques des générations suivantes sont déterminés par la génération courante et une série de matrices de transition. Ainsi, les individus participant à la génération suivante sont choisis aléatoirement selon leur fitness. Cette modélisation est plus complexe et plus adaptée à des questions prédictives que les approches rétrospectives. Des simulateurs comme *SFS_code* (Hernandez, 2008) et *GenomePop* (Carvajal-Rodríguez, 2008) permettent de simuler l'évolution de séquences codantes de façon prospective. Cependant, comme l'histoire de la population entière est simulée, cette approche est plus lourde



(a) Approche prospective



(b) Approche rétrospective

Figure II.2 – Principe des simulateurs en génétique des populations avec les deux types d'approches (prospective ou rétrospective)
Dans l'approche prospective, les simulateurs sont initiés avec des génotypes ancestraux et une série d'évènements démographiques et évolutifs à simuler. En fonction de ces évènements, des paramètres de migration et de mutation et recombinaison, les simulateurs simulent l'évolution des différents génotypes et de leur distribution dans la population jusqu'à atteindre les génotypes finaux. L'algorithme présenté est inspiré de celui de *SFS_code* (Hernandez, 2008).
Avec l'approche rétrospective, les simulateurs partent d'échantillons de génotypes et une série d'évènements démographiques et évolutifs à simuler. En fonction de ces évènements et des paramètres de migration, les simulateurs simulent la généalogie de l'échantillon en remontant dans le temps pour obtenir les histoires évolutives des échantillons. Dans une seconde étape, les simulateurs simulent l'évolution des séquences le long de ces généalogies en fonction des paramètres de mutation et recombinaison. L'algorithme présenté est inspiré de celui de *Simcoal* (Excoffier *et al.*, 2000).

computationnellement que l'approche rétrospective.

Les simulations en génétique des populations sont principalement utilisées de façon prédictive avec des suppositions spécifiques prédéfinies. Dans une revue sur les simulations des données moléculaires avec divers scénarios, Miguel Arenas propose plusieurs exemples pratiques de la simulation de séquences génétiques en utilisant des simulateurs de génétique des populations, couplés à des simulateurs en phylogénie moléculaire (Arenas, 2012). Ainsi, dans son premier exemple, il propose un scénario d'évolution des séquences nucléotidiques sous sélection naturelle, scénario souvent appliqué pour identifier les cibles de la sélection positive dans des jeux de données réels. À sa connaissance, aucun simulateur basé sur la coalescence ne permet de simuler les données sous sélection naturelle même en utilisant des modèles markoviens de substitutions de l'ADN. Il propose ainsi de combiner deux programmes. D'abord, les arbres de coalescence sont simulés avec des programmes comme *msms* (Ewing et Hermisson, 2010) ou *SelSim* (Spencer et Coop, 2004), bien que ces outils simulent un seul locus sous sélection. Ensuite, les séquences nucléotidiques évoluent le long de ces arbres avec *Seq-Gen* (Rambaut et Grass, 1997). Une autre possibilité proposée par Arenas (2012) est d'appliquer un simulateur basé sur une approche prospective qui implémente une sélection complexe et tous les modèles de substitution de l'ADN (*SFS_code* (Hernandez, 2008), par exemple).

Pour *Prochlorococcus*, les simulations en génétique des population pourraient permettre de tester l'impact de changements d'environnement ou de structure de populations. D'après Hoban *et al.* (2012), il faudrait utiliser les simulateurs *msms* si nous souhaitons tester les changements d'environnement et de sélection, *mlcoalsim* ou *mbs* pour le test du changement d'environnement seul ou *genomePop* pour tester les changements de sélection seuls. Ce dernier est le seul des quatre simulateurs proposés ayant une approche prospective, ce qui semble l'approche la plus adaptée au cas de *Prochlorococcus*. Cependant, pour utiliser les simulateurs de génétique des populations, il nous faudrait des données de populations pour les différents écotypes pour comparer les résultats des simulateurs aux données réelles. Les seules données de population disponibles à ce jour sont ceux de l'écotype de *Prochlorococcus* MIT9312, utilisés pour estimer la taille efficace de population (Kashtan *et al.*, 2014). Mais, même avec des données de populations, ces simulateurs ne pourraient être utilisés pour simuler l'évolution réductive. La plupart de ces simulateurs n'ont pas été développés dans l'optique d'être utilisés avec des populations immenses, principalement asexuées, comme celles de *Prochlorococcus*. Mais surtout, la grande majorité des simulateurs ne permettent pas à l'architecture génomique de muter. C'est donc l'inconvénient le plus important de ces simulateurs pour étudier l'évolution réductive, c'est-à-dire une réduction de la taille des génomes par la perte de gènes et de bases non codantes.

II.3 Expériences d'évolution *in vivo*

Les bactéries, comme tous les organismes vivants, évoluent par évolution darwinienne : des modifications génétiques aléatoires suivies d'une sélection des individus les plus adaptés.

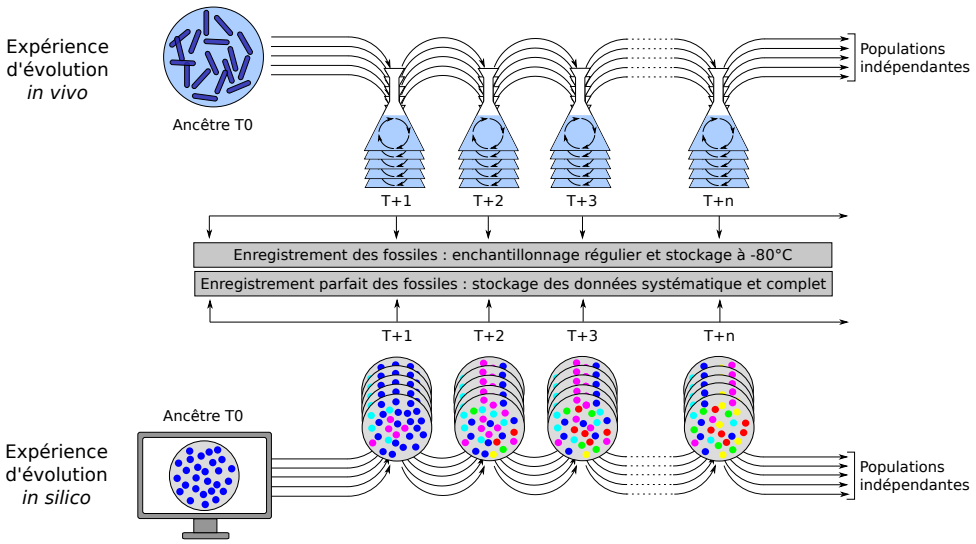


Figure II.3 – Expériences d'évolution *in vivo* (haut du graphique) et *in silico* (bas du graphique)
Les organismes ancestraux microbiens (haut) et artificiels (bas) sont propagés dans des environnements humides ou informatiques, respectivement.
Le principal avantage de ces expériences est la disponibilité d'un ancêtre et des populations évoluées qui sont échantillonnées tout au long de l'évolution. Tous les organismes vivants et artificiels sont gelés ou stockés dans des bases de données.
La figure est inspirée de la Figure 1 de Hindré *et al.* (2012)

Ce mécanisme d'évolution peut être étudié par des expériences d'évolution en laboratoire, consistant à propager des lignées d'organismes dans un environnement contrôlé (Figure II.3).

Les états ancestraux sont connus grâce à la congélation à intervalles réguliers d'une partie de la population (Figure II.3), permettant d'avoir accès, par séquençage, à la dynamique de fixation des mutations ou des réarrangements. Ainsi, l'évolution expérimentale *in vivo* permet d'apporter des connaissances sur la dynamique des processus évolutifs (Barrick *et al.*, 2009; Conrad *et al.*, 2011), l'émergence de traits complexes (Çakar *et al.*, 2005; Stanley *et al.*, 2010), le rôle des interactions épistatiques durant l'évolution (Cooper *et al.*, 2008), la reproductibilité et l'ordre de fixation des mutations (Toprak *et al.*, 2012), la quantification des paramètres évolutifs fondamentaux comme les taux de mutation, la distribution des effets de fitness (Sanjuán, 2010; Peris *et al.*, 2010; Sanjuán *et al.*, 2004; Domingo-Calap *et al.*, 2009), etc. En dépit de certaines contraintes expérimentales, cette approche permet d'étudier l'évolution en action, sans *a priori* sur la façon dont la sélection et la dérive filtrent les mutations, contrairement aux approches de phylogénie moléculaire ou de génétique des populations.

Cependant, ce type de méthodologie est coûteuse en temps. Les expériences sur des mois ou des années permettent d'avoir accès seulement à quelques centaines ou milliers de générations. Cette méthodologie semble ainsi peu adaptée à la question de l'évolution réductive. Après 25 ans, les expériences d'évolution menées sur *E. coli* par le laboratoire de R. Lenski ont atteint 60 000 générations. Ce nombre est relativement faible à l'échelle

de l'évolution et, en particulier, de l'évolution réductive. En effet, l'évolution réductive de *Buchnera* a commencé il y a 180 millions d'années (Moran *et al.*, 2008), correspondant à environ $5 \cdot 10^9$ générations[1]. De plus, bien que l'environnement soit contrôlé, des pressions évolutives différentes co-existent, compliquant l'interprétation d'une évolution réductive. Enfin, les mécanismes ayant induit une évolution réductive sont peu connus et de nombreuses espèces dont des lignées ont subi une évolution réductive ne sont pas facilement cultivables, *Pelagibacter* par exemple (Giovannoni *et al.*, 2014). Les conditions de culture de *Prochlorococcus* sont, quant à elles, complexes et diffèrent selon les écotypes (Moore *et al.*, 2007), mais il existe des collections de culture de *Prochlorococcus*, comme celles du laboratoire de cultures des cyanobactéries du MIT aux États-Unis ou celles de la station biologique de Roscoff en France.

II.4 L'évolution expérimentale *in silico*

L'évolution expérimentale *in silico* suit la même stratégie que les expériences *in vivo* mais avec des organismes digitaux (Figure II.3). Des organismes artificiels simulés dans un ordinateur sont soumis à un modèle minimal de l'évolution darwinienne (variation et sélection), sans hypothèse *a priori* sur la force de la sélection sur les différents gènes (Figure II.4). Cette méthodologie permet d'étudier l'émergence de propriétés et de structures particulières en fonction des conditions contrôlées d'évolution et avec des mécanismes facilement identifiables. De plus, toutes les lignées, même celles éteintes, sont conservées, permettant une analyse *a posteriori* des dynamiques évolutives.

Les simulateurs d'évolution expérimentale *in silico* sont individu-centrés. Chaque organisme artificiel possède un matériel génétique qui lui est propre. Celui-ci est interprété par des programmes implémentant une "chimie artificielle" simple pour calculer le phénotype des individus. Au contraire des simulations en génétique des populations ou en phylogénie moléculaire, un phénotype est calculé à partir du génotype. En fonction de ce phénotype et de la réalisation d'une tâche donnée, les individus sont sélectionnés et un taux de reproduction leur est attribué. Durant la reproduction, le matériel génétique de chaque individu peut être soumis à divers types de mutations. Ainsi, dans l'évolution expérimentale *in silico*, des populations d'organismes digitaux évoluent et s'adaptent à leur environnement, c'est-à-dire à leur tâche à accomplir. Les mutations ont lieu sur le génotype et la sélection sur le phénotype, rendant l'évolution moins contrainte qu'avec les autres familles de simulateurs.

Dans les expériences d'évolution *in silico*, les effets du phénomène évolutif modélisé sont directement observés sur les organismes (séquences ou phénotype suivant le modèle), afin de comprendre comment le fait d'appartenir à une population évoluant dans des conditions spécifiques façonne les individus. De nombreuses études expérimentales ont utilisé cette approche, par exemple, pour étudier la relation entre la robustesse et l'évolvabilité[1]

[1]D'après Clark *et al.* (1999), *Buchnera* aurait environ 30 générations par an.

[1]L'évolvabilité correspond à la capacité d'une population d'organismes de générer de la diversité

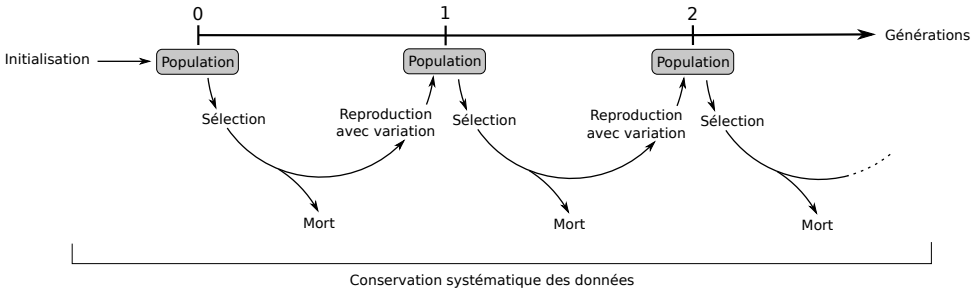


Figure II.4 – Principe de l'évolution expérimentale *in silico*
Les organismes sont modélisés par des structures de données, selon divers formalismes.
Durant une expérience, un ensemble d'organisme (population) est initialisé, aléatoirement, manuellement ou à partir d'une population issue d'une expérience précédente. L'évolution est modélisée par un cycle générationnel. A chaque génération, les organismes de la population sont évalués, généralement sur la base de leur capacité à effectuer des tâches prédéfinies, durant une phase de sélection qui détermine les individus se reproduisant et ceux qui meurent. Ensuite, les individus sélectionnés se reproduisent, avec des variations potentielles (mutations), pour créer une nouvelle population. Une nouvelle génération et un nouveau cycle démarrent alors.
Tous les organismes de toutes les générations peuvent être stockés dans des bases de données pour des analyses.
La figure est inspirée de la Figure 2 de Hindré *et al.* (2012)

(Wagner, 2008; Elena et Sanjuán, 2008), l'évolution de l'évolvabilité (Draghi et Wagner, 2008; Crombach et Hogeweg, 2008), l'évolution de la complexité (Soyer et Bonhoeffer, 2006) ou l'effet de la stochasticité sur l'adaptation microbienne (Jenkins et Stekel, 2010).

Dans un modèle d'évolution *in silico*, l'implémentation de la variation et de la sélection dépend de la façon dont sont codés le matériel génétique et la compétition pour les ressources. Ainsi, plusieurs types de formalismes sont disponibles d'après la nomenclature des modèles d'évolution *in silico* proposée par Hindré *et al.* (2012)[2].

Formalisme "Génome-programme" Le génome est une séquence d'instructions élémentaires dans un langage pseudo-assembleur interprété par un processus virtuel. *Avida*, le principal programme implémentant ce formalisme (Ofria et Wilke, 2004), a permis l'étude de la robustesse (Wilke *et al.*, 2001), de l'évolvabilité (Elena et Sanjuán, 2008), des effet des faibles taux de mutation (Nelson et Sanford, 2011), de la radiation adaptative (Chow *et al.*, 2004), de l'évolution de la complexité (Lenski *et al.*, 2003), de la modularité (Misevic *et al.*, 2006), etc.

Formalisme "Génome-graphe" Les individus sont caractérisés par un réseau, généralement de taille fixe, pouvant représenter un réseau protéique, un réseau de régula-

génétique adaptative.

[2]Un des formalismes de la nomenclature de Hindré *et al.* (2012) n'est pas présenté dans la suite car il s'approche des simulateurs en génétique des populations. En effet, dans le formalisme "Génome-collection d'allèles", le génome correspond à un nombre fixe de gènes. Chaque gène peut exister dans un nombre fini ou infini d'allèles, et chaque individu est ainsi caractérisé par ses allèles. Les études des lignées hypermutatrices (Taddei *et al.*, 1997; Tenaillon *et al.*, 1999) et de la spéciation bactérienne en l'absence de sélection (Hanage *et al.*, 2006) ont été effectuées avec ce formalisme.

tion de gènes, un réseau neuronal ou un circuit logique, et n'ont pas de séquences ADN. Ce formalisme a été utilisé pour l'étude de la modularité (Kashtan et Alon, 2005; Espinosa-Soto et Wagner, 2010), des relations entre la robustesse aux mutations et la robustesse au bruit (Kaneko, 2011), l'évolution de la communication et de l'altruisme (Floreano *et al.*, 2007; Waibel *et al.*, 2011), l'évolvabilité (Draghi et Wagner, 2009), la complexité des voies de signalisation (Soyer et Bonhoeffer, 2006), le comportement prédictif (Tagkopoulos *et al.*, 2008), l'effet des transferts horizontaux dans les réseaux génétiques (Mozhayskiy et Tagkopoulos, 2012b), le taux d'évolution selon l'environnement (Mozhayskiy et Tagkopoulos, 2012a), l'impact de la variation environnementale (Tsuda et Kawata, 2010), l'émergence de la robustesse (Krishnan *et al.*, 2008), etc.

Formalisme "Génome-collier de perles" Le génome est représenté par une chaîne de longueur variable d'éléments, dont la collection est prédéfinie. Ce formalisme a permis l'étude de l'évolvabilité des génomes (Crombach et Hogeweg, 2007), de l'évolvabilité des réseaux de gènes (Crombach et Hogeweg, 2008), de la spéciation en l'absence de barrière géographique (Tusscher et Hogeweg, 2009) et de la transformation des ressources dans un écosystème (Crombach et Hogeweg, 2009).

Formalisme "Génome-séquence de nucléotides" Le génome est représenté par une chaîne de caractères de longueur variable représentant les nucléotides. Ce formalisme a été utilisé pour l'étude de l'évolution du nombre de gènes et de l'ADN non codant (Knibbe *et al.*, 2007a), l'évolution de l'organisation des gènes en opérons (Parsons *et al.*, 2010), l'évolution de la taille et de la topologie des réseaux de régulation (Dwight Kuo *et al.*, 2006; Mattiussi et Floreano, 2007; Beslon *et al.*, 2010), l'évolution de la coopération (Frénoy *et al.*, 2012; Misevic *et al.*, 2012; Frénoy *et al.*, 2013), etc.

Un formalisme avec un niveau ADN permet plus de flexibilité pour incorporer des représentations réalistes des différents types de mutations. Le modèle *aevol* a été conçu, par Guillaume Beslon et Carole Knibbe, pour étudier l'évolution de l'organisation fonctionnelle des génomes et des réseaux de gènes, avec le formalisme "Génome-séquence de nucléotides". Les individus ont des génomes circulaires et une machinerie de transcription et de traduction explicitement basée sur des séquences signal et très inspirée de la génétique bactérienne. La structure des génomes est ainsi très proche de l'organisation des génomes bactériens avec du codant et du non codant, un nombre variable de gènes, une longueur variable des gènes, des opérons, etc. *aevol* semble ainsi être un bon outil de simulation des scénarios d'évolution réductive, mais surtout de l'étude de l'impact de ces scénarios sur la complexité génomique. Ce modèle permet de se détacher des hypothèses d'évolution moléculaire utilisées pour caractériser les changements et ne requiert pas de données de populations, actuellement peu disponibles. De plus, contrairement à d'autres modèles d'expériences d'évolution *in silico*, comme *Avida* (Adami *et al.*, 1994) ou le modèle de Crombach et Hogeweg (2007), *aevol* possède une notion de gène au sens biologique, avec une séquence, et de l'ADN intergénique permettant ainsi d'étudier l'érosion des gènes et du non codant observé chez *Prochlorococcus* (Chapitre IX et Section VII.1).

Chapitre III

Tester les hypothèses proposées pour l'évolution réductive avec *aevol*

III.1 *aevol* : modèle de l'évolution de la taille et de l'organisation des génomes bactériens

La plate-forme *aevol* a été conçue pour étudier l'évolution de la taille et de l'organisation des génomes bactériens dans des scénarios divers. Elle simule l'évolution d'une population de N organismes artificiels en utilisant un cycle variation-reproduction (Figure III.1). Par défaut, la population est de taille constante dans le temps et est totalement renouvelée à chaque pas de temps. Chaque organisme possède un chromosome circulaire, double-brin contenant une chaîne de nucléotides binaires. Ce chromosome contient des séquences codantes (gènes) séparées par des régions non codantes. Chaque séquence codante est détectée par un processus de transcription-traduction (voir ci-dessous). Elle est ensuite traduite en une "protéine" contribuant à un ensemble de traits phénotypiques abstraits. L'interaction de toutes les protéines donne les valeurs des différents traits phénotypiques. L'adaptation d'un individu est alors mesurée en comparant les valeurs de ses traits phénotypiques à des valeurs optimales (arbitrairement choisies) pour la survie dans l'environnement. A chaque pas de temps, N nouveaux individus sont créés en reproduisant préférentiellement les individus les plus adaptés de la génération parentale. Après cela, tous les individus de la population parentale meurent. Avec ce modèle de reproduction générationnel, un individu peut avoir plus de deux descendants et une génération dans *aevol* correspond ainsi à plusieurs générations d'une vraie bactérie. Quand un chromosome est répliqué, il peut subir des mutations ponctuelles, des petites insertions et des petites délétions, mais aussi de grands réarrangements chromosomiques comme des duplications, des grandes délétions, des inversions et des translocations[1]. Ainsi, les mutations peuvent modifier les gènes existants mais aussi créer de nouveaux gènes, éliminer des gènes existants,

[1]Ce que nous appelons "translocation" dans ce manuscrit est l'excision d'un segment chromosomique suivie de sa réinsertion à un autre endroit du chromosome

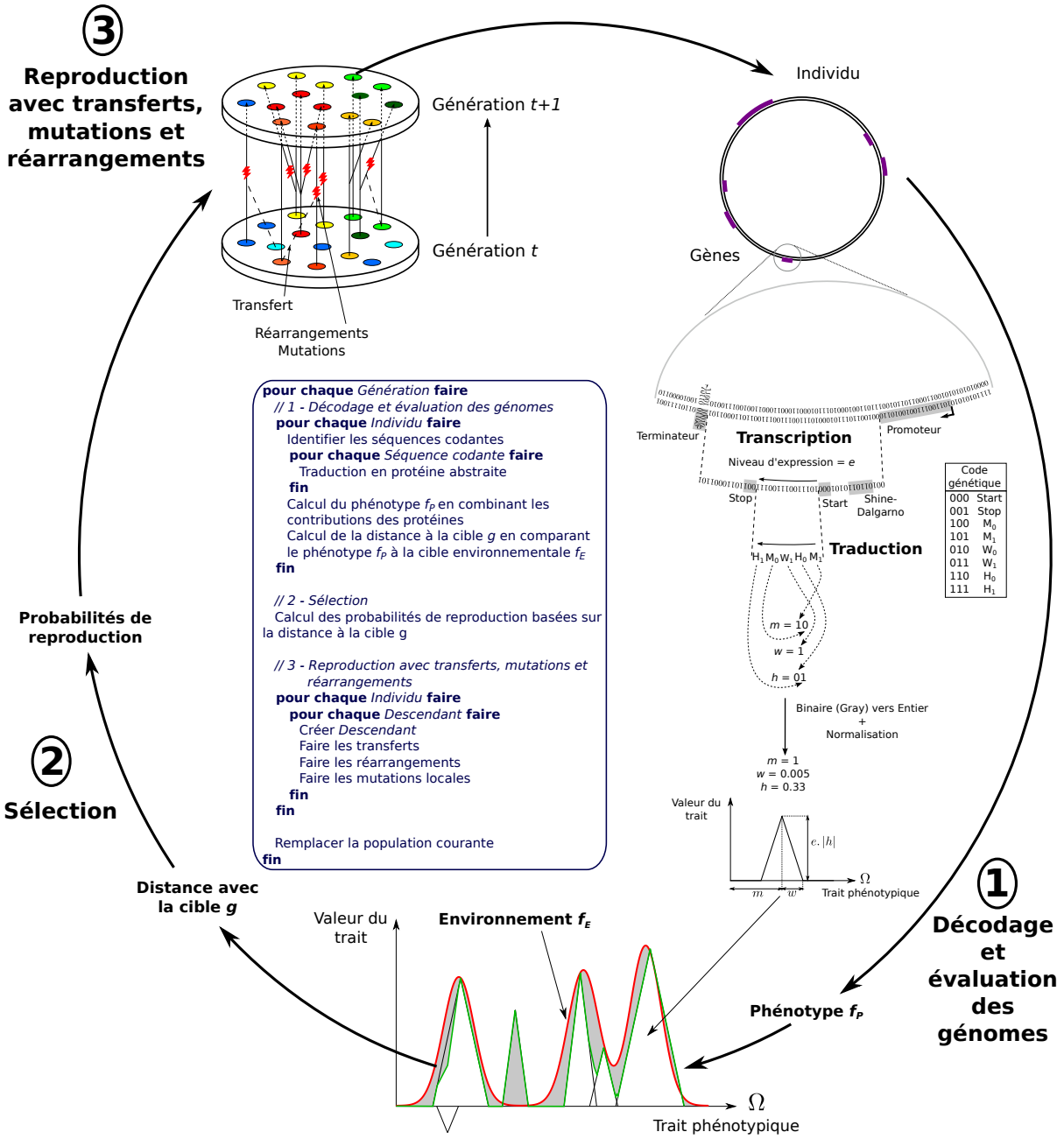


Figure III.1 – Représentation graphique de la plate-forme *aevol*
L'algorithme sous-jacent itère en trois étapes principales : (1) décodage et évaluation des génomes, (2) sélection des meilleurs individus et (3) reproduction avec mutations, réarrangements et transferts. Ces étapes sont détaillées dans la suite du texte principal.

modifier la longueur des régions intergéniques, modifier l'ordre des gènes, etc.

Avant de donner davantage de détails sur chaque niveau du modèle, il est utile de dire quelques mots de la plate-forme de simulation. L'implémentation du modèle prend la forme d'une suite logicielle codée en C++ (environ 48 000 lignes de code), exécutable en ligne de commande sur Unix et MacOS X. Ce code source est disponible sous licence GPL

sur le site `http://www.aevol.fr`. Outre le programme principal qui simule l'évolution d'une population pendant des milliers de générations, la suite comporte plusieurs programmes de pré- et post-traitements. L'installation repose sur les outils GNU et plusieurs configurations de compilation sont disponibles (avec ou sans régulation, avec ou sans sortie graphique, etc) ce qui permet d'adapter la compilation aux plate-formes, en particulier lors des expériences systématiques menées sur des clusters de calcul. Un package debian (avec les options de compilation par défaut) est également disponible dans les dépôts officiels Debian. À l'heure actuelle, le logiciel n'est pas encore parallélisé, ce travail est en cours dans l'équipe. Il est pour l'instant possible de simuler l'évolution de populations d'au maximum quelques milliers d'individus et des génomes atteignant plusieurs centaines de milliers de caractères. Une petite dizaine de contributeurs, principalement de l'équipe Inria-LIRIS Beagle mais aussi de l'équipe Inserm U1001, font évoluer le code source en fonction de leurs besoins et de ceux des utilisateurs qui exploitent le code sans le modifier.

III.1.1 Calcul du phénotype

Dans *aevol*, chaque individu possède un chromosome[1]. Celui-ci est une séquence binaire, double brin, circulaire. Le calcul du phénotype démarre à partir du génotype en cherchant, sur les deux brins de chromosomes, des séquences promotrices et terminatrices délimitant les régions transcrites. Les promoteurs sont des séquences dont la distance de Hamming d avec un consensus prédéfini est inférieure ou égale à d_{max}. Dans la suite du manuscrit, le consensus est 01010110011110010010110 (22 paires de bases) et jusqu'à $d_{max} = 4$ différences sont autorisées. Les terminateurs sont des séquences capables de former une structure tige-boucle, comme les terminateurs bactériens ρ indépendants, avec une tige de 4 bases et une boucle de 3 bases. Promoteurs et terminateurs délimitent les régions transcrites, qui peuvent être chevauchantes lorsque plusieurs promoteurs partagent un même terminateur. En fonction de la distance d entre le promoteur et le consensus, le niveau d'expression du transcrit change : $e = 1 - \frac{d}{1+d_{max}}$.

Lorsque toutes les régions transcrites ont été localisées, des signaux d'initiation et de terminaison du processus de traduction sont recherchés dans les transcrits. Ces signaux délimitent les séquences codantes. Le signal d'initiation est le motif $011011****000$, soit un signal de type Shine-Dalgarno suivi d'un codon start (000 dans notre cas). Le signal de terminaison est simplement le codon stop 001. A chaque fois qu'un signal d'initiation est trouvé, les positions suivantes sont lues trois par trois (codon par codon) jusqu'à la rencontre d'un codon stop sur le même cadre de lecture. Si aucun codon stop n'est trouvé dans la région transcrite après le signal d'initiation, aucune protéine n'est produite. Une région transcrite peut contenir plusieurs séquences codantes (chevauchantes ou non), permettant ainsi l'existence et l'évolution de structures opéroniques.

Pour déterminer la contribution phénotypique de chaque séquence codante, un formalisme mathématique est utilisé. L'ensemble abstrait $\Omega = [0, 1] \in \mathbb{R}$ représente l'ensemble

[1] Il est possible de configurer *aevol* pour que chaque individu possède également un ou plusieurs plasmides, mais cette possibilité n'a pas été utilisée dans le présent travail.

des traits phénotypiques possibles. Un "trait phénotypique" est donc simplement ici un nombre réel entre 0.0 et 1.0. À chaque trait est associé un niveau de réalisation, qui sera également compris entre 0.0 et 1.0. Chaque protéine peut contribuer à un sous-ensemble de traits phénotypiques, avec un degré variable selon les traits. Formellement, la contribution d'une protéine à chaque trait phénotypique est représentée par une fonction mathématique $f : \Omega \rightarrow [0, 1]$. Pour chaque trait x, $f(x)$ définit la contribution de la protéine à x. Dans *aevol*, la fonction choisie est une fonction affine par morceaux avec une forme triangulaire symétrique (Figure III.1). Trois paramètres sont donc nécessaires pour caractériser complètement une telle fonction : la position m ("moyenne") du triangle sur l'axe, qui correspond au trait phénotypique principal de la protéine, la hauteur H du triangle qui détermine le degré du trait principal et la demi-largeur w du triangle qui représente l'étendue fonctionnelle de la protéine, un moyen de quantifier sa pléitropie. Ainsi la protéine est impliquée dans les traits allant de $m - w$ à $m + w$, avec une contribution maximale pour le trait à la position m. Le sous-ensemble des traits que la protéine affecte est ainsi l'intervalle $]m - w, m + w[\subset \Omega$. m et w sont spécifiés par la séquence codante, tandis que H est un paramètre composite prenant en compte à la fois le niveau d'expression de la séquence et l'efficacité intrinsèque de la protéine : $H = e.|h|$, où e est le niveau d'expression du transcrit et h l'efficacité de la protéine codée, comme m ou w, dans la séquence génique. Le signe de h détermine si la protéine contribue positivement ou négativement aux traits $]m - w, m + w[$. Ainsi, l'importance de la contribution phénotypique d'une protéine donnée est réglée par sa séquence primaire (h), la qualité (e) de son(ses) promoteur(s) et la variation possible du nombre de copies du gène (effet de concentration).

En termes computationnels, la séquence codante est interprétée comme l'entrelacement des codes Gray[1] des trois paramètres m, w et h. En termes biologiques, la séquence codante est lue codon par codon et un code génétique artificiel (Figure III.1) est utilisé pour traduire les codons en trois nombres réels m, w et h. Dans ce code génétique, deux codons sont assignés à chaque paramètre. Par exemple, w est calculé à partir des codons $W_0 = 010$ et $W_1 = 011$. Tous les codons W trouvés pendant la lecture de la séquence codante vont déterminer le code Gray de w. Le premier bit du code Gray de w est un 0 (respectivement un 1) si le premier codon W de la séquence est un W_0 (respectivement W_1) et ainsi de suite. Ainsi, si la séquence codante contient n_W codons de type W, elle code un entier compris entre 0 et $2^{n_W} - 1$. Une normalisation permet alors de ramener la valeur du paramètre dans la gamme autorisée par la simulation. Le paramètre w, qui détermine la largeur du triangle, est normalisée entre 0 et w_{max}, où w_{max} est un paramètre défini au début de la simulation. La valeur brute est ainsi multipliée par $\frac{w_{max}}{2^{n_W}-1}$. Les valeurs des paramètres m et h sont obtenues de façon similaire, m étant normalisée entre 0 et 1 et h entre -1 et 1.

Les triangles des protéines peuvent se chevaucher partiellement ou complètement. Plusieurs protéines peuvent donc contribuer aux mêmes traits phénotypiques. Les valeurs finales des traits phénotypiques sont obtenues en sommant les contributions de toutes les protéines, tout en écrétant le résultat entre 0 et 1. Si f_i^+ est la fonction de contribution

[1]Le code Gray, ou code binaire réfléchi, est une variante du code binaire où deux valeurs décimales successives diffèrent seulement d'un bit. Il permet d'éviter les falaises de Hamming du code binaire traditionnel

de la i^e protéine positive (protéine avec $h > 0$) et f_j^- la fonction de contribution de la j^e protéine négative ($h < 0$), le phénotype de l'individu est représenté par la fonction $f_p : \Omega \rightarrow [0, 1]$ telle que $f_p(x) = max(min(\sum_i f_i^+(x), 1) - min(\sum_j f_j^-(x), 1), 0)$.

III.1.2 Sélection

L'environnement dans lequel les bactéries virtuelles évoluent est modélisé indirectement par une fonction f_E sur l'intervalle $[0, 1]$. f_E spécifie la valeur optimale de chaque trait phénotypique dans cet environnement et peut être nulle pour certains traits. Cette distribution est choisie au début de la simulation et peut fluctuer dans le temps si le protocole expérimental le nécessite (voir ci-dessous). L'adaptation d'un individu est alors mesurée par l'écart $g = \int_0^1 |f_E(x) - f_P(x)|dx$ entre son phénotype f_P et la cible f_E. Cet écart, appelé "écart avec la cible" ou erreur métabolique, pénalise aussi bien les traits "trop peu" réalisés que les traits "trop" réalisés.

La population est asexuée, de taille fixe N et complètement renouvelée à chaque pas de temps. Il faut donc affecter une probabilité de reproduction à chaque individu, en fonction de son "écart à la cible" g, et tirer le nombre de reproductions effectives par un tirage multinomial. Différentes méthodes sont implémentées dans la plate-forme pour calculer la probabilité de reproduction d'un individu à partir de la distribution des valeurs de g pour l'ensemble de la population. Avec la méthode *Fitness-proportionate* (utilisée dans tout le présent travail), la probabilité de reproduction est directement fonction de la valeur de g : $\frac{e^{-kg}}{\sum_{i=1}^{N} e^{-kg_i}}$, où k est un paramètre contrôlant la force de la sélection. Si k est faible, la plupart des mutations seront quasiment sans effet sur la probabilité de reproduction, même si elles modifient beaucoup le phénotype et donc l'écart à la cible. k module à quel point la probabilité de reproduction dépend de la réalisation de la tâche et donc la quantité de dérive génétique. En effet, un faible k diminue la rapidité de fixation dans la population de mutations avantageuses mais augmente parallèlement la probabilité de fixation de mutations légèrement délétères, qui ne sont pas éliminées par une sélection faible. Le choix de la valeur de ce paramètre est discuté dans la suite du manuscrit. Le nombre de descendants de chaque individu est ensuite tiré selon une loi multinomiale de paramètres $\left(N, \left(\frac{e^{-kg_1}}{\sum_{i=1}^{N} e^{-kg_i}}, \frac{e^{-kg_2}}{\sum_{i=1}^{N} e^{-kg_i}}, ..., \frac{e^{-kg_N}}{\sum_{i=1}^{N} e^{-kg_i}}\right)\right)$.

Dans les expériences de ce manuscrit, la cible f_E est construite comme la somme de trois fonctions gaussiennes (Figure III.2). La hauteur de chacune de ces fonctions varie à chaque pas de temps autour d'une hauteur moyenne $\bar{h_i}$, selon un processus autorégressif d'ordre 1 de paramètres σ et τ[1] : $h_i(t+1) = \bar{h_i} + \Delta h_i(t+1)$ avec $\Delta h_i(t+1) = \Delta h_i(t)\left(1 - \frac{t}{\tau}\right) + \frac{\sigma}{\tau}\sqrt{2\tau - 1}\varepsilon(t)$ (Figure III.2). Les tirages $\varepsilon(t) \sim N(0, 1)$ pour chaque gaussienne i sont indépendants les uns des autres et sont normalement distribués. σ contrôle l'amplitude

[1]Un processus autorégressif d'ordre 1 est l'équivalent en temps discret d'un processus d'Ornstein-Uhlenbeck en temps continu. Au cours du temps, ce processus tend à revenir vers sa moyenne à long terme (*mean-reverting process*). Il peut être vu comme une modification de la marche aléatoire dans laquelle on tend à revenir vers un endroit central, avec une plus grande attraction quand on se trouve loin de cet endroit.

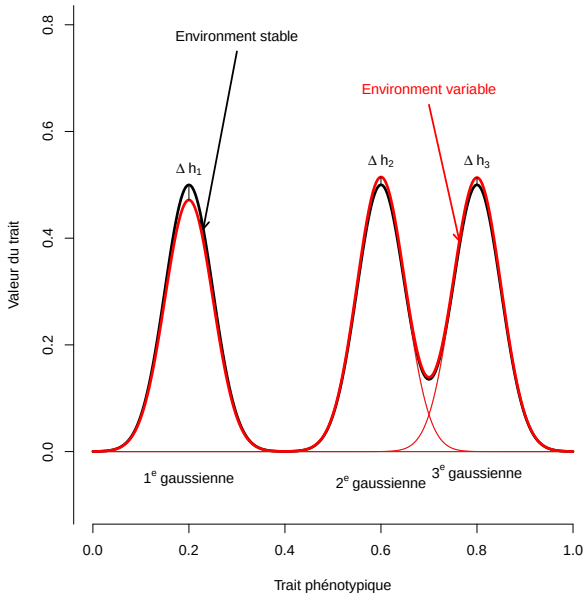


Figure III.2 – Fluctuation des hauteurs des trois fonctions gaussiennes formant la distribution cible f_E
La cible f_E, ou environnement, est la somme des trois gaussiennes. Lorsque l'environnement est stable (courbe en noir), les trois gaussiennes ont une hauteur identique, considérée comme la hauteur moyenne des gaussiennes. L'environnement ou cible résultant est l'environnement de référence. Lorsque l'environnement varie (courbe rouge), les hauteurs des gaussiennes varient indépendamment les unes des autres autour des hauteurs moyennes avec Δh_i suivant un processus auto-régressif d'ordre 1.

de la fluctuation et τ la vitesse à laquelle h_i tend à retourner à $\bar{h_i}$. Leurs valeurs pour les simulations sont discutées dans la section III.2.1.

III.1.3 Mutations

Quand un individu se reproduit, son génome est répliqué et peut subir des mutations. Il peut s'agir de mutations dites locales, concernant quelques nucléotides, de grands réarrangements chromosomiques ou d'échanges de portions de génome entre individus.

Pour une mutation ponctuelle, une position aléatoire est changée de 0 vers 1 ou inversement. Pour une petite insertion (respectivement une petite délétion), une courte séquence aléatoire (de taille uniforme entre 1 et 6 bases) est insérée (respectivement délétée) à une position aléatoire. Pour une grande délétion (respectivement une inversion), deux positions p_1 et p_2 sont tirées uniformément sur le chromosome et le segment $p_1, ..., p_2$ est délété (respectivement inversé). Pour une duplication (respectivement une translocation), trois positions p_1, p_2 et p_3 sont tirées uniformément sur le chromosome et le segment $p_1, ..., p_2$ est copié (respectivement déplacé) à la position p_3 dans son orientation originale.

Pour chaque type de mutations (mutations locales et réarrangements), un taux spontané

par nucléotide u_{type} est choisi au début de la simulation. u_{type} représente la probabilité d'une mutation de type *type* pour chaque nucléotide du génome. Le nombre moyen de mutations de type *type* subies par un organisme au cours d'un évènement de réplication est donc $u_{type}L$ avec L la longueur du chromosome de l'organisme. Techniquement, quand un individu se reproduit, après l'étape de transferts, détaillée par la suite, les nombres de réarrangements que le génome va subir sont calculés. Le nombre de grandes délétions est tiré selon la loi binomiale $B(L, u_{GrandesDeletions})$, le nombre de duplications selon la loi binomiale $B(L, u_{Duplications})$, ... Tous les réarrangements sont alors effectués dans un ordre aléatoire. Ensuite, les nombres de mutations locales (mutations ponctuelles, petites insertions et petites délétions) sont tirés et ces évènements sont effectués dans un ordre aléatoire. La longueur du génome peut donc varier pendant ce processus.

III.1.4 Transferts

Chez les bactéries, trois mécanismes sont à l'origine du transfert horizontal : la conjugaison, la transformation et la transduction. Ils entraînent soit des remplacements soit des insertions de séquences au sein du chromosome receveur. Chez *Prochlorococcus*, l'insertion de séquences, en particulier de gènes, a eu lieu principalement par de la transduction médiée par des bactériophages (Zeidner *et al.*, 2005; Sullivan *et al.*, 2005, 2003; Lindell *et al.*, 2004). *Prochlorococcus* semble aussi capable de transfert par remplacement (Section VII.3). Les deux types de transferts sont modélisés dans *aevol*, mais seul le transfert par remplacement sera utilisé ici car il est moins étudié chez *Prochlorococcus* et peut donner des clés de compréhension sur le processus d'évolution moléculaire.

L'étape de transfert dans *aevol* a lieu au début de la réplication. Un paramètre μ_t détermine la proportion d'individus pour lesquels des transferts vont être tentés. Quand un individu est sélectionné pour une tentative de transfert, il devient receveur et son donneur est sélectionné aléatoirement dans la population. Si le transfert réussit, un segment du génome du donneur sera copié et transféré chez le receveur, où il remplacera un segment similaire en séquence et en longueur.

Pour cela, une série de recherches locales est effectuée entre des points choisis aléatoirement pour chaque génome (donneur et receveur), selon une distribution uniforme le long des génomes. En d'autres termes, pour chaque tentative d'appariement, une position aléatoire est tirée dans le génome du donneur et une autre dans celle du receveur. Un alignement est recherché au voisinage des positions tirées. Le nombre maximal de paires de points candidats à tester est proportionnel à la taille du génome receveur : $nb\_paires = \mu_n \times L_{receveur}$, avec μ_n le taux de voisinage. Le voisinage dans lequel la recherche locale de similarité est effectuée correspond aux nucléotides à une certaine distance, définie par le paramètre $demi\_largeur$, des points candidats. Les séquences du donneur et du receveur dans cet espace sont face à face, mais peuvent aussi glisser l'une par rapport jusqu'à un seuil fixé par le paramètre $decalage\_maximal$. Ainsi, chaque nucléotide dans la zone de recherche d'une séquence est testé contre son vis-à-vis direct mais aussi ses $decalage\_maximal$ voisins en amont et en aval. Ceci produit une extension de chaque côté de l'espace de re-

cherche pour garantir que chaque nucléotide appartenant à l'espace de recherche est testé contre le même nombre de vis-à-vis. Un score est calculé par comparaison des paires de nucléotides avec un gain pour l'appariement des nucléotides ($gain\_appariement = 1$) et un coût au mésappariement ($cout\_mesappariement$ = -2). La probabilité que l'alignement soit accepté dépend de ce score de similarité : $P(accepter\ alignement) = \frac{1}{exp\left(\frac{score-\alpha}{\lambda}\right)+1}$. Les simulations présentées dans ce manuscrit ont été réalisées avec $\lambda = 4$, $\alpha = 50$, $\mu_n = 0.1$, $demi\_largeur = 50$ bases, $decalage\_maximal = 20$ bases.

Dans l'implémentation initiale du transfert avec remplacement dans *aevol* (Parsons, 2011), un transfert a lieu lorsque deux alignements distincts A_1 et A_2 sont trouvés entre les chromosomes du donneur et du receveur. Le segment entre les points de cassures de A_1 et A_2 du chromosome donneur remplace le segment entre les points de cassures de A_1 et A_2 du chromosome receveur. Les deux alignements sont recherchés indépendamment l'un de l'autre, n'importe où sur les génomes. Ainsi, la taille du segment reçu peut être bien différente de celle du segment remplacé. Ce type de transfert modélise donc des échanges de portions de génomes plus ou moins grandes, touchant potentiellement un grand nombre de gènes, avec une inégalité d'échange entre le donneur et le receveur.

Or, dans ce travail, nous souhaitons plutôt modéliser la recombinaison allélique telle qu'étudiée dans la section VII.3, c'est-à-dire le remplacement d'un allèle par un autre allèle du même gène. Le segment remplacé et le segment reçu doivent donc être approximativement de même longueur et être homologues sur toute leur longueur. Nous avons donc modifié le transfert avec remplacement dans *aevol*, de la façon suivante. Lorsqu'un premier alignement est trouvé entre le donneur et le receveur, nous tentons de l'étendre jusqu'à ce que la similarité de séquence soit perdue. Pour cela, nous testons si un alignement peut être accepté sur le même brin que le premier alignement, à une distance définie comme $3 \times demi\_largeur$ du centre du premier alignement. Les mécanismes de recherche des alignements sont identiques à ceux du premier alignement. Si un alignement n'est pas trouvé, le transfert n'aura pas lieu. Dans le cas contraire, tant qu'un alignement est trouvé et qu'un nombre tiré dans une loi uniforme est inférieur à une probabilité de détachement, un alignement est recherché à une distance de $3 \times demi\_largeur$ du dernier alignement trouvé. Lorsque les conditions de recherche ne sont plus vérifiées, la zone d'homologie entre les deux génomes est délimitée par le premier alignement et le dernier alignement trouvé et un transfert de cette zone peut avoir lieu entre le chromosome du donneur et celui du receveur. Le segment entre les deux alignements du chromosome receveur est remplacé par le segment entre les deux alignements du chromosome donneur.

III.1.5 Sorties et post-traitements des simulations

Les sorties principales d'une simulation avec *aevol* sont les séries temporelles, au cours de l'évolution, de valeurs comme la taille du génome ou le nombre de gènes, à chaque génération, pour le meilleur individu de la génération et pour la moyenne de la population.

Cependant, le meilleur individu à une génération donnée peut appartenir à une lignée

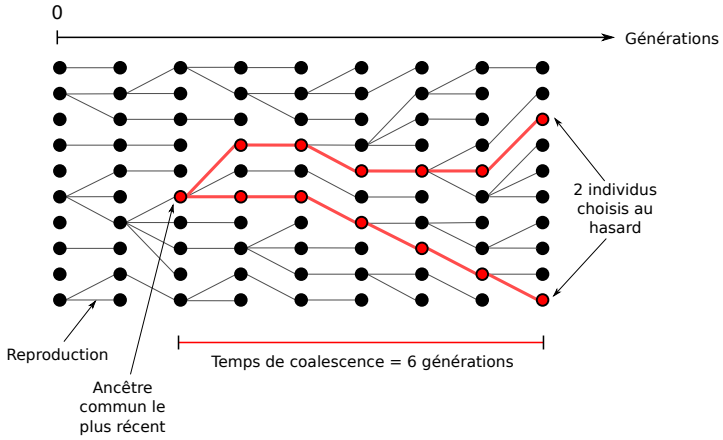
qui sera éteinte en fin de simulation. Or, nous nous intéressons plus particulièrement à l'évolution de caractéristiques fixées, c'est-à-dire celles conduisant aux individus en fin de simulation. À l'aide des arbres généalogiques, *aevol* permet de reconstruire la lignée ancestrale du meilleur individu obtenu en fin de simulation et nous nous intéressons à l'évolution des caractéristiques génomiques le long de cette lignée. De plus, pour comparer les caractéristiques en fin de simulation, les différents indicateurs sont étudiés sur l'ancêtre commun le plus récent de tous les individus de la dernière génération afin d'avoir accès à des caractéristiques partagées par tous les individus.

Les évènements mutationnels le long de la lignée ancestrale, c'est-à-dire gagnante, sont aussi étudiés afin de différencier les évènements spontanés des évènements fixés, conservés par la sélection. La lignée et ses évènements mutationnels sont aussi utilisés pour reconstruire des familles de gènes et leur évolution (création, perte, duplication d'un gène, ...). Contrairement aux méthodes de génomique comparative, les familles de gènes ne sont donc pas issues de la recherche d'homologie des séquences à partir des génomes "finaux" avec une méthode rétrospective. Elles sont construites avec une approche prospective en partant des gènes initiaux et en tenant compte de tous les évènements susceptibles de modifier les familles de gènes.

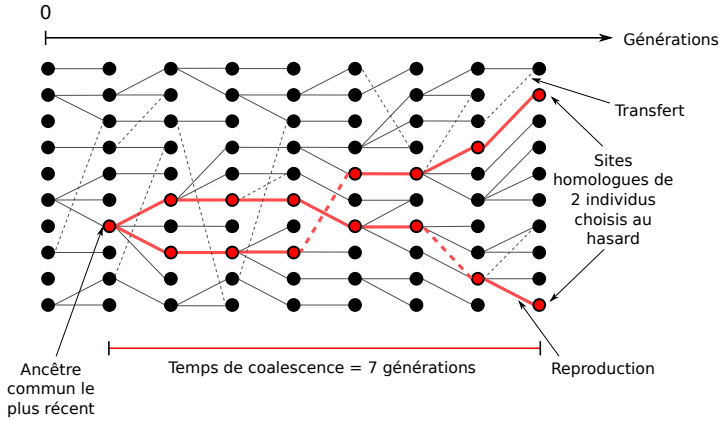
À intervalles réguliers, l'ensemble de la population est enregistré dans des fichiers de sauvegarde. À partir des informations contenues dans ces fichiers, des essais reproductifs peuvent être effectués pour l'ensemble de la population. Les distributions de certains caractères liés à la reproduction des individus au sein de la population sont ainsi accessibles, comme le nombre d'individus se reproduisant, le nombre de descendants par individus, les proportions de descendants identiques à leur parent en terme de fitness. Ces informations donnent des indications sur la génétique des populations dans *aevol*.

Autre indicateur important en génétique des populations, la taille efficace de population peut aussi être inférée. La taille efficace de population, notée N_e, correspond à la taille d'une population idéale évoluant seulement par dérive génétique qui présenterait le même niveau de diversité génétique qu'une population réelle. Cette mesure détermine ainsi le taux de changement de la composition d'une population causé par dérive génétique. N_e peut être estimé empiriquement à partir du temps de coalescence, c'est-à-dire le nombre de générations nécessaires pour remonter à l'ancêtre commun des sites génétiques de deux individus pris au hasard. Ainsi, la taille efficace de population à l'instant t est égale à $^{E[T]}/_{2}$ avec $E[T]$ l'espérance du temps de coalescence de deux individus. Cependant, dans de nombreux cas biologiques, le temps de coalescence est difficile à estimer parce que tous les évènements ne sont pas connus. Avec *aevol*, tous les évènements de reproduction sont enregistrés et le temps de coalescence de toutes les paires d'individus peut être calculé afin d'en déduire N_e.

En l'absence de transfert, le temps de coalescence peut être calculé de façon exacte (Figure III.3a). En revanche, en présence de transfert, le temps de coalescence de deux individus est plus compliqué à estimer. En effet, une partie du génome d'un individu peut provenir d'un autre individu que son parent principal. Le temps de coalescence de deux individus n'est alors plus calculable. Le temps de coalescence est alors calculé pour des ensembles de sites



(a) Sans transfert



(b) Avec transfert

Figure III.3 – Estimations des temps de coalescence dans aevol avec une connaissance exacte des évènements de reproduction et transfert
Tous les évènements de reproduction et de transfert sont enregistrés dans des fichiers de sauvegarde. En utilisant une méthode rétrospective, nous pouvons donc remonter à l'ancêtre commun le plus récent de deux individu choisis au hasard en suivant la lignée ancestrale de chacun des individus choisis.
Avec du transfert, le génome d'un individu n'est pas issu seulement de celui de n'importe quel de ces ancêtres directs, certaines portions ayant été acquises par transfert. Pour accéder au temps de coalescence, nous ne remontons pas jusqu'à l'ancêtre commun de deux individus mais jusqu'à l'ancêtre commun de sites homologues de deux individus en suivant la lignée ancestrale lorsque les sites ont été transmis par reproduction et en suivant les donneurs lorsque les sites sont transmis par transfert.

"homologues" entre des paires d'individus (Figure III.3b). Pour chaque paire d'individus, leurs génomes sont alignés afin de trouver des portions conservées entre les deux génomes. Pour chaque paire de portions dites "homologues", les évènements de reproduction et de transfert sont rejoués depuis la fin de simulation jusqu'à trouver l'ancêtre commun le plus récent de la paire. N_e est alors la moyenne pour toutes les paires d'individus de la moyenne du temps de coalescence de toutes les paires de portions "homologues", divisée par 2.

III.2 Méthodologie : Tester les hypothèses proposées pour l'évolution réductive

Les expériences d'évolution *in silico* sont habituellement réalisées de la façon suivante. Un paramètre dont l'impact doit être étudié est déterminé ainsi que les valeurs à tester. Les autres paramètres sont fixés et ne changent pas. Pour chaque valeur du paramètre à tester, plusieurs simulations sont effectuées avec des graines du générateur aléatoire différentes afin de prendre en compte l'effet statistique des observations. Toutes les simulations démarrent à la génération 0 avec des populations d'organismes "naïfs" générés aléatoirement ou manuellement. Dans le cas d'*aevol*, par exemple, la population est généralement initialisée avec des génomes aléatoires comprenant au moins un gène fonctionnel. Elles évoluent pendant un grand nombre de générations, sans changement des paramètres au cours de la simulation.

Pour tester les différentes hypothèses pouvant induire l'évolution réductive, un tel plan d'expérience n'est pas adapté. En effet, l'évolution réductive correspond à la réduction des génomes dans certaines lignées pour des organismes ayant déjà évolué et dont le génome est composé d'un grand nombre de gènes. C'est pourquoi nous avons proposé une nouvelle méthodologie pour utiliser la plate-forme (Batut *et al.*, 2013) (Figure III.5). Elle se base sur la construction de populations de génomes artificiels, nommées populations souches, par évolution pendant 150 000 générations. A partir de ces populations évoluées, différents changements de paramètres sont effectués, un par un, et l'évolution est prolongée pendant 50 000 générations avec ces nouveaux paramètres pour étudier leur impact sur la structure des génomes et tester s'ils induisent une évolution réductive. Ces populations avec changements sont appelés scénarios car ils symbolisent des hypothèses de scénarios pour expliquer l'évolution réductive, chez les endosymbiotes ou chez *Prochlorococcus*. Afin de pouvoir quantifier l'impact des changements, les populations souches continuent leur évolution en parallèle des différents scénarios. Elles sont nommées simulations de contrôle dans la suite du manuscrit.

Pour un minimum de puissance statistique, dix populations souches sont construites avec, à l'exception de la graine du générateur aléatoire, les même paramètres détaillés par la suite. Onze scénarios, soit onze changements de paramètres sont testés. Au total, en dehors des simulations de contrôles, 110 simulations, soit $5.5 \cdot 10^6$ générations, sont simulées. Cette campagne de simulation a nécessité un total d'environ 4 ans et 116 jours de calcul. Les campagnes préliminaires ayant permis de choisir les valeurs des paramètres (voir section

suivante) ont par ailleurs nécessité environ 30 ans de calcul en comptant la campagne présentée dans l'annexe A.

III.2.1 Choix des paramètres pour la construction des populations souches

Pour la construction des populations souches, dix simulations avec des paramètres identiques (Table III.1) sont jouées avec des populations de $N = 1000$ individus, démarrant d'une population clonale avec une séquence aléatoire de 5 000 bases et au moins un gène fonctionnel[1]. Elles sont simulées pendant 150 000 générations pour constituer les populations souches pour les scénarios, puis pendant 50 000 générations supplémentaires afin d'avoir des contrôles pour les scénarios. Certaines des valeurs des paramètres utilisés pour les populations souches sont différentes de celles couramment utilisées dans les simulations avec *aevol* et leur choix est donc discuté dans la suite.

Dans la plupart des campagnes de simulations effectuées avec *aevol*, le calcul des probabilités de reproduction se fait selon un schéma basé sur les rangs des individus (*Exponential ranking*). En effet, ces méthodes de sélection sont moins sensibles au phénomène de convergence prématurée vers un optimum local, observé avec des méthodes basées sur les valeurs brutes d'adaptation (*Fitness-proportionate*). Cependant ces dernières sont plus proches de la façon dont la sélection est modélisée dans les modèles de génétique des populations, et surtout de la biologie "réelle". En effet, un individu dix fois mieux adapté qu'un autre individu devrait se reproduire dix fois plus que l'autre individu. Pour cette campagne de simulations, le choix se porte donc sur la méthode de sélection dite *Fitness-proportionate*, où la probabilité de reproduction est directement fonction de la valeur d'écart à la cible g. Le paramètre k permet de contrôler la force de la sélection, en déterminant la vitesse à laquelle le coefficient de sélection s décroît quand l'écart à la cible g augmente (Figure III.4). Il est fixé à 750 dans les populations souches et sera augmenté ou diminué dans les scénarios.

Comme étudié dans Knibbe (2006), avec une méthode de sélection *Fitness-proportionate*, les génomes tendent à se raccourcir progressivement au cours du temps, principalement par la perte de bases non codantes. Ce phénomène peut s'expliquer par le rôle du non codant dans les réarrangements génomiques et donc dans la variabilité mutationnelle du phénotype. En effet, au fur et à mesure de l'évolution, les gains d'adaptation dus aux mutations favorables deviennent de plus en plus faibles, alors que les pertes d'adaptation dues aux mutations délétères peuvent rester conséquentes. En conséquence avec le mode de sélection basé sur les valeurs brutes d'adaptation, les mutations favorables ne sont plus sélectionnées alors que les mutations délétères sont, elles, contre-sélectionnées, passant ainsi d'une sélection directionnelle à une sélection stabilisatrice. Cela implique que le niveau de

[1]Pour chaque simulation, des séquences de 5 000 bases sont créées aléatoirement et testées jusqu'à ce qu'elles contiennent au moins un gène fonctionnel, c'est-à-dire codant pour un triangle de largeur et de hauteur strictement positives. La première séquence répondant à ce critère est alors donnée comme génome pour l'ensemble des individus de la population initiale d'une simulation.

Paramètres	Symbole	Valeur
Taille de population	N	1000
Taille du génome initial (aléatoire)	L_{init}	5 000 paires de bases
Séquence promotrice		0101011001110010010110 avec $d_{max} = 4$ mésappariements
Séquence terminatrice		$abcd***\overline{dcba}$
Signal d'initiation de la traduction		$011011****000$
Signal de terminaison de la traduction		001
Code génétique		Figure III.1
Ensemble global des processus cellulaires	Ω	$[0, 1]$
Pléiotropie maximale des protéines	w_{max}	$5 \cdot 10^{-3}$
Cible moyenne de l'environnement	f_E	Figure III.1
Variation de l'environnement : temps caractéristique	τ	5000
Variation de l'environnement : déviation standard	σ	0.05
Intensité de sélection	k	750
Taux de mutation ponctuelle	$u_{MutationPonctuelle}$	$5 \cdot 10^{-6}$ par pb
Taux de petite insertion	$u_{PetiteInsertion}$	$5 \cdot 10^{-6}$ par pb
Taux de petite délétion	$u_{PetiteDeletion}$	$1 \cdot 10^{-5}$ par pb
Taux de grande délétion	$u_{GrandeDeletion}$	$5 \cdot 10^{-5}$ par pb
Taux de duplication	$u_{Duplication}$	$5 \cdot 10^{-5}$ par pb
Taux d'inversion	$u_{Inversion}$	$5 \cdot 10^{-5}$ par pb
Taux de translocation	$u_{Translocation}$	$5 \cdot 10^{-5}$ par pb
Longueurs des petits indels		Loi uniforme entre 1 et 6 pb
Proportion d'essais de transferts	u_t	0.5 par individus
Taux de détachement		0.3

Table III.1 – Valeurs des paramètres utilisés pour la construction des populations souches
Ces valeurs ont été choisies après des analyses préliminaires. Certains paramètres comme les signaux structuraux n'ont pas d'impact sur la structure du génome. L'impact de w_{max} a été étudié (Knibbe *et al.*, 2007b) tout comme l'impact des taux de mutation et particulièrement les taux de réarrangement (Knibbe *et al.*, 2007a). Taux de mutation et w_{max} ont été choisis pour obtenir une densité de gènes assez proche de la densité de gènes bactérienne et avec suffisamment de gènes pour permettre des expériences sur l'évolution réductive. L'intensité et la fréquence des variations environnementales (σ et τ respectivement) ont été choisis suite à une large campagne d'expériences (Annexe A). k a été testé dans Batut *et al.* (2013).

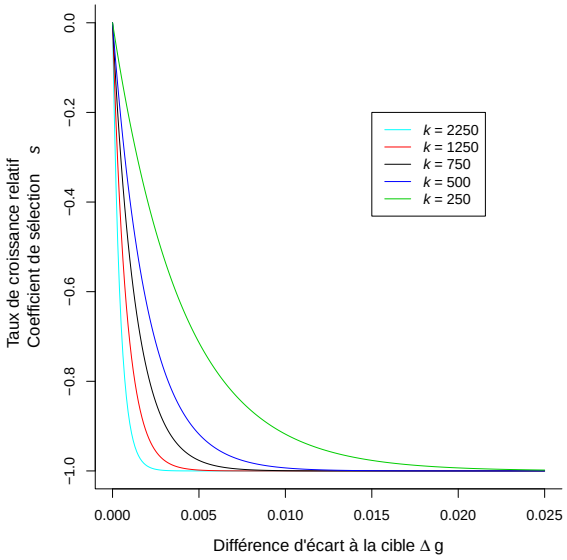


Figure III.4 – Taux de croissance relatif ou coefficient de sélection s en fonction de la différence d'écart à la cible entre deux individus
Le taux de croissance relatif est $\frac{e^{-kg_2}}{e^{-kg_1}} - 1$, soit le rapport entre la probabilité de reproduction pour des individus ayant une différence d'écart à la cible $\Delta g = g_1 - g_2$.

variabilité mutationnelle indirectement sélectionné diminue. Il devient alors avantageux de réduire les régions non codantes qui sont mutagènes pour les réarrangements[1] mais ne participent pas au phénotype, donc à la fitness des individus.

Afin de rester dans une sélection directionnelle et d'éviter l'érosion du non codant, les génomes doivent subir des changements fréquents des conditions d'évaluation de leur adaptation afin qu'ils soient confrontés à des tâches différentes à accomplir. Ainsi, dans ce travail, nous faisons fluctuer la cible environnementale à chaque pas de temps par changement des hauteurs des trois gaussiennes constituant la cible (Figure III.2) selon un processus régressif d'ordre 1 de paramètres σ et τ. σ contrôle l'amplitude de la fluctuation et τ la vitesse à laquelle une hauteur de gaussienne tend à retourner vers la hauteur moyenne de la gaussienne (Section III.1.2).

Une campagne de simulations a eu lieu durant cette thèse pour tester l'impact de σ et τ sur la structure des génomes (Annexe A)[2]. Ainsi, τ et la taille du génome ont une relation

[1] En effet, c'est la taille totale du génome et pas seulement la partie codante qui détermine le nombre de réarrangements spontanés subis à chaque reproduction ($n_{rear} = u_{rear} \times L$, avec n_{rear} le nombre de réarrangement, u_{rear} le taux de réarrangement spontané par base, L la taille du génome). Comme par ailleurs un réarrangement entre deux séquences non codantes affecte tous les gènes situés entre ces deux séquences (une délétion, par exemple), l'ADN non codant est *de facto* mutagène pour les gènes avoisinnants.

[2] On notera que cette campagne a été effectuée avec une cible environnementale un peu différente de celle utilisée dans ce présent travail et où ce sont les positions des gaussiennes qui fluctuent au cours du temps et non les hauteurs. En outre, les trois gaussiennes utilisées étaient chevauchantes, avec une gaussienne négative alors que les trois gaussiennes utilisées ici sont positives et peu chevauchantes. L'avantage de cet environnement est qu'il est plus simple à modifier pour les scénarios. De plus, la variation

en forme de cloche, avec des petits génomes pour les petites et grandes valeurs de τ et des grands génomes pour les valeurs moyennes de τ, principalement par des changements dans la quantité de bases non codantes. La forme de la cloche est exacerbée par des valeurs croissantes de σ. Bien que l'environnement et sa fluctuation soient différents entre la campagne précédente de simulations et les simulations de construction des contrôles, les impacts de σ et τ restent similaires, mais avec des plages quelques peu différentes. Les valeurs utilisées ici ($\sigma = 0.05$ et $\tau = 5000$) ont été choisies pour qu'au moins 80% des bases des génomes soient codantes, reflétant ainsi la forte densité en gènes des génomes bactériens, et que l'environnement varie assez lentement pour les génomes aient le temps de s'adapter à ces variations.

Afin d'atteindre un nombre assez important de gènes dans les souches pour espérer voir ensuite une réduction du nombre de gènes, nous avons choisi dans ce travail une pléiotropie maximale des protéines inférieure à celle utilisée par défaut dans *aevol*. Ce paramètre correspond à la largeur maximale des triangles w_{max}. En diminuant cette valeur, chaque triangle couvre une surface moins importante et plus de triangles sont donc nécessaires pour approcher au mieux la cible environnementale (Knibbe *et al.*, 2007b). Les valeurs de $w_{max} \in [0.01; 0.3]$ utilisées jusqu'à présent permettait d'obtenir environ 70 gènes, ce qui est trop faible pour des expériences d'évolution réductive. La valeur choisie $w_{max} = 0.005$ permet d'obtenir une centaine de gènes au minimum.

Les taux de mutation et de réarrangement ont aussi un impact sur le nombre de gènes et surtout sur la quantité de bases non codantes (Knibbe *et al.*, 2007a). Le génome est d'autant plus compact et pauvre en gènes que ces taux sont élevés. Comme mentionné précédemment, les génomes des populations souches doivent avoir assez de gènes pour simuler une évolution réductive mais aussi avoir une densité de gènes assez élevée. Des expériences préliminaires ont montré qu'il est possible d'obtenir au moins 80% des bases incluses dans des gènes, lorsque les taux de réarrangement (duplication, grande délétion, translocation, inversion) sont un ordre de grandeur supérieurs aux taux de mutation locale (mutation ponctuelle, petite insertion, petite délétion) et que ces derniers sont de l'ordre de $5 \cdot 10^{-6}$.

La compaction des génomes chez les bactéries semble principalement due à un biais mutationnel favorisant les délétions sur les insertions (Kuo et Ochman, 2009; Mira *et al.*, 2001). Afin de favoriser cette compaction, nous avons introduit un biais semblable à celui observé dans les bactéries dans les taux spontanés de mutation locale. Ainsi, dans ce travail, le taux spontané de petite délétion est systématiquement deux fois plus fort que le taux de petite insertion et le taux de mutation ponctuelle.

Enfin, alors que la plupart des bactéries libres sont capables d'effectuer des recombinaisons entre leur ADN et l'ADN d'autres bactéries (Takuno *et al.*, 2012), les endosymbiotes sont isolés génétiquement au sein d'une cellule eucaryote et donc peu exposés à de l'ADN exogène. Certaines espèces ont même perdu la faculté de recombiner. Cet arrêt de la re-

environnementale des hauteurs est plus facile à appréhender. En effet, la variation des moyennes dans l'environnement précédent entrainait une variation des hauteurs des pics de la cible environnementale car les gaussiennes sont chevauchantes. Cela rend l'impact des valeurs σ et τ plus difficile à interpréter.

combinaison est l'une des causes évoquées pour expliquer leur évolution réductive. Afin de pouvoir tester ce scénario, nous avons donné aux populations souches la capacité de recombiner entre elles de façon homologue, par transfert de portions homologues de génome entre un donneur et un receveur selon la procédure décrite dans la section III.1.3, avec un taux d'essais de transfert $\mu_t = 0.5$. Le coefficient de proportionnalité μ_n déterminant le nombre d'essais effectués pour trouver le premier alignement est fixé à $1 \cdot 10^{-5}$, la demi-largeur de l'espace de recherche à 50 bases et le décalage maximal entre les alignements à 20 bases. La probabilité de détachement lors de l'extension de la zone d'homologie par recherche d'alignement est de 0.3. Ces valeurs permettent d'obtenir environ $N/4$ transferts par génération, de taille généralement comprise entre 100 et 400 bases, ce qui correspond environ à la longueur d'un à quatre gènes dans nos génomes artificiels.

III.2.2 Scénarios : tests des hypothèses d'évolution réductive

L'objectif de cette partie est de tester des hypothèses émises dans la littérature ou issues de nos analyses pour tenter d'expliquer l'évolution réductive chez *Prochlorococcus*. Pour cela, les différentes hypothèses sont décomposées en sous-hypothèses. Ces dernières sont testées indépendamment les unes des autres grâce à *aevol* afin d'analyser leur impact sur les structures des génomes des populations souches. Si une évolution réductive est induite, les changements observés sont comparés aux génomes de *Prochlorococcus* et des endosymbiotes afin de déterminer si l'hypothèse testée peut expliquer une évolution réductive de type endosymbiose ou une évolution réductive similaire à celle observée pour *Prochlorococcus*.

A partir des populations souches, les dix simulations sont rejouées entre $t = 150000$ et $t = 200000$ dans onze scénarios, avec des changements de paramètres (Figure III.5). Les scénarios sont rassemblés en trois grandes catégories correspondant aux principales hypothèses émises pour expliquer l'évolution réductive.

III.2.2.1 Cliquet de Muller

L'évolution réductive chez les endosymbiotes est généralement expliquée par le cliquet de Muller (Wernegreen et Moran, 1999; van Ham *et al.*, 2003; Wernegreen, 2002; Moran, 1996), qui est un processus dégénératif touchant les populations non recombinantes, avec une petite taille efficace de population (Felsenstein, 2005; Muller, 1964). En effet, les endosymbiotes sont une cible idéale du cliquet de Muller : une petite taille de population touchée par des goulets d'étranglements fréquents (Mira et Moran, 2002) et un manque d'opportunités de recombinaison.

La dynamique des populations de *Prochlorococcus* est très différente de celle des endosymbiotes. Les estimations de N_e chez *Prochlorococcus* sont quatre ordres de grandeurs supérieures à celles pour *E. coli* (Baumdicker *et al.*, 2012; Charlesworth et Eyre-Walker,

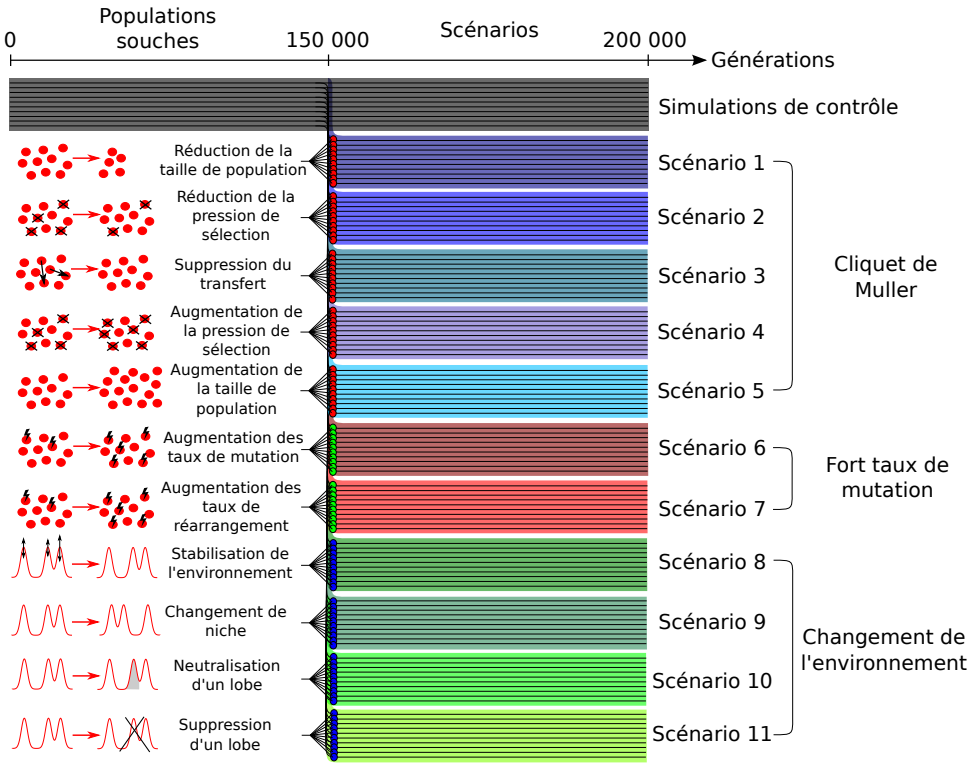


Figure III.5 – Méthodologie d'étude de l'évolution réductive par la construction de populations souches et simulation de différents scénarios concernant les paramètres évolutifs
Les scénarios de l'évolution réduction sont rassemblés en trois catégories : cliquet de Muller (en bleu), fort taux de mutation (en rouge) et changement de l'environnement (en vert). Les couleurs des scénarios correspondent à celles utilisées dans la suite du manuscrit pour désigner ces scénarios.

2006). En outre, contrairement à de nombreux endosymbiotes, *Prochlorococcus* a conservé l'essentiel de sa machinerie de recombinaison (Section VII.3). Il y a donc peu de raisons de supposer que *Prochlorococcus* pourrait être sujet au cliquet de Muller.

Afin d'éliminer cette hypothèse pour *Prochlorococcus* et la confirmer pour les endosymbiotes, elle est testée en simulation. Dans un premier scénario, la taille de population est divisée par deux, passant de $N = 1000$ à $N = 500$. Un autre scénario consiste à arrêter la recombinaison.

Sous l'effet du cliquet de Muller, la sélection naturelle est dépassée par la dérive génétique et est donc diminuée. Dans un troisième scénario, la force de la sélection k est donc diminuée de 750 à 250[1]. En diminuant la force de sélection k, le taux de croissance relatif des individus par rapport au meilleur individu est différent de -1 pour un plus grand nombre d'individus (Figure III.6) et le nombre d'individus dont la probabilité de reproduction n'est pas nulle augmente. Ainsi les résultats observés précédemment d'une évolution réductive similaire à celle de *Prochlorococcus* lorsque k est diminué (Batut *et al.*, 2013)

[1]Cette hypothèse, lors de tests préliminaires, a entraîné une réduction de la taille des génomes cohérente avec les observations faites chez *Prochlorococcus* mais pas chez les endosymbiotes (Batut *et al.*, 2013). Cependant, la recombinaison n'était pas présente dans ces simulations et pourrait changer les conclusions.

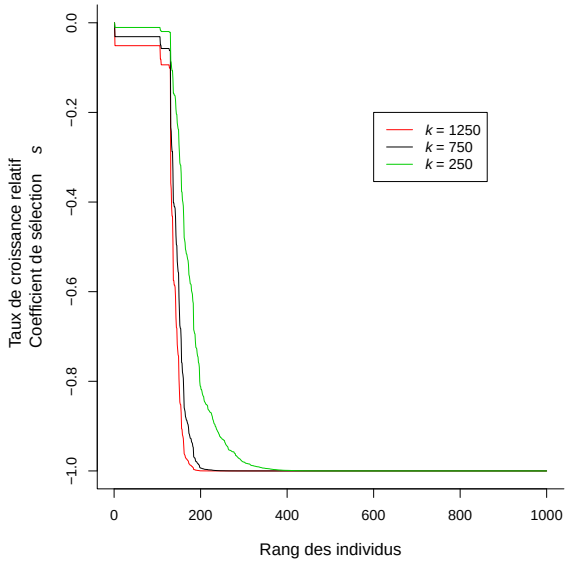


Figure III.6 – Taux de croissance relatif ou coefficient de sélection s en fonction du rang des invidus pour une population issue d'une simulation de population souche à $t = 150000$
Le taux de croissance relatif est $\frac{e^{-kg_2}}{e^{-kg_1}} - 1$, soit le rapport entre la probabilité de reproduction pour des individus ayant une différence d'écart à la cible $g_1 - g_2$.

pourraient s'expliquer par le fait que la taille efficace de population est augmentée avec la diminution de k, les *selective sweeps* étant moins forts. Pour tester cette hypothèse, il est alors intéressant de simuler l'augmentation de k afin d'analyser les conséquences d'une diminution de la taille efficace due à de forts *selective sweeps*. La valeur de k choisie pour l'augmentation est 1250 car les différences de taux de croissance entre les premiers individus sont du même ordre que celles pour $k = 250$ (Figure III.6).

Comme la diminution de k entraine une augmentation de N_e mais aussi une évolution réductive similaire à ce qui est observé chez *Prochlorococcus*, simuler l'augmentation de la taille de population permettrait de comparer les cas d'augmentation de N et d'augmentation de N_e. Ainsi, dans un scénario de contrôle, la taille de population est multipliée par deux, passant de $N = 1000$ à $N = 2000$.

III.2.2.2 Changements de l'environnement

La plupart des écotypes de souches réduites de *Prochlorococcus* se trouvent dans les eaux de surface des eaux marines tropicales et sub-tropicales, qui sont pauvres en nutriments toute l'année. La petite taille de génomes pourrait être une adaptation à la vie dans un environnement pauvre en nutriments (Rocap *et al.*, 2003; Dufresne *et al.*, 2005). Un plus petit génome signifie moins d'ADN dans la cellule et donc moins de besoins en nitrogène et en phosphore, deux élèments rares dans les eaux de surface. Cette hypothèse pour l'évolution réductive prédit une corrélation entre les réduction des génomes et les niches écologiques. Cependant, certaines souches réduites de *Prochlorococcus* vivent dans

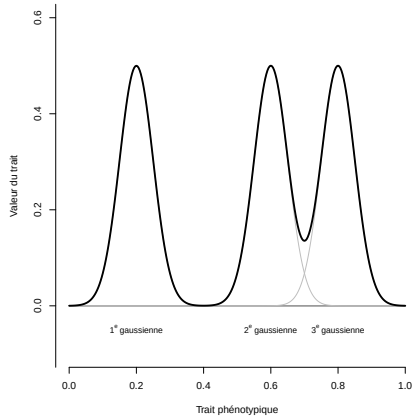
des niches similaires aux souches non réduites de *Prochlorococcus*. L'hypothèse adaptative pourait expliquer une amplification de la réduction des génomes mais les facteurs adaptatifs initiant le processus restent mystérieux. Des simulations de changements environnementaux pourraient permettre d'identifier des caractéristiques initiant une évolution réductive, telles que celles qui pourraient être à l'origine de l'évolution réductive pour l'ensemble des souches de *Prochlorococcus*. Pour cela, quatre scénarios sont proposés.

- L'environnement de *Prochlorococcus* en plus d'être pauvre en nutriments est stable tout au long de l'année, par contraste avec les eaux tempérées où vit *Synechococcus*. Un premier scénario consiste donc à stabiliser la cible environnementale, qui fluctue à chaque génération dans les populations souches.

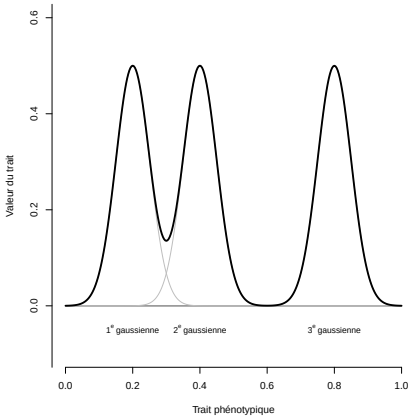
- La plupart des souches réduites de *Prochlorococcus* se trouvent dans un environnement différent des souches non réduites de *Prochlorococcus*. Le changement de niches pourrait avoir entrainé une réorganisation des gènes qui dans un certain contexte mutationnel déclencherait une évolution réductive. Pour ce scénario, une des gaussiennes constituant la cible environnementale est déplacée (Figure III.7b).

- L'évolution réductive chez les endosymbiotes s'est initiée avec le passage d'un mode de vie libre à un mode de vie intracellulaire. De nombreuses fonctions précédemment accomplies deviennent inutiles car les produits sont fournis par l'hôte. Les gènes correspondant sont éliminés. De façon similaire, si des tâches accomplies par certaines bactéries peuvent profiter à une communauté bactérienne entière, la plupart des bactéries perdent la capacité d'effectuer la tâche. C'est l'hypothèse de la Reine Noire (Morris *et al.*, 2012). Ainsi, les souches de *Prochlorococcus* ont perdu la capacité de produire le katG nécessaire pour leur protection et sembler bénéficier du katG sécrété par *Synechococcus* (Morris *et al.*, 2012). Il est cependant difficile de savoir combien de gènes peuvent être sujets à l'hypothèse de la Reine Noire et dans quelle proportion cela a contribué à la réduction des génomes chez *Prochlorococcus*. Pour tester cette hypothèse, la tâche à accomplir est simplifiée par la suppression d'un lobe. Deux alternatives sont alors possibles pour le sous-ensemble de traits phénotypiques désactivés. Il peut devenir délétère, c'est-à-dire que toute valeur non nulle pour ces traits phénotypiques de cet intervalle est contre sélectionnée (Figure III.7c), mais il peut aussi être neutralisé : les valeurs dans cette zone ne sont pas prises en compte dans le calcul de l'adaptation des individus (Figure III.7d). Ces deux cas sont testés, dans deux scénarios différents.

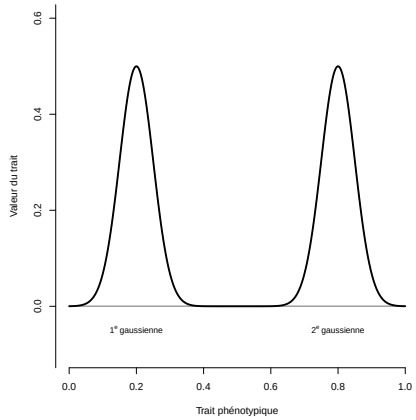
Enfin, chez *Prochlorococcus*, les réseaux de régulation semblent s'être simplifiés avec la perte de gènes de régulation (Rocap *et al.*, 2003; Dufresne *et al.*, 2003; García-Fernández *et al.*, 2004) du fait d'un environnement stable et simplifié. L'impact de ces changements d'environnement sur des individus avec un réseau de régulation des gènes ne peut pas être testé avec *aevol*. Cependant, dans une extension d'*aevol*, *raevol*, des réseaux de régulations des gènes ont été ajoutés. Cette extension et la méthodologie utilisée dans ce cas sont développés spécifiquement dans le chapitre V.



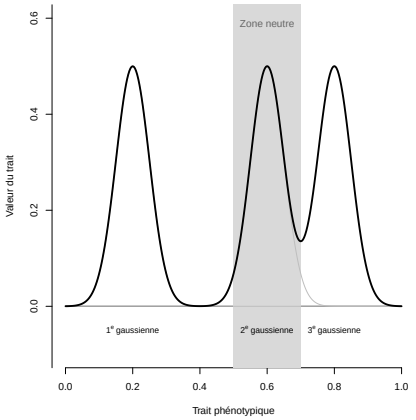
(a) Cible environnementale des populations souches



(b) Cible environnementale du scénario "Changement de niche" par déplacement d'une gaussienne



(c) Cible environnementale du scénario "Simplification de l'environnement" par suppression d'une gaussienne



(d) Cible environnementale du scénario "Simplification de l'environnement" par neutralisation d'une gaussienne

Figure III.7 – Cibles environnementales de populations souches et des scénarios liés aux changements de l'environnement
Chaque cible environnementale est la combinaison de deux ou trois gaussiennes, plus ou moins chevauchantes.

III.2.2.3 Fort taux de mutation

Les répertoires des gènes de réparation de l'ADN se sont modifiés au cours de l'évolution de *Prochlorococcus* (Partensky et Garczarek, 2010) et les souches réduites semblent avoir perdu des gènes de réparation de l'ADN. Cette perte et l'augmentation des taux de mutation concomitante peuvent expliquer les pertes importantes de gènes, l'enrichissement en AT et l'évolution rapide des séquences (Marais *et al.*, 2008). Pour vérifier ces hypothèses, deux scénarios indépendants sont simulés : la multiplication par cinq des taux de mutation locales, la multiplication par deux des taux de réarrangement[1].

III.3 Conclusion

Pour tester les hypothèses proposées pour l'évolution réductive, une nouvelle méthodologie a ainsi été mise en place et les valeurs des paramètres ont dû être ajustées afin d'être plus réalistes biologiquement. Par exemple, le transfert par homologie a été implémenté pour mieux correspondre à la recombinaison allélique telle qu'elle est envisagée dans les analyses génomiques de ce travail et surtout pour tester le scénario d'arrêt de la recombinaison, souvent énoncé comme cause de l'évolution réductive chez les endosymbiotes. Cependant, la présence des transferts dans les simulations complique la reconstruction des évènements *a posteriori*, en particulier le destin des familles de gènes. Ces situations ont pu être résolues mais l'estimation de la taille efficace de population en présence de transfert est encore en cours d'implémentation à ce jour.

Avec cette méthodologie, le nombre de répétitions doit être suffisant pour avoir une puissance statistique minimale mais ne doit pas être trop important. En effet, les simulations, du fait des paramètres choisis, sont coûteuses en temps de calcul mais aussi en espace disque. Chacune des 10 simulations de contrôle ont tourné pendant environ 44 jours pour 200 000 générations et utilisent environ 60 Go. À cela, s'ajoutent les 110 simulations des scénarios, dont les temps d'évolution dépendent des paramètres modifiés.

[1]Des simulations faites avec une augmentation par cinq des taux de réarrangement montrent que les taux de réarrangement sont trop importants. En effet, au moment de la reproduction, les individus ne peuvent éviter les réarrangements. Or les réarrangements ont des chances de modifier fortement la structure des génomes. Les individus sont donc moins bons. L'augmentation des taux de réarrangement est trop importante pour que les individus restent adaptés et une augmentation plus faible a donc finalement été choisie.

Chapitre IV

Résultats

Onze scénarios ont été testés, avec 10 répétitions chacun, pour déterminer si les changements effectués dans ces scénarios peuvent être à l'origine d'une évolution réductive semblable à celle observée chez *Prochlorococcus* (Tableau IV.1).

Les données disponibles pour *Prochlorococcus*, c'est-à-dire les génomes de 12 souches actuelles, sont différentes de celles accessibles dans les simulations avec *aevol* (Tableau IV.1). Ainsi, avec le peu de données pour *Prochlorococcus* et en particulier l'absence des états ancestraux, certaines caractéristiques sont inférées indirectement (force de la sélection et taille efficace de population avec le $^{d_N}/_{d_S}$, par exemple). Or, la présence de seulement 2 nucléotides et l'absence de redondance du code génétique dans *aevol* (Figure III.1) ne permettent pas d'avoir de telles mesures pour les simulations. Cependant, les caractéristiques de structure des populations et les pressions évolutives souhaitées par les mesures indirectes comme le $^{d_N}/_{d_S}$ sont plus facilement accessibles dans les simulations avec *aevol*. En effet, les évènements de reproduction, les évènements mutationnels et les séquences ancestrales sont enregistrés régulièrement et peuvent être utilisés pour reconstruire l'évolution de la structure de la population, les pressions évolutives en jeu, le destin des familles de gènes, etc.

Pour déterminer si des scénarios induisent une évolution similaire à celle observée pour *Prochlorococcus*, nous utilisons certains indicateurs simples de la structure génomique, comme la taille des génomes ou la proportion de bases codantes, en fin de simulation, c'est-à-dire après 50 000 générations d'évolution. Nous les comparons à des simulations dites de contrôle où aucun changement n'a été effectué. Un scénario induit une évolution réductive similaire à *Prochlorococcus* si la taille des génomes s'est réduite avec la perte de gènes et de bases non codantes de telle sorte que la proportion de bases codantes soit plus forte et la taille des gènes plus faible, conformément à ce qui est observé chez *Prochlorococcus* (Tableau IV.1). En effet, comme nous le détaillerons dans les chapitres VII et IX respectivement, la proportion de bases non codantes et la longueur des gènes conservés se réduisent le long de l'arbre phylogénétique de *Prochlorococcus*. Nous nous

Caractéristiques	*Prochlorococcus* réduite vs *Prochlorococcus* non réduite	Observable avec *aevol*
Taille du génome	Réduction	Observable directement
Proportion d'ADN codant	Augmentation (Section VII.1)	
Distances intergéniques	Réduction (Section VII.1)	
Couverture par les opérons	Augmentation (Section VII.2)	
%GC	Réduction	Seulement 0/1 comme nucléotides
Nombre de gènes	Réduction	Observable directement
Familles de gènes	Réduction	
Longueur des gènes	Réduction (Chapitre IX)	
Pseudogènes	Réduction	
Gènes de réparation	Pertes et gains	Pas de distinction des gènes selon leur fonction
Vitesse d'évolution	Augmentation	Observable par taux de fixation des évènements mutationnels
d_N/d_S	Stable	Pas de données d'usage des codons mais caractéristiques souhaitées accessibles plus directement
Changement dans la constitution en acides aminés	Enrichissement en bases AT (Section X.3.1)	
Architecture des génomes	Stable	Observable directement
Recombinaison intragénique	Stable (Section VII.3)	Accessible par les évènements de transfert fixés
Codons optimaux	Réduction (Section X.4)	Pas de données d'usage des codons
Gènes ARN_t	Réduction (Section X.5)	Pas de distinction des gènes selon leur fonction

Table IV.1 – Motifs des changements des caractéristiques de *Prochlorococcus* et comparaison avec les données issues de *aevol*
Le tableau est inspiré du tableau I.1 et de résultats qui seront détaillés dans la seconde partie du manuscrit.

contentons de ces caractéristiques génomiques générales pour déterminer des scénarios d'intérêt. Ces derniers sont ensuite analysés plus en profondeur via les données obtenues *a posteriori* des simulations, comme les mutations et réarrangements fixés ou le taux de reproduction neutre d'un individu[1], afin d'analyser les pressions évolutives à l'origine des changements observés.

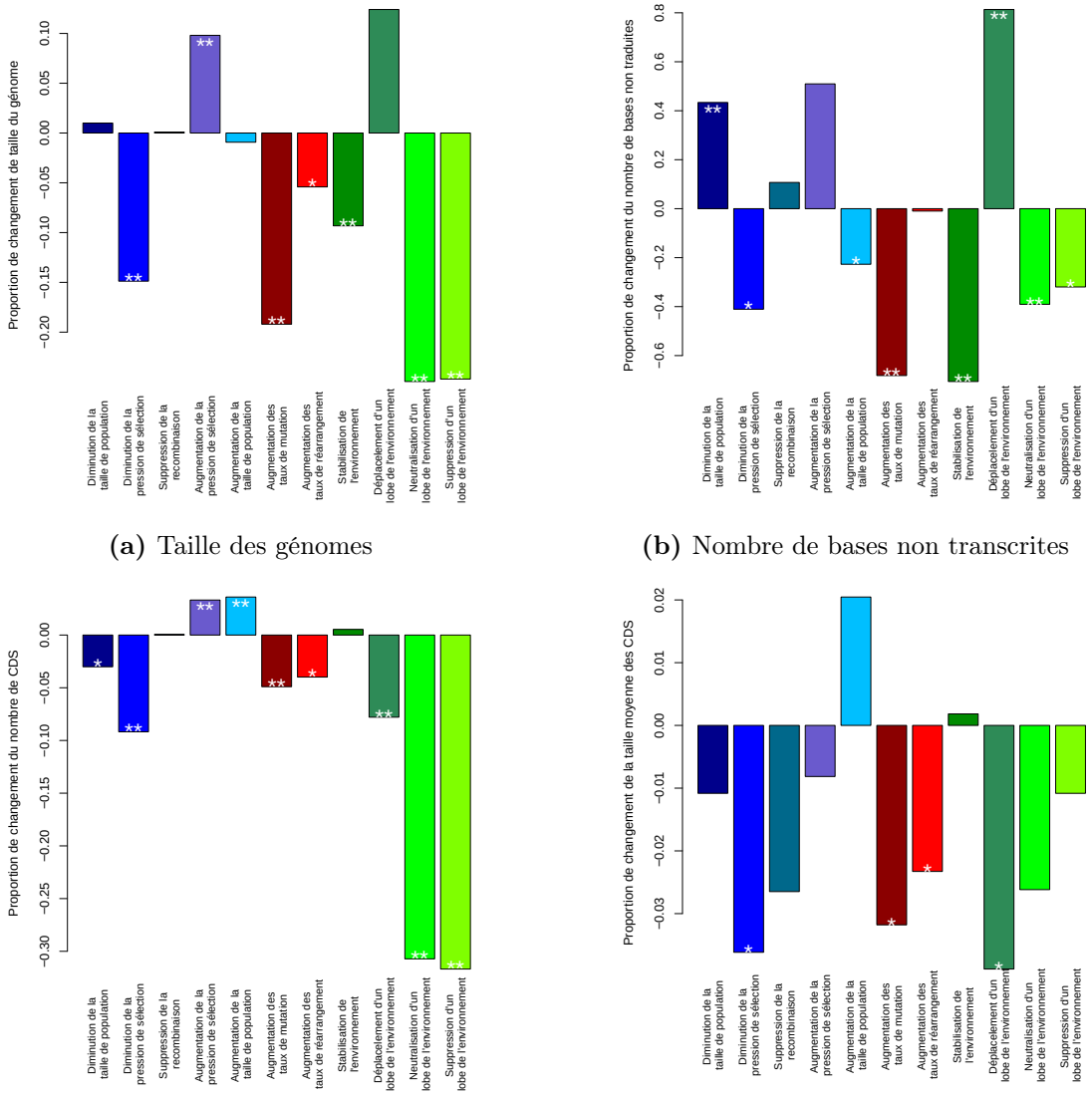
[1]Le taux de reproduction neutre d'un individu correspond à la proportion de descendants d'un individu qui sont neutres, c'est-à-dire qui n'ont pas subi de changements ou seulement des changements n'ayant pas d'impact sur la fitness.

IV.1 Comparaison des différents scénarios

Les différents scénarios testés n'induisent pas les mêmes changements génomiques et tous n'entrainent pas une évolution réductive (Figure IV.1). Ainsi, la taille des génomes ne diminue pas pour les scénarios de diminution de la taille de population, d'augmentation de la pression de sélection, d'arrêt de la recombinaison et de déplacement d'un lobe de l'environnement (Figure IV.1a). Pourtant, un biais spontané vers les petites délétions a été introduit dans toutes les simulations. Selon les raisonnements classiques trouvés dans la littérature, nous nous attendons donc à ce que tous les scénarios liés au cliquet de Muller se traduisent par une érosion du génome. Nos simulations montrent que ce n'est pas le cas : parmi ces scénarios, seul celui d'une diminution de la pression de sélection conduit à un génome significativement plus court. Une population deux fois moins grande ou un arrêt de la recombinaison ne suffisent pas à révéler ce biais mutationnel. Notons qu'une réduction plus drastique de la taille de population, avec une population divisée par dix, avait été testée dans une campagne préliminaire, et qu'elle n'avait pas non plus conduit à une réduction significative du génome, malgré le biais vers les petites délétions. Cela suggère que même avec une population de seulement 100 individus, la sélection peut rester suffisamment efficace dans nos simulations. Elle peut contrer aussi le biais vers les petites délétions, même dans les portions non codantes. Ainsi, dans les simulations de contrôle, la quantité de bases non transcrites ne s'érode pas (Figure IV.2b). L'ADN non codant n'a pas d'influence directe sur le phénotype mais il est une source de réarrangements et de mutations qui peuvent être utiles pour l'évolvabilité des individus (Knibbe *et al.*, 2007a), surtout lorsque l'environnement varie au cours des générations. Une certaine proportion d'ADN non codant est donc utile et cette quantité ne semble pas touchée par le biais vers les petites délétions.

L'évolution des caractéristiques, autres que la taille des génomes, dans les scénarios pour lesquels la taille des génomes se réduit, n'est pas toujours identique. Ainsi, le nombre de gènes augmente pour le scénario de stabilisation de l'environnement mais diminue pour le scénario de neutralisation d'un lobe de l'environnement (Figure IV.1c).

Les caractéristiques génomiques observées en fin de simulation ne décrivent pas totalement le comportement des scénarios. En effet, les dynamiques de changement peuvent être différentes alors que les résultats en fin de simulation sont similaires. C'est le cas pour les scénarios de neutralisation d'un lobe de l'environnement et de suppression d'un lobe de l'environnement. Pour ceux-ci, la taille des génomes, le nombre de gènes, le nombre de bases non transcrites sont inférieurs, en fin de simulation, aux valeurs dans les simulations de contrôle (Figure IV.1). Cependant, la dynamique d'évolution de ces caractéristiques est différente entre les deux scénarios (Figure IV.2). Le nombre de gènes chute brusquement pour le scénario de suppression d'un lobe alors que pour le scénario de neutralisation, la perte de gènes est plus lente (Figure IV.2c). Cette différence d'évolution vient de la définition des scénarios. En effet, bien que dans ces deux scénarios, un lobe de l'environnement ait disparu (Figures III.7d et III.7c), dans un cas la suppression entraîne une perte



(a) Taille des génomes

(b) Nombre de bases non transcrites

(c) Nombre de CDS

(d) Taille moyenne des CDS

Figure IV.1 – Proportion de changement de structure génomique entre les simulations de contrôle et les simulations des 11 scénarios pour l'ancêtre commun à l'ensemble des populations en fin de simulation

Les graphiques correspondent aux moyennes des proportions de changement entre scénario et contrôle pour les 10 simulations.

Les scénarios sont représentés par les différentes couleurs, avec en bleu les scénarios liés au cliquet de Muller, en rouge les scénarios liés aux forts taux de mutation et en vert les scénarios de changement d'environnement.

Les étoiles correspondent à la significativité des tests des rangs signés de Wilcoxon de comparaison entre scénarios et contrôles, avec une étoile pour une p-value comprise en 5% et 1% et deux étoiles pour une p-value inférieure à 1%.

de fitness à compenser[1] mais pas dans l'autre cas[2]. Ainsi, dans le premier cas, afin de rester proches de la cible, les individus doivent perdre rapidement les gènes qui codaient pour des triangles dans la partie de l'environnement supprimée. Dans le second cas, la perte des gènes codant pour des triangles dans la partie neutralisée n'est ni avantageuse, ni délétère à court terme. Il n'y a donc pas une pression qui pousse à leur perte rapide. Cependant, les séquences de ces gènes "inutiles" sont peu à peu éliminées par la dérive génétique. De plus, cet ADN, devenu "inutile", c'est-à-dire n'ayant pas d'influence directe sur le phénotype des individus, est mutagène pour les gènes qui le flanquent (Knibbe *et al.*, 2007a). En effet, la probabilité de perte d'un gène augmente s'il est entouré d'ADN inutile, car le nombre de réarrangements est proportionnel à la taille du génome. L'ADN "inutile" s'érode donc progressivement du fait d'une pression indirecte pour la robustesse des individus.

Les scénarios impactent donc les génomes de façon différente (Tableau IV.2), avec des dynamiques mutationnelles différentes (Tableau IV.3), parfois de façon opposée à ce qui aurait été attendu. Ainsi, la diminution de la pression de sélection (k) augmentant vraisemblablement la taille efficace de population, les résultats des scénarios de diminution de k et d'augmentation de la taille de population devraient être similaires, tout comme les scénarios d'augmentation de k et de diminution de la taille de population. Or les changements observés pour le scénario d'augmentation de la taille de population sont différents de ceux pour le scénario de diminution de k (Tableau IV.2), l'un entraînant une évolution réductive et pas l'autre. Cependant, les dynamiques mutationnelles sont relativement proches (Tableau IV.3). Mais, la diminution (respectivement l'augmentation) de la taille efficace de population n'a pas les même effets que la diminution (respectivement l'augmentation) de la taille brute de population sur les caractéristiques génomiques des organismes.

De façon surprenante, l'arrêt de la recombinaison ne semble pas avoir d'effet sur les caractéristiques génomiques des individus (Tableau IV.2). L'ajout de la recombinaison dans ces simulations n'apporterait donc rien. Or, nous constatons que la très grande majorité des familles de gènes dans les simulations de contrôle ont subi au moins un évènement de transfert. Ainsi, presqu'aucune famille de gène d'un individu n'est parvenue en fin de simulation par la lignée d'ascendance directe de l'individu. 80% des évènements mutationnels fixés sont des transferts, un transfert étant fixé toutes les cinq générations environ, et 82% des transferts sont neutres. Ainsi, même s'ils n'impactent pas les génomes directement, les transferts doivent permettre une évolution plus rapide des génomes en cas de changement, particulièrement en facilitant la fixation dans la population des gènes avantageux ou en limitant les effets d'auto-stop et donc la fixation de mutations légèrement délétères situées à proximité des mutations avantageuses.

In fine, sur les onze scénarios testés, seuls deux induisent une évolution réductive similaire à ce qui est observé chez *Prochlorococcus* avec la réduction du nombre de gènes, l'augmen-

[1] Tout gène codant pour un triangle dans cette partie de l'environnement devient délétère pour l'organisme.

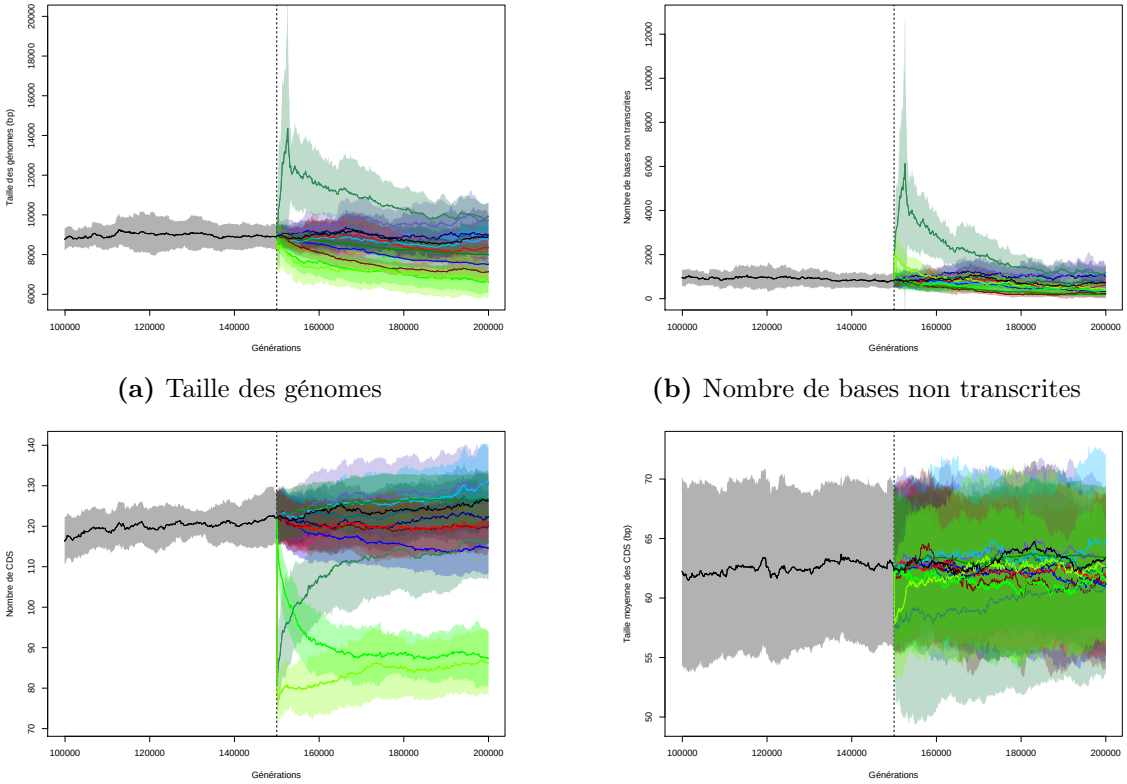
[2] Tout gène codant pour un triangle dans cette partie de l'environnement devient neutre pour l'organisme.

	↓ de N	↓ de k	∅ de la recombinaison	↑ de k	↑ de N	↑ des taux de mutation	↑ des taux de réarrangement	Stabilisation de l'environnement	Déplacement d'un lobe	Neutralisation d'un lobe	Suppression d'un lobe
Fitness	-	-	+	+	-	∼	-	+	-	+	+
Taille du génome	+	-	∼	+	-	-	-	-	+	-	-
Nombre de bases non codantes	+	-	+	+	-	-	∼	-	+	-	-
Proportion de bases codantes	-	+	-	-	+	+	+	+	-	∼	-
Nombre de CDS	-	-	∼	+	+	-	-	+	-	-	-
Taille moyenne des CDS	-	-	-	-	+	-	-	+	-	-	-
Nombre d'opérons	+	-	+	+	+	-	-	+	+	-	-
Nombre moyen de gènes par opérons	-	-	+	+	-	+	+	-	-	-	-
Proportion de gènes dans un opéron	+	-	-	+	-	+	+	+	-	-	-

Table IV.2 – Comparaison entre les différents scénarios et les contrôles pour des caractéristiques génomiques
Les comparaisons ont été faites pour les ancêtres communs à l'ensemble des individus des populations en fin de simulation.
Les "+" symbolisent des valeurs en moyenne supérieure pour les scénarios par rapport aux contrôles et les "-", l'inverse. Une case est colorée lorsque le test des rangs signés de Wilcoxon de comparaison entre les 10 simulations du scénario étudié et les 10 simulations de contrôle est significatif avec une p-value de 5%.

			↓ de N	↓ de k	∅ de la recombinaison	↑ de k	↑ de N	↑ des taux de mutation	↑ des taux de réarrangement	Stabilisation de l'environnement	Déplacement d'un lobe	Neutralisation d'un lobe	Suppression d'un lobe
Mut. loc.		Taux	+	-	+	+	-	+	-	-	+	+	+
		Proportion	+	-	+	+	-	+	-	-	+	-	-
	Prop.	Neutres	+	-	-	+	-	-	+	+	+	-	+
	Prop.	Délétères	-	+	-	-	+	+	-	+	-	+	~
	Prop.	Avantageux	-	-	+	+	+	-	-	-	-	-	-
	Imp.	Délétères	-	+	-	-	+	+	+	-	+	+	+
	Imp.	Avantageux	+	+	-	-	-	-	+	-	+	-	+
Réarr.		Taux	+	-	+	+	-	-	+	-	+	~	+
		Proportion	+	-	+	+	-	-	+	-	+	-	+
	Prop.	Neutres	+	-	-	+	-	+	+	+	+	-	-
	Prop.	Délétères	+	+	-	-	+	+	+	+	-	+	+
	Prop.	Avantageux	-	-	+	~	+	-	-	-	-	+	+
	Imp.	Délétères	+	+	+	+	+	-	+	+	+	-	+
	Imp.	Avantageux	+	-	~	+	+	-	+	+	+	-	+
Transferts		Taux	+	+	-	-	+	+	+	+	-	+	+
		Proportion	-	+	-	-	+	-	~	+	-	+	~
	Prop.	Neutres	+	-	\	+	-	-	+	~	+	-	-
	Prop.	Délétères	-	+	\	-	+	+	-	+	-	+	+
	Prop.	Avantageux	-	+	\	-	+	+	-	-	-	+	+
	Imp.	Délétères	-	-	\	+	-	-	+	-	-	-	-
	Imp.	Avantageux	+	+	\	-	-	-	+	+	-	-	-

Table IV.3 – Comparaison entre les différents scénarios et les contrôles pour les évènements mutationnels ayant eu lieu entre le changement de paramètre et les ancêtres communs à l'ensemble des individus des populations en fin de simulation
Pour les différents types d'évènements (mutations locales, réarrangements et transferts), les évènements peuvent être soit neutres, soit délétères, soit avantageux. Dans ce tableau, sont comparés le taux (nombre d'évènements rapporté à la taille des génomes) et la proportion des évènements par rapport à l'ensemble des évènements, les proportions d'évènements neutres, délétères et avantageux et l'impact moyen des évènements délétères et avantageux.
Les "+" symbolisent des valeurs en moyenne supérieure pour les scénarios par rapport aux contrôles et les "-", l'inverse. Une case est colorée lorsque le test des rangs signés de Wilcoxon de comparaison entre les 10 simulations du scénario étudié et les 10 simulations de contrôle est significatif avec une p-value de 5%.



(a) Taille des génomes

(b) Nombre de bases non transcrites

(c) Nombre de CDS

(d) Taille moyenne des CDS

Figure IV.2 – Évolution de certaines caractéristiques génomiques le long de la lignée ancestrale du meilleur individu de la génération 200 000 des scénarios
Les lignes représentent les moyennes pour les 10 simulations de chaque scénario et la plage colorée à l'écart-type sur les 10 simulations.
Les simulations de contrôle sont en noir et celles des scénarios en couleurs, avec en bleu les scénarios liés au cliquet de Muller, en rouge les scénarios liés aux forts taux de mutation et en vert les scénarios de changement d'environnement. Dans l'annexe B, les figures sont représentées scénario par scénario.

tation de la proportion de bases codantes et la réduction de la taille des gènes, mais aussi une accélération de l'évolution des séquences principalement par l'augmentation des taux de fixation des mutations locales (Tableau IV.3). Ce sont les scénarios de réduction de la pression de sélection k et d'augmentation des taux de mutation[1]. Ces scénarios sont des sous-hypothèses de deux des trois grandes hypothèses principales de l'évolution réductive que sont le cliquet de Muller, les forts taux de mutation et le changement d'environnement.

Ainsi, la réduction des pressions de sélection accompagnant le cliquet de Muller semble induire une évolution réductive similaire à celle observée chez *Prochlorococcus*, au contraire des autres scénarios testés (diminution de la taille de population, arrêt de la recombinaison,

[1]Notons en effet que les scénarios de simplification de l'environnement (suppression ou neutralisation d'un lobe) et celui d'augmentation des taux de réarrangement conduisent certes à une réduction de l'ADN codant et de l'ADN non codant (Tableau IV.2), mais dans les mêmes proportions, de sorte que le génome ne devient pas plus compact : la *proportion* d'ADN codant n'augmente pas significativement (Tableau IV.2), contrairement à ce qui est observé chez *Prochlorococcus* (détaillé dans la seconde partie du manuscrit au chapitre VII.1)

etc). Il faut cependant noter que, de façon peut-être contre-intuitive, diminuer la pression de sélection dans les simulation revient en fait à augmenter la taille efficace de population. En diminuant l'intensité des *selective sweeps*, nous nous rapprochons plus de la population "idéale" de Wright-Fisher, dans laquelle il n'y a pas de sélection du tout, et N_e devient plus proche de N. Or, le scénario de diminution de la pression de sélection semble reproduire une évolution similaire à celle observée chez *Prochlorococcus*, où les tailles efficaces de population seraient grandes, contrairement aux endosymbiotes. Cependant, le scénario d'augmentation de la taille de population, qui devrait aussi augmenter la taille efficace de population, ne donne pas les mêmes résultats. Le scénario de diminution de la pression de sélection mérite donc d'être approfondi afin de mieux comprendre les changements observés.

Dans les deux scénarios de l'hypothèse liée à des taux de mutation forts (augmentation des taux de mutation et augmentation des taux de réarrangement), les génomes se réduisent par la perte de gènes, la réduction de la longueur des gènes et du non codant. La proportion d'ADN codant augmente, mais pas de façon significative pour le scénario d'augmentation des taux de réarrangement (Tableau IV.3). De plus, les dynamiques mutationnelles à l'origine des changements (Tableau IV.3) sont différentes. Les moteurs derrières les changements pourraient donc être différents. Une étude plus poussée est ainsi nécessaire.

Dans l'hypothèse du changement d'environnement, 4 scénarios ont été testés : la stabilisation de l'environnement, le déplacement d'un lobe de l'environnement, la suppression d'un lobe de l'environnement et la neutralisation d'un lobe de l'environnement. Malgré une réduction des génomes observée pour 3 des scénarios (stabilisation de l'environnement, suppression et neutralisation d'un lobe de l'environnement), elle n'est pas entièrement similaire à celle observée chez *Prochlorococcus* : pas de pertes de gènes dans un cas, pas d'augmentation de la proportion de bases codantes dans les deux autres cas. Dans le scénario de déplacement d'un lobe de l'environnement, l'hypothèse la plus proche théoriquement de celle du changement de niche écologique émise dans la littérature, les génomes ne sont pas réduits malgré une réduction du nombre de gène et de la taille des gènes (Tableau IV.2). Il semble intéressant d'étudier pourquoi ces différents scénarios, et en particulier celui de déplacement d'un lobe de l'environnement, n'induisent pas une évolution réductive.

IV.2 Analyses détaillées de l'évolution réductive dans les scénarios

Afin de mieux comprendre les observations précédentes mais aussi l'évolution réductive chez *Prochlorococcus*, les scénarios induisant une évolution réductive similaire à celle observé pour *Prochlorococcus* sont analysés plus en détails. Nous souhaitons aussi comprendre pourquoi le scénario le plus proche des hypothèses proposées pour *Prochlorococcus* (déplacement d'un lobe de l'environnement) n'induit pas d'évolution réductive.

IV.2.1 Augmentation des taux de mutation et des taux de réarrangement

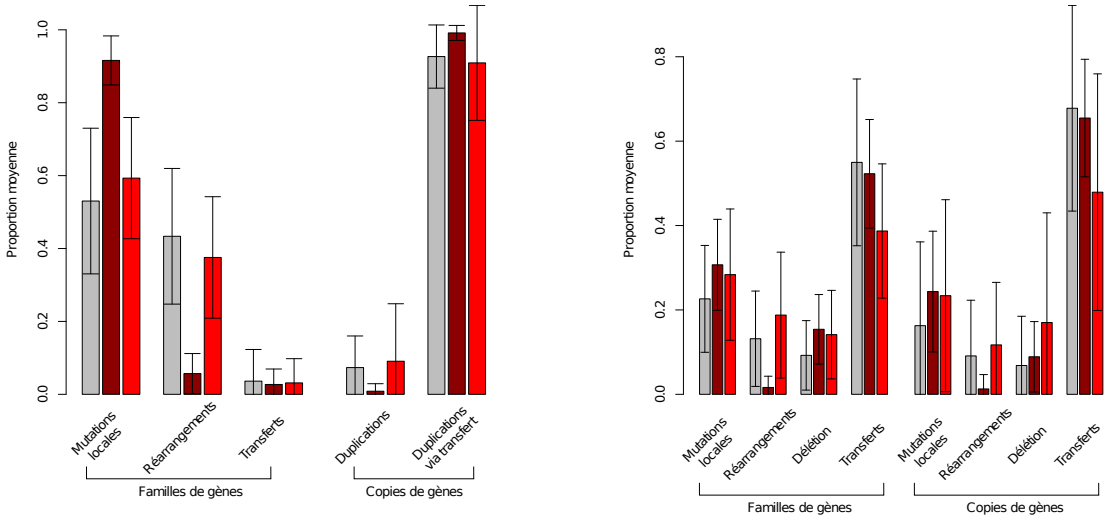
Dans *aevol*, l'augmentation des taux de mutation et celle des taux de réarrangement entraînent une évolution des génomes qui pourrait s'approcher de l'évolution réductive observée chez *Prochlorococcus*, avec des pertes de gènes, des gènes conservés plus petits (Tableau IV.2) et des taux d'évolution des séquences plus élevés (Tableau IV.3). Les modifications des structures génomiques ont lieu tout au long des 50 000 générations d'évolution après les changements de paramètres (Figure IV.2). Un état stable pourrait ne pas être encore atteint.

La réduction du nombre de gènes peut s'expliquer par l'effet direct de l'augmentation des taux de mutation et de réarrangement. La plupart des mutations étant délétères, le taux de mutation/réarrangement par base, comme dans nos simulations, peut imposer une limite supérieure au nombre de bases codantes qui vont pouvoir être conservées à l'identique (Eigen, 1971; Maynard Smith, 1983; Hurst, 1995; Pál et Hurst, 2004). Dans *aevol*, plus le taux de mutation/réarrangement est élevé, plus le nombre de gènes à l'équilibre est faible (Knibbe *et al.*, 2007a). De ce fait, l'augmentation, en cours de simulation, des taux de mutation ou de réarrangement entraîne une réorganisation du génome pour atteindre le nombre de gènes maximal étant donnés les nouveaux taux de mutation/réarrangement.

Les changements ne se limitent pas seulement à la perte de gènes. Les gènes présents en fin de simulations sont différents de ceux présents au moment du changement. Ainsi, pour l'augmentation des taux de mutation, seules environ 48% des familles de gènes en fin de simulation sont issues de familles présentes au moment du changement, pourcentage significativement inférieur au 72% de familles conservées des simulations de contrôle ($P = 0.031$, test des rangs signés de Wilcoxon sur les 10 répétitions). Dans le cas de l'augmentation des taux de réarrangement, les différences sont moindres mais restent significatives : environ 64% des familles en fin de simulation sont issues de familles présentes au moment du changement, pourcentage significativement inférieur aux valeurs des simulations de contrôle ($P = 0.031$, test des rangs signés de Wilcoxon).

La dynamique du répertoire génique est ainsi très forte, particulièrement dans le scénario d'augmentation des taux de mutation. Ainsi, dans ce scénario, 2.63 ± 0.29 fois plus de familles sont créées entre le changement des taux de mutation et la fin des simulations dans les simulations du scénario que dans celles de contrôle, et 2.33 ± 0.16 fois plus de familles perdues. Au sein des familles de gènes, 2.23 ± 0.28 fois plus de copies de gènes sont créés et 1.75 ± 0.38 fois plus de copies perdues dans le scénario d'augmentation des taux de mutation. Les familles et les copies de gènes sont créées et perdues principalement par des mutations locales et moins par des réarrangements que dans les simulations de contrôle (Figure IV.3), reflétant ainsi l'augmentation des taux de mutation.

Pour le scénario d'augmentation des taux de réarrangement, les renouvellements sont moindres et équivalents à ceux des simulations de contrôle. De plus, les pertes et les créations des familles et des copies de gènes au sein des familles ne sont pas significativement



(a) Évènements à l'origine de gains de familles de gènes et de copies de gènes

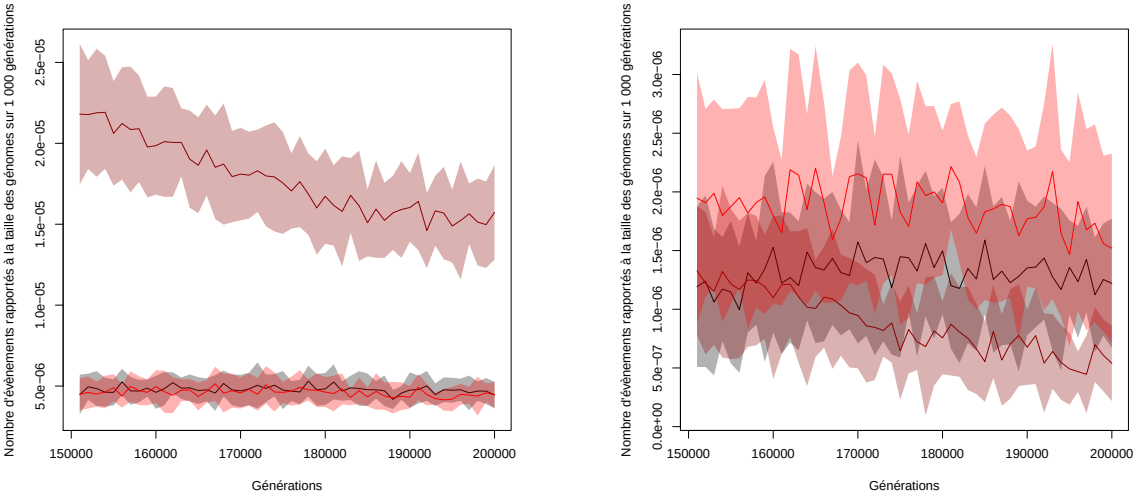
(b) Évènements à l'origine de gains de familles de gènes et de copies de gènes

Figure IV.3 – Évènements à l'origine des gains et pertes de familles de gènes et des copies de gènes au sein des familles pour les simulations de contrôle, celles du scénario d'augmentation des taux de mutation et celles du scénario d'augmentation des taux de réarrangement
Les valeurs représentées sont les moyennes sur les 10 simulations de chaque type de simulation avec l'écart-type sur les 10 simulations symbolisé par la barre d'erreur.
Les données pour les simulations de contrôle sont en gris, celles du scénario d'augmentation des taux de mutation en rouge foncé et celles du scénario d'augmentation des taux de réarrangement en rouge.

plus imputables à des réarrangements dans les simulations du scénario d'augmentation des taux de réarrangement que dans celles de contrôle (Figure IV.3). Les mutations locales restent la source majeure de création de familles de gènes.

Ces différences entre les scénarios s'expliquent par l'impact différent d'une mutation locale et d'un réarrangement. Un réarrangement étant plus délétère en moyenne qu'une mutation locale, les individus ayant subi des réarrangements seront davantage contre-sélectionnés que les individus ayant subi des mutations locales, même si les taux spontanés de réarrangement sont supérieurs aux taux spontanés de mutation locale. Cette différence de reproduction se reflète dans les mutations et réarrangements conservés et fixés dans les populations (Figure IV.4).

Les changements de répertoire et les pertes de gènes peuvent entraîner des pertes de fitness. En effet, avec moins de gènes, la tâche devient plus compliquée à accomplir. Ainsi, dans les scénarios d'augmentation des taux de réarrangement, la fitness des individus est diminuée (Tableau IV.2). De façon surprenante, ce n'est pas le cas pour le scénario d'augmentation des taux de mutation (Tableau IV.2). Malgré le plus faible nombre de gènes, les individus remplissent la cible environnementale de façon équivalente aux simulations de contrôle. Pour cela, les gènes présents en fin de simulation codent pour des triangles dont l'aire est supérieure ($P = 0.0244$, test des rangs signés de Wilcoxon sur les aires moyennes des triangles des 10 simulations). Ils peuvent couvrir ainsi plus facilement la



(a) Mutations locales **(b)** Réarrangements

Figure IV.4 – Évolution du taux de mutation et réarrangement fixé au cours des simulations de contrôle, celles du scénario d'augmentation des taux de mutation et celles du scénario d'augmentation des taux de réarrangement
Est représenté le nombre d'évènements rapporté à la taille des génomes par fenêtre de 1000 générations le long de la lignée ancestrale du meilleur individu final. Les lignes représentent les moyennes pour les 10 simulations de chaque scénario et la plage colorée représente l'écart-type sur les 10 simulations. Les données pour les simulations de contrôle sont en gris, celles du scénario d'augmentation des taux de mutation en rouge foncé et celles du scénario d'augmentation des taux de réarrangement en rouge.

cible environnementale. La baisse du nombre de gènes est donc compensée par une réorganisation du "métabolisme" des organismes. Pour le scénario d'augmentation des taux de réarrangement, l'aire moyenne des triangles n'est pas significativement différente de celle observée dans les simulations de contrôle ($P = 0.1611$, test des rangs signés de Wilcoxon sur les aires moyennes des triangles des 10 simulations). Les gènes perdus touchent ainsi tous les gènes, ceux de petite comme ceux de grande aire.

Malgré les pertes de gènes observées, la réduction des génomes dans le scénario d'augmentation des taux de mutation se fait principalement par la perte de bases non codantes (Figure IV.1b). Ainsi, en fin de simulation, les proportions de bases codantes sont supérieures aux proportions observées dans les simulations de contrôle, alors qu'elles sont conservées pour le scénario d'augmentation des taux de réarrangement (Figure IV.5).

Dans le simulateur, l'ADN non codant n'a pas d'influence directe sur la fitness d'un individu mais il est mutagène pour les gènes qu'il avoisine (Knibbe *et al.*, 2007a). La probabilité qu'une mutation locale n'affecte aucune région codante est égale à $(1 - \%L_1)$ et pour un réarrangement $\frac{1}{4}(2 - \%L_1)\left((1 - \%L_1)^2 + \frac{1}{2L^2}\sum_{i=1}^{N_G}\lambda_i(\lambda_i + 1)\right)$, avec $\%L_1$ la proportion de bases codantes, L la taille du génome, N_G le nombre de régions codante et λ_i la taille de la i^e région codante[1]. Or, plus les taux de mutation et de réarrangement sont

[1]Les probabilités $\widetilde{\nu}_{type}$ qu'une mutation aléatoire de type *type* n'affecte aucune région fonctionnelle

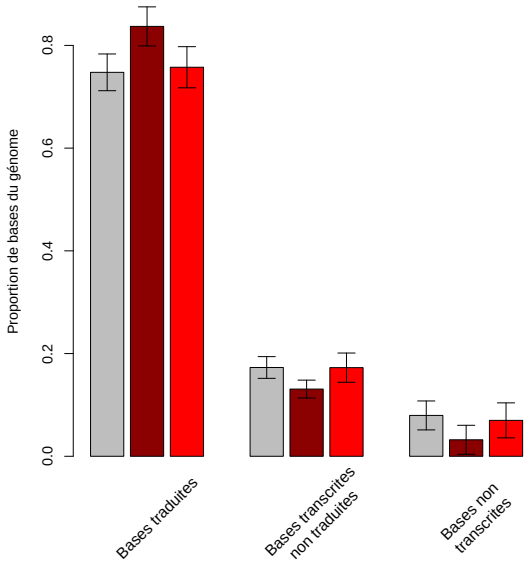


Figure IV.5 – Répartition des bases des génomes chez les ancêtres communs les plus récents des populations finales pour les simulations de contrôle, celles du scénario d'augmentation des taux de mutation et celles du scénario d'augmentation des taux de réarrangement
Les valeurs représentées sont les moyennes sur les 10 simulations de chaque type de simulation avec l'écart-type sur les 10 simulations symbolisé par la barre d'erreur.
Les données pour les simulations de contrôle sont en gris, pour celles du scénario d'augmentation des taux de mutation en rouge foncé et pour celles du scénario d'augmentation des taux de réarrangement en rouge.

élevés, à taille de génome constante, plus le nombre de mutations et de réarrangements est important, les taux étant définis par base. Les génomes subissant plus d'environ 1 évènement par génération sont contre-sélectionnés car trop instables pour transmettre leur répertoire génique intact (Fischer, 2013) et la proportion de non codant dépend des taux spontanés de mutation/réarrangement par le biais d'une sélection indirecte pour un niveau approprié de variabilité mutationnelle (Knibbe *et al.*, 2007a). Les lignées conservées

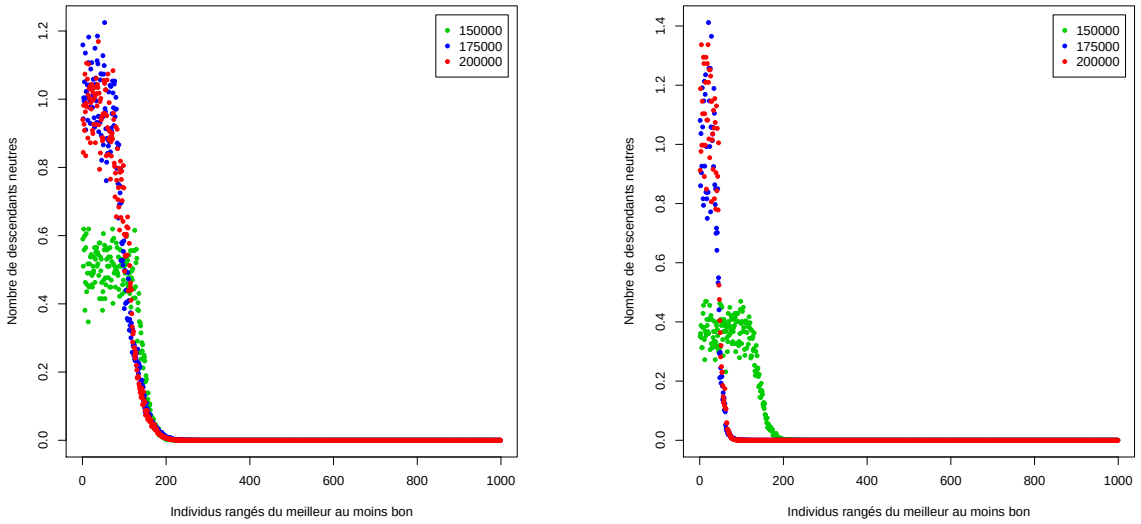
sont issues de Knibbe (2006). Ainsi, pour les mutations locales (substitutions, petites insertions et petites délétions), $\widetilde{\nu}_{ponct} = \widetilde{\nu}_{ins} = \widetilde{\nu}_{del} = 1 - \%L_1$. La probabilité qu'une mutation locale n'affecte aucune région codante est $P(\textit{mutation locale neutre}) = P(\textit{neutre}|\textit{mutation ponctuelle})P(\textit{mutation locale} = \textit{mutation ponctuelle}) + P(\textit{neutre}|\textit{insertion})P(\textit{mutation locale} = \textit{insertion}) + P(\textit{neutre}|\textit{délétion})P(\textit{mutation locale} = \textit{délétion}) = \widetilde{\nu}_{ponct}\frac{u_{ponct}}{u_{ponct}+u_{ins}+u_{del}} + \widetilde{\nu}_{ins}\frac{u_{ins}}{u_{ponct}+u_{ins}+u_{del}} + \widetilde{\nu}_{del}\frac{u_{del}}{u_{ponct}+u_{ins}+u_{del}} = 1 - \%L_1$. Pour les réarrangements (inversions, translocations, grandes délétions et duplications), la probabilité qu'un réarrangement n'affecte aucune région fonctionnelle est $P(\textit{réarrangement neutre}) = P(\textit{neutre}|\textit{grande délétion})P(\textit{réarrangement} = \textit{grande délétion}) + P(\textit{neutre}|\textit{duplication})P(\textit{réarrangement} = \textit{duplication}) + P(\textit{neutre}|\textit{inversion})P(\textit{réarrangement} = \textit{inversion}) + P(\textit{neutre}|\textit{translocation})P(\textit{réarrangement} = \textit{translocation}) = \widetilde{\nu}_{gdel}\frac{u_{gdel}}{u_{gdel}+u_{dupl}+u_{inv}+u_{trans}} + \widetilde{\nu}_{dupl}\frac{u_{dupl}}{u_{gdel}+u_{dupl}+u_{inv}+u_{trans}} + \widetilde{\nu}_{inv}\frac{u_{inv}}{u_{gdel}+u_{dupl}+u_{inv}+u_{trans}} + \widetilde{\nu}_{trans}\frac{u_{trans}}{u_{gdel}+u_{dupl}+u_{inv}+u_{trans}}$. Or, dans nos simulations, $u_{gdel} = u_{dupl} = u_{inv} = u_{trans} = u_{rear}$. Ainsi, $P(\textit{réarrangement neutre}) = \frac{u_{rear}}{4u_{rear}}(\widetilde{\nu}_{gdel} + \widetilde{\nu}_{dupl} + \widetilde{\nu}_{inv} + \widetilde{\nu}_{trans})$. De plus, comme $\widetilde{\nu}_{inv} = (1 - \%L_1)^2$, $\widetilde{\nu}_{trans} = (1 - \%L_1)^3$, $\widetilde{\nu}_{gdel} = \frac{1}{2L^2}\sum_{i=1}^{N_G}(\lambda_i\lambda_i + 1)$, $\widetilde{\nu}_{dupl} = (1 - \%L_1)\frac{1}{2L^2}\sum_{i=1}^{N_G}\lambda_i(\lambda_i + 1)$, $P(\textit{rarrangementneutre}) = \frac{1}{4}(2 - \%L1)\left((1 - \%L1)^2 + \frac{1}{2L^2}\sum_{i=1}^{N_G}\lambda_i(\lambda_i + 1)\right)$.

au cours de l'évolution sont celles qui présentent un compromis entre la robustesse de réplication (peu de mutations/réarrangements) et l'évolvabilité (capacité à "innover"). Ce compromis est atteint lorsque les individus ont environ un descendant neutre, c'est-à-dire un descendant n'ayant subi aucune mutation ou seulement des mutations neutres, permettant ainsi de transmettre la qualité des individus parents. Plus formellement, une lignée sera conservée tout au long des générations si $F_\nu W \sim 1$, avec F_ν la proportion de descendants neutres et W le nombre de descendants.

Au moment où les taux de mutation/réarrangement augmentent, la proportion de descendants neutres diminue : il est plus difficile de reproduire à l'identique un individu. Ainsi, à la génération 150 000, c'est-à-dire au moment du changement de paramètres, la proportion de descendants neutres F_ν des individus se reproduisant est significativement inférieure dans les simulations des scénarios par rapport à celles de contrôle ($P = 9.77 \cdot 10^{-4}$, tests des rangs signés de Wilcoxon) sur les même populations (après les changements, avant le début de l'évolution). Le nombre de descendants neutres pour l'ensemble des individus de la population est ainsi réduit (Figure IV.6).

Si le principal moteur de l'évolution de la taille des génomes est le compromis entre la robustesse et l'évolvabilité, $F_\nu W$ devrait ré-augmenter pour les scénarios d'augmentation des taux de mutation et d'augmentation des taux de réarrangement. De fait, en fin de simulation, $F_\nu W$ a retrouvé sa valeur d'environ 1 descendant neutre par génération (Figure IV.6). Cependant, les valeurs de F_ν des individus reproducteurs restent inférieures à celles des simulations de contrôle ($P = 4.89 \cdot 10^{-3}$ pour le scénario d'augmentation des taux de mutation, $P = 9.77 \cdot 10^{-4}$ pour le scénario d'augmentation des taux de réarrangement, tests des rangs signés de Wilcoxon). Simultanément, la *proportion* de descendants neutres a augmenté, principalement pour le scénario d'augmentation des taux de mutation. Comme mentionné précédemment, la probabilité qu'une mutation locale n'affecte aucune région fonctionnelle dépend directement de la proportion de bases codantes. Le principal levier d'action est donc le non codant. L'augmentation des taux de mutation entraîne donc des génomes plus compacts (Figure IV.5). Ce n'est cependant pas le cas dans le scénario d'augmentation des taux de réarrangement, où la taille des génomes s'est réduite par la perte de gènes et de non codant, tout en conservant des proportions de bases non codantes similaires à celles observées dans les simulations de contrôle (Figure IV.5). Bien que supérieures aux valeurs à la génération 150 000, les proportions de descendants neutres des individus reproducteurs en fin de simulation restent inférieures dans le scénario d'augmentation des taux de réarrangement par rapport au scénario d'augmentation des taux de mutation ($P = 9.77 \cdot 10^{-4}$, tests des rangs signés de Wilcoxon).

Dans le scénario d'augmentation des taux de réarrangement, l'augmentation de $F_\nu W$ s'est plutôt effectuée par l'augmentation de W des individus se reproduisant (valeurs significativement supérieures à celles des simulations de contrôle et du scénario d'augmentation des taux de mutation, $P = 9.77 \cdot 10^{-4}$, test des rangs signés de Wilcoxon). Or, la somme des W sur l'ensemble de la population est fixée au nombre d'individu (1000) tout au long des générations. L'augmentation de W des individus se reproduisant diminue ainsi le nombre de reproducteur (valeurs significativement inférieures à celles des simulations de contrôle et d'augmentation des taux de mutation, $P = 9.77 \cdot 10^{-4}$ et $P = 2.94 \cdot 10^{-3}$,



(a) Augmentation des taux de mutation **(b)** Augmentation des taux de réarrangement

Figure IV.6 – Nombre de descendants neutres pour l'ensemble des individus des populations aux générations 150 000, 175 000 et 200 000 pour le scénario d'augmentation des taux de mutation et le scénario d'augmentation des taux de réarrangement
Le nombre de descendants neutres est calculé *a posteriori* de l'évolution par multiplication de deux caractéristiques d'un individu à une génération donnée : le nombre de descendants, estimé à partir de la distribution des erreurs métaboliques des individus de la population et la proportion de descendants neutres. Cette dernière est estimée sur 1 000 essais reproductifs de chaque individu de la population et en déterminant dans ces 1 000 descendants virtuels, la proportion dont l'erreur métabolique est identique à l'erreur métabolique parentale.
Les couleurs correspondent aux différentes générations avec en vert 150 000, en bleu 175 000 et en rouge 200 000.
Une seule des 10 simulations est représentée dans les deux cas, les autres ayant des tendances similaires.

test des rangs signés de Wilcoxon).

Avec la multiplication par deux des taux de réarrangement, les génomes subissent en moyenne plus de réarrangements. Ainsi, juste après le changement de paramètres, les meilleurs individus des simulations de contrôle sont sujets à 0.89 ± 0.06 réarrangements de chaque type (duplication, inversion, translocation, grande délétion) par génération[1]. Or, la probabilité qu'un réarrangement n'affecte aucune région fonctionnelle est nettement plus faible que celle pour une mutation locale et ne dépend pas seulement de la quantité de non codant mais aussi du nombre de gènes, de la distance entre les gènes, de la longueur du génome $(\frac{1}{4}(2 - \%L1)\left((1 - \%L1)^2 + \frac{1}{2L^2}\sum_{i=1}^{N_G}\lambda_i(\lambda_i + 1)\right))$. Il est alors plus difficile d'augmenter la probabilité de ne pas dégrader la fitness des individus en modifiant seulement la compaction de génomes déjà très compact, sans aucune perte de fitness, les

[1]Nous avons aussi testé ce scénario avec une multiplication par 5 des taux de réarrangement. Dans ce cas, la taille des génomes fluctue avec des phases d'explosion des génomes et des phases de forte réduction. Le nombre de chaque type de réarrangement (2.25 ± 0.15 au moment du changement de paramètres) est alors trop important pour maintenir une lignée avec un descendant neutre. Nous avons donc choisi une augmentation moins forte des taux de réarrangement.

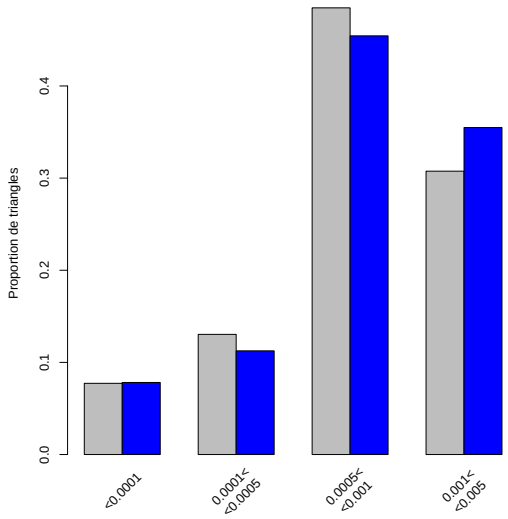


Figure IV.7 – Distribution de l'aire des triangles des meilleurs individus finaux pour les simulations de contrôle et celles du scénario de diminution de la pression de sélection
Les données pour les simulations de contrôle sont en gris et celles du scénario de diminution de la pression de sélection en bleu.

réarrangements étant moins fins que les mutations locales pour ajuster la quantité d'ADN non codant. Les modifications des génomes sont alors globales (nombre de gènes, etc) mais sont marginales par rapport à celles pour les simulations du scénario d'augmentation des taux de mutation (Figure IV.1). Les changements touchent alors principalement la structure de la population avec des reproducteurs moins nombreux mais bien meilleurs que le reste de la population.

IV.2.2 Diminution de la pression de sélection

Dans le scénario de diminution de la pression de sélection, les génomes subissent une évolution réductive avec des pertes de gènes (Figure IV.1c). Sous l'effet de ce scénario, les gènes conservés tendent à coder pour des triangles de plus grande aire (Figure IV.7).

En diminuant k, la population est homogénéisée en terme de probabilité de reproduction (Figure III.6), la probabilité de reproduction d'un individu étant $e^{-kg}/\sum_{i=1}^{N} e^{-kg_i}$ avec g l'erreur métabolique et $N = 1000$ le nombre d'individus de la population. Des individus mal-adaptés peuvent donc se reproduire au détriment des bons individus. Les mutations locales et les réarrangements légèrement délétères ne sont pas nécessairement éliminés. Ainsi, les évènements fixés sont en proportion plus délétères et l'impact moyen d'un évènement délétère est plus fort dans le scénario de diminution de la pression de sélection que dans les simulations de contrôle (Tableau IV.3). En conséquence, les gènes ayant de petits impacts sur le phénotype ont plus de chances que les autres d'être perdus par dérive génétique que les autres (Figure IV.7).

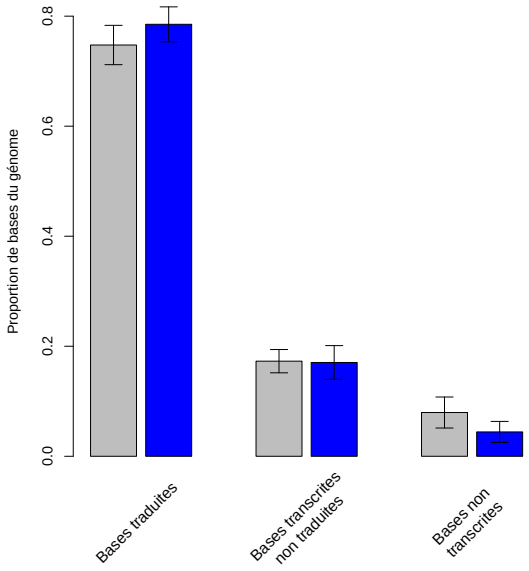
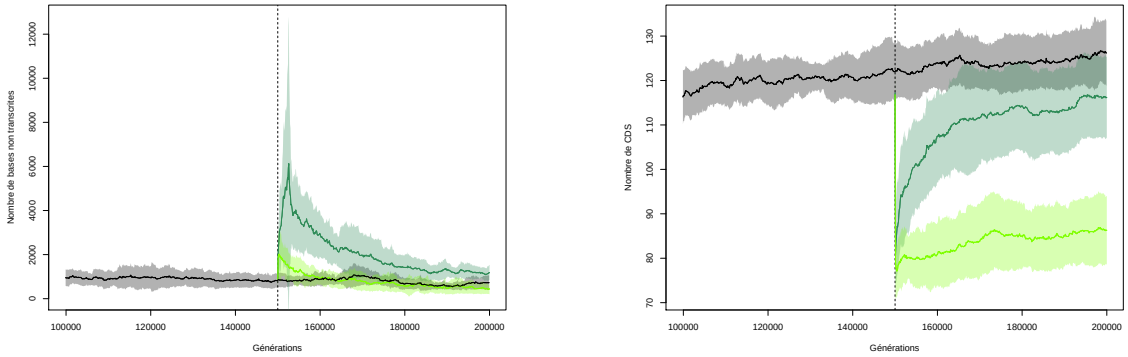


Figure IV.8 – Répartition des bases des génomes chez les ancêtres communs des populations finales pour les simulations de contrôles et celles de diminution de la pression de sélection
Les valeurs représentées sont les moyennes sur les 10 simulations pour chaque type de simulation avec l'écart-type sur les 10 simulations symbolisé par la barre d'erreur.
Les données pour les simulations de contrôle sont en gris et celles du scénario de diminution de la pression de sélection en bleu.

La cause principale de la réduction des génomes n'est cependant pas la perte de gènes, mais la perte de bases non codantes. Ainsi, la proportion de bases non transcrites est inférieure à celle des simulations de contrôles (Figure IV.8). Cette réduction du non codant s'explique par l'augmentation de la proportion de descendants neutres des individus se reproduisant. Le nombre de descendants d'un bon individu décroit avec la diminution de k et l'homogénéisation des probabilités de reproduction. Ainsi, en fin de simulation, les nombres moyens de descendants des individus pouvant se reproduire sont inférieurs aux valeurs observées dans les simulations de contrôle ($P = 9.77 \cdot 10^{-4}$, test de rangs signés de Wilcoxon). Les lignées gagnantes sont celles qui ont rétabli $F_\nu W \sim 1$ en compensant la baisse de W par une augmentation de F_ν, elle-même permise par la perte de bases non codantes. Ainsi, dans ce scénario, l'évolution réductive passe donc principalement par la réduction du non codant.

IV.2.3 Changement de niche

Contrairement aux autres scénarios détaillés, aucun des scénarios de changement de niche (déplacement, suppression ou neutralisation d'un lobe de l'environnement) n'induit une évolution réductive similaire à celle observée chez *Prochlorococcus* (Tableau IV.2), pourtant les principales hypothèses proposées pour l'évolution réductive chez *Prochlorococcus* reposent sur un tel changement de niche.



(a) Nombre de bases non transcrites **(b)** Nombre de CDS

Figure IV.9 – Évolution du nombre de bases non transcrites et du nombre de gènes au cours des simulations de contrôle, de celles du scénario de déplacement d'un lobe de l'environnement et de celles du scénario de suppression d'un lobe de l'environnement
Les lignes représentent les moyennes pour les 10 simulations de chaque type de simulation et la plage colorée à l'écart-type sur les 10 simulations.
Les simulations de contrôle sont en gris, celles du scénario de déplacement d'un lobe de l'environnement en vert foncé et celles du scénario de suppression d'un lobe de l'environnement en vert clair.

Dans les scénarios de suppression et de neutralisation d'un lobe de l'environnement, les génomes se réduisent mais la proportion de bases codantes n'augmente pas et la taille des gènes n'est pas réduite (Tableau IV.2). Les principales modifications des structures génomiques ont lieu peu après les changements d'environnement (Figure IV.2), mais avec des dynamiques un peu différentes entre les deux scénarios (Section IV.1). La réduction observée des génomes (bien que différente de celle observée chez *Prochlorococcus*) est liée à la perte de gènes devenus inutiles, car ils codaient pour des triangles situés dans le lobe supprimé ou neutralisé. Ces gènes sont un fardeau mutationnel et sont donc éliminés, par la pseudogénéisation puis l'érosion progressive des séquences. La proportion d'ADN codant commence par diminuer fortement, au moment de la perte des gènes par pseudogénéisation et une augmentation concomitante du nombre de bases non codantes, puis augmente au fur et à mesure de l'érosion des séquences et pourrait atteindre un niveau similaire, voire supérieur, à celui des simulations de contrôle avec plus de générations d'évolution.

Le scénario de déplacement d'un lobe de l'environnement n'induit pas de réduction des génomes (Tableau IV.2). Ainsi, la taille des génomes des individus le long de la lignée ancestrale du meilleur individu final augmente fortement peu après le déplacement du lobe, principalement par l'augmentation des bases non codantes (bases non transcrites) (Figure IV.9a). Ces augmentations sont précédées de nombreuses pertes de gènes (Figure IV.9b), les gènes codant des triangles remplissant le lobe à son ancienne position. De nouveaux gènes sont ensuite acquis progressivement (Figure IV.9b), pour remplir le lobe à sa nouvelle position. En moins de 500 générations, les triangles remplissant le lobe à son ancienne position sont perdus et des nouveaux triangles remplissent peu à peu le lobe à sa nouvelle position (Figure IV.10).

Ainsi, lors du déplacement du lobe, les gènes ne sont pas modifiés dans leur séquence afin

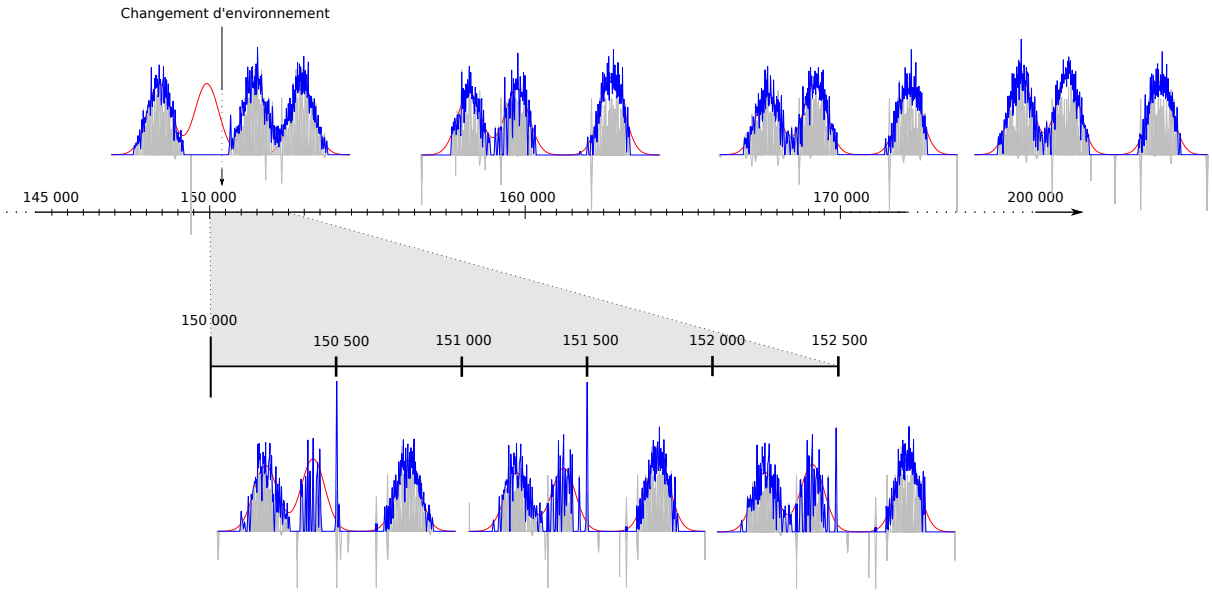


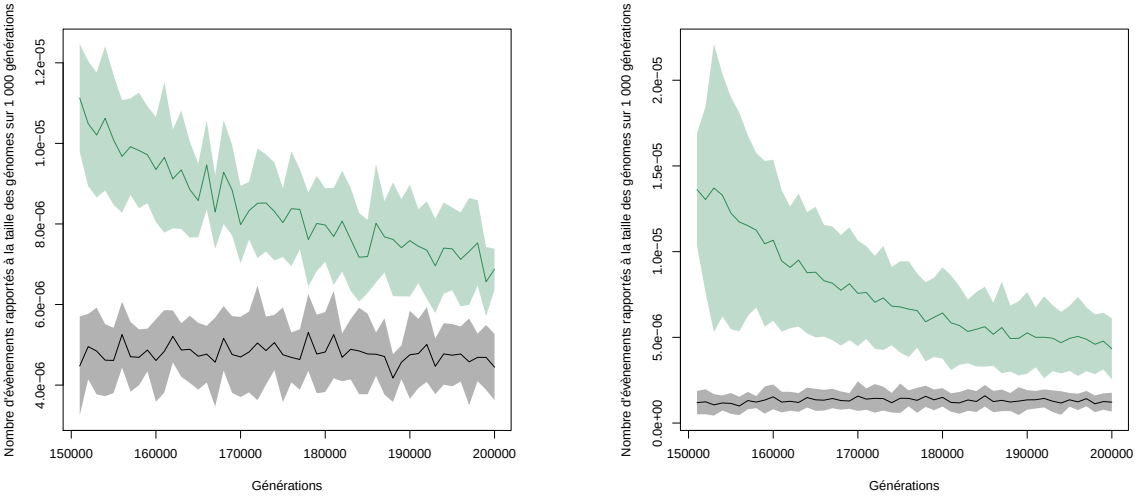
Figure IV.10 – Évolution du phénotype au cours d'une simulation du scénario de déplacement d'un lobe de l'environnement
La courbe en rouge correspond à la cible environnementale, la courbe en bleu au phénotype de l'individu représenté. En gris, sont représentés les différents triangles correspondant aux gènes de l'individu représenté.
L'individu représenté est le meilleur individu de la génération étudiée.

de suivre le lobe[1]. Les gènes sont éliminés et de nouveaux gènes sont créés. Les pertes de gènes dans les premières générations après le déplacement du lobe sont en effet similaires à celles observées dans le scénario de suppression d'un lobe (Figure IV.9b). Les gènes à l'ancienne position du lobe sont éliminés rapidement car ils sont fortement délétères pour la fitness des individus : ils éloignent le phénotype de la cible environnementale. La quantité de bases non codantes augmente du fait de ces pseudogénéisations, mais aussi parce que les nouveaux gènes nécessaires pour remplir le lobe à sa nouvelle position sont principalement acquis par duplication de gènes existants (voir le nombre de réarrangements fixés, Figure IV.11b). Or, dans un évènement de duplication de gène, le gène en question est rarement dupliqué seul : il est accompagné du non codant l'entourant.

À partir d'un certain point (environ 2 500 générations après le déplacement du lobe), les principaux gènes permettant de remplir le lobe à sa nouvelle position ont été acquis (Figure IV.10). Les acquisitions de gènes continuent (Figure IV.9b), probablement des gènes codant des triangles dont l'aire est plus petite afin d'affiner le remplissage du lobe. À ce moment là, la grande quantité de non codant présent dans les génomes est une source d'instabilité mutationnelle, dangereux une fois la nouvelle cible approchée. Le non codant s'érode donc peu à peu (Figure IV.9a), ce qui entraîne une baisse du nombre de mutations et de réarrangements fixés (Figure IV.11).

En fin de simulation, à la génération 200 000, l'état stable ne semble pas encore atteint.

[1]La position des triangles sur l'axe est déterminée par le paramètre m qui est codé dans la séquence du gène (Figure III.1)



(a) Mutations locales **(b)** Réarrangements

Figure IV.11 – Évolution du taux de mutation et réarrangement fixés au cours des simulations de contrôle et de celles du scénario de déplacement d'un lobe de l'environnement
Est représenté le nombre d'évènements rapporté à la taille des génomes sur 1000 générations le long de la lignée ancestrale du meilleur individu en fin de simulation, les lignes représentant les moyennes pour les 10 simulations de chaque type de simulation et la plage colorée à l'écart sur les 10 simulations. Les simulations de contrôle sont en gris et celles du scénario de déplacement d'un lobe de l'environnement en vert.

Les taux de mutation et réarrangement fixés sont encore élevés (Figure IV.11). L'érosion du non codant continue tout comme l'acquisition de gènes (Figure IV.9). Les proportions de bases codantes restent inférieures à celles dans les simulations de contrôle (Tableau IV.2). Ces valeurs pourraient probablement atteindre celles des simulations de contrôle si les simulations étaient prolongées pendant quelques milliers de générations.

⋆ ⋆ ⋆

Dans les scénarios d'augmentation des taux de mutation et de réarrangement, la principale cause de l'évolution réductive semble être la sélection indirecte de $F_\nu W \sim 1$ après la réduction de F_ν due à l'augmentation des taux de mutation et de réarrangement (Figure IV.6). C'est aussi le cas pour les autres scénarios étudiés (Figure IV.12), même si les dynamiques, les points de départ et d'arrivée changent. Ainsi, en fin de simulation (à 200 000), tous les scénarios ont atteint $F_\nu W \sim 1$ (Figure IV.12). Pour les simulations de contrôle, les changements sont moindres entre les générations 150 000 et 200 000 et sont toujours dans la même zone du graphique (Figure IV.12).

Pour les scénarios d'augmentation des taux de mutation et de réarrangement, les trajectoires entre 150 000 et 200 000, démarrant à 150 000 par une proportion de descendants neutres réduite par rapport aux simulations de contrôle (Figure IV.12), se rapprochent rapidement de $F_\nu W \sim 1$ par l'augmentation de W mais ne s'arrêtent pas quand $F_\nu W \sim 1$ est atteint. Ainsi, pour le scénario d'augmentation des taux de mutation, après l'augmen-

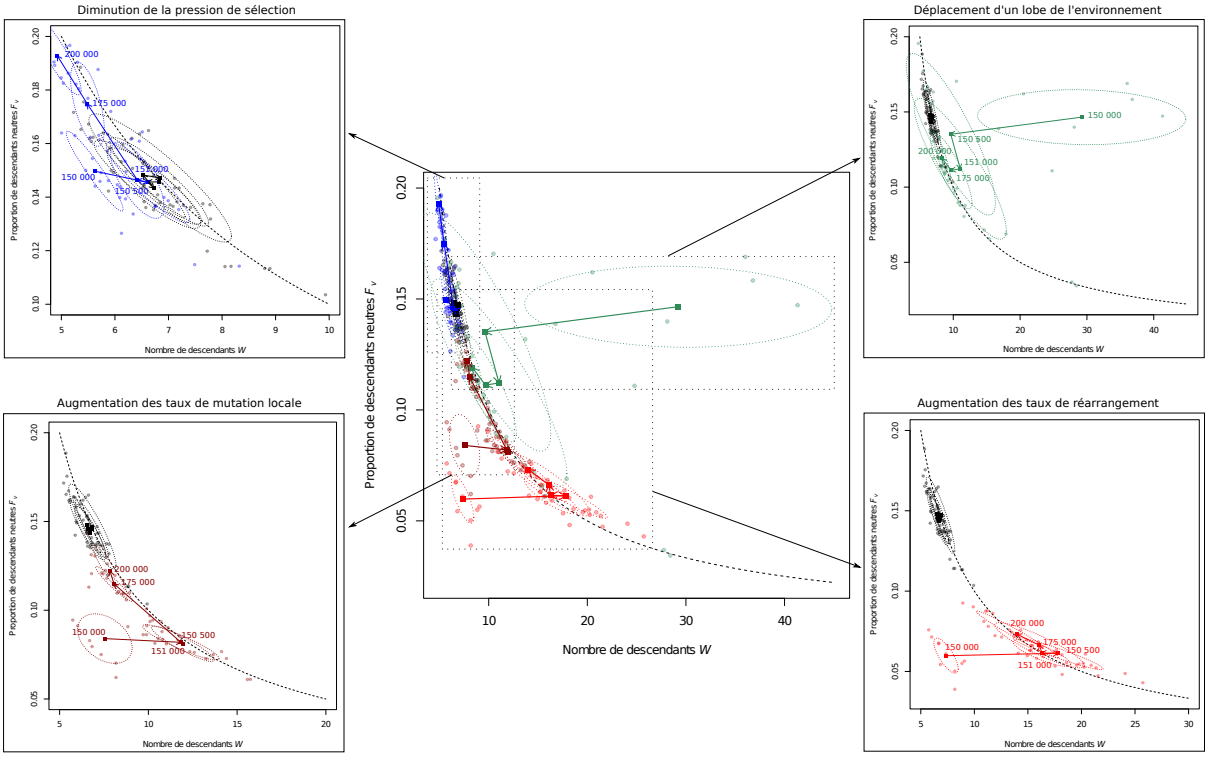


Figure IV.12 – Évolution de la relation entre le nombre de descendants et la proportion de descendants neutres pour les scénarios d'augmentation des taux de mutation, d'augmentation des taux de réarrangement, de diminution de la pression de sélection et de déplacement d'un lobe de l'environnement

Le nombre de descendants W et la proportion de descendants neutres F_ν sont calculés pour les dix simulations de contrôle et des scénarios (points en transparence) pour les générations 150 000, 150 500, 151 000, 175 000 et 200 000. Les ellipses englobent 95% des points des 10 simulations d'un scénario à une génération donnée dont le centre de l'ellipse (point carré) est la moyenne sur les 10 simulations. Les flèches relient les moyennes pour les 10 simulations entre les différentes générations dans l'ordre chronologique. La droite en pointillés correspond à $F_\nu W = 1$.

Les simulations de contrôle sont en noir, celles du scénario d'augmentation des taux de mutation en rouge foncé, celles du scénario d'augmentation des taux de réarrangement en rouge, celles du scénario de diminution de la pression de sélection en bleu et celles du scénario de déplacement d'un lobe de l'environnement en vert. Pour chaque scénario, un zoom est fait sur les données liées à ce scénario, en éliminant les autres scénarios (graphiques en haut et en bas).

tation de W, F_ν augmente puis W diminue pour se rapprocher des simulations de contrôle. La tendance est similaire pour l'augmentation des taux de réarrangement mais dans une moindre mesure. Ainsi, à 200 000, les valeurs de W et F_ν sont différentes pour les deux scénarios. Les simulations du scénario d'augmentation des taux de mutation sont proches des contrôles alors que celles du scénario d'augmentation des taux de réarrangement sont encore éloignées (Figure IV.12).

Pour le scénario de diminution de la pression de sélection, le changement initial (à 150 000) vient d'une diminution de W due à l'homogénéisation de la reproduction au sein de la population (impact de k). $F_\nu W \sim 1$ est rapidement atteint avec des valeurs proches de celles des simulations de contrôle (Figure IV.12). Cependant, les simulations du scénario s'éloignent de celles du contrôle avec une diminution de W et une augmentation de F_ν.

Dans le scénario de déplacement d'un lobe de l'environnement, le changement brusque de l'environnement induit une perte d'efficacité des individus de la population vis à vis de la tâche à accomplir. Quelques individus, probablement mauvais avant le changement, sont capables de se reproduire, entraînant une augmentation forte de W et un éloignement de $F_\nu W \sim 1$ (Figure IV.12). Le rapprochement de $F_\nu W \sim 1$ se fait d'abord par la diminution de W : les "bons" individus se sont répandus dans la population, la reproduction est donc plus homogène. La population s'adapte aussi au nouvel environnement. Cependant, cette adaptation s'accompagne d'une augmentation du non codant, liée à une diminution de F_ν. Lorsque les individus sont adaptés au nouvel environnement, ils perdent le non codant et F_ν augmente. Les valeurs de F_ν et W sont alors proches de celles des simulations de contrôle.

Atteindre le compromis entre la robustesse et l'évolvabilité semble donc être un moteur important de l'évolution des structures génomiques après des changements de paramètres. Cependant, toutes les valeurs F_ν et W telles que $F_\nu W \sim 1$ ne semblent pas convenir. Ainsi, bien que $F_\nu W \sim 1$ soit atteint pour les scénarios à 175 000 (voire avant), les valeurs de F_ν et W à 200 000 sont différentes de celles à 175 000. Il semble y avoir des couples (F_ν, W) optimaux, qui peuvent différer selon les paramètres de simulation. Ainsi, les scénarios tendent vers les valeurs des simulations de contrôle. Cependant, dans le scénario de diminution de la pression de sélection, les valeurs de contrôle sont atteintes rapidement mais les valeurs F_ν et W changent ensuite pour s'éloigner des valeurs des simulations de contrôle : l'optimum du couple (F_ν, W) tel que $F_\nu W \sim 1$ semble donc différent lorsque k change. Il serait possible que cela soit aussi le cas pour les autres scénarios si les simulations étaient continuées pendant quelques milliers de générations supplémentaires.

Chapitre V

Scénarios avec régulation

Les conditions dans les eaux tropicales et subtropicales où vivent les souches de *Prochlorococcus* ne changent pas significativement tout au long de l'année, contrairement aux eaux tempérées de *Synechococcus*. De plus, les concentrations en nutriments dans les eaux de surface tropicales et subtropicales sont constamment faibles. Dans cet environnement, une machinerie de régulation sophistiquée et potentiellement coûteuse, utile pour répondre aux variations des concentrations de nutriments, peut être perdue à faibles coûts. De fait, chez *Prochlorococcus*, certains gènes impliqués dans la régulation ont été perdus (Tableau I.1) (Rocap *et al.*, 2003; Dufresne *et al.*, 2003; García-Fernández *et al.*, 2004). La réduction des besoins en régulation fine aurait eu lieu au moment de la divergence entre *Prochlorococcus* et *Synechococcus* du fait de la stabilisation de l'environnement et pourrait ainsi avoir initié l'évolution réductive.

Dans les scénarios précédents, nous avons testé la stabilisation de l'environnement avec *aevol*, c'est-à-dire en l'absence de régulation. Dans ce scénario, les génomes se réduisent par la perte de bases non codantes mais le nombre de gènes augmente (Tableau IV.2). Ce scénario n'induit donc pas une évolution réductive similaire à celle observée chez *Prochlorococcus*. Cependant, la perte des mécanismes de régulation fine induite par un changement d'environnement semble être un scénario plausible dans le cas de *Prochlorococcus*. Nous le testons ici avec une extension d'*aevol*, *raevol*, qui inclut un processus explicite de régulation transcriptionnelle. Cette extension est actuellement en développement et les temps de simulation sont nettement plus longs que ceux avec *aevol*. De plus, pour des scénarios n'impliquant pas directement la régulation (taux de mutation, structure de la population, ...), il a été montré que les changements génomiques observés sont similaires entre *aevol* et *raevol* (Sanchez-Dehesa, 2009; Beslon *et al.*, 2010). Devant l'impossibilité pratique de conduire de multiples expériences avec *raevol*, seuls deux scénarios sont testés ici. Ces deux scénarios correspondent respectivement à une simplification et à un arrêt de la variation de l'environnement dans lequel évoluent les bactéries artificielles. Ce travail a été réalisé en collaboration avec Yoram Vadee Le Brun, dans le cadre de son doctorat.

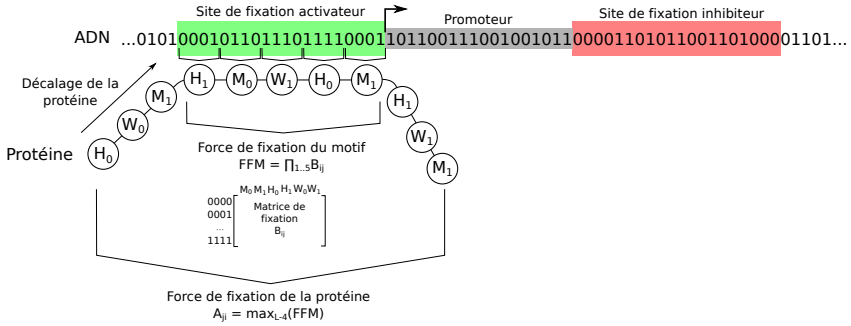


Figure V.1 – Calcul de l'affinité entre les facteurs de transcription et les sites de régulation
La séquence primaire de la protéine se place devant les 20 bases d'un site de régulation et tous les motifs de 5 acides aminés de long sont testés. Pour chaque couple (acide aminé, quadruplet de bases), la valeur de fixation B_{ij} est lue dans la matrice de fixation B. La force de fixation du motif est la multiplication des 5 valeurs de B_{ij}. On considère ensuite que la force de fixation de la protéine est celle de son motif le plus fortement fixable.

V.1 *raevol* : modélisation de l'évolution des réseaux de régulation dans *aevol*

raevol est une extension d'*aevol* incluant un processus explicite de régulation transcriptionnelle, initialement développée par Yolanda Sanchez-Dehesa durant sa thèse (Sanchez-Dehesa, 2009). Dans *raevol*, un modèle de régulation procaryotique est ajouté dans la chimie artificielle d'*aevol* pour modéliser les interactions entre des facteurs de transcription et les promoteurs. Toutes les autres composantes de *raevol* sont similaires à celles d'*aevol* et ne sont donc pas décrites ici.

Chez les procaryotes, la transcription de l'ADN en ARN peut être augmentée ou diminuée par des facteurs de transcription qui se fixent à l'ADN sur des sites proches des promoteurs des ARNs, facilitant ou réprimant la transcription par l'ARN polymérase. La concentration des ARNs et donc les taux de traduction des gènes et la concentration des protéines correspondantes peuvent ainsi varier au cours de la vie de l'organisme.

Dans *raevol*, les sites de fixation sont des séquences de 20 paires de bases et encadrent directement les promoteurs (Figure V.1). En amont du promoteur, le site de fixation est activateur : les facteurs de transcription qui s'y fixent augmentent l'activité transcriptionnelle de l'ARN. En aval du promoteur, le site de fixation est inhibiteur : tout facteur de transcription fixé sur ce site diminue l'activité transcriptionnelle du promoteur correspondant. Toute protéine encodée dans le génome peut être un facteur de transcription et se fixer sur des sites de fixation, à condition que sa séquence contienne au moins un domaine de régulation, c'est-à-dire un motif de cinq acides aminés ayant une force de fixation strictement positive sur un ou plusieurs sites de fixation du génome (Figure V.1). L'affinité élémentaire entre un acide aminé et un quadruplet de bases est déterminée par une matrice d'affinité, matrice de 7 acides aminés et 16 quadruplets initialisée suivant une distribution fixée en paramètres (Sanchez-Dehesa, 2009). Toutes les simulations présentées

ici ont été réalisées avec la même matrice, initialisée avec un tirage uniforme de valeurs entre 0 et 1 puis un certain pourcentage de valeurs sont fixées à zero. Ce pourcentage, ici 75%, permet de gérer la proportion de motifs qui ont une capacité régulatrice.

L'activité transcriptionnelle d'un promoteur peut potentiellement être régulée par l'ensemble des protéines du protéome. Sans régulation, l'expression basale β_i d'un ARN i (et donc de ses gènes) correspond au niveau d'expression défini dans *aevol*, c'est-à-dire la distance entre le promoteur et le consensus : $\beta_i = 1 - {}^{d_i}/_{(1+d_{max})}$ avec d_i la distance de Hamming entre la séquence du promoteur et la séquence consensus et $d_{max} = 4$ la distance de Hamming maximale admise entre un promoteur et le consensus. En présence de protéines régulatrices, l'activité transcriptionnelle du promoteur dépend aussi de :

$$A_i(t) = \sum_j c_j(t) A_{ji} \text{ et } I_i(t) = \sum_j c_j(t) A_{ji} \tag{V.1}$$

avec $A_i(t)$ l'activité activatrice des facteurs de transcription sur le promoteur i au temps t, $I_i(t)$ l'activité inhibitrice des facteurs de transcription sur le promoteur i au temps t, $c_j(t)$ la concentration de la protéine j au temps t, A_{ji} l'affinité de la protéine j avec le site activateur du promoteur i et I_{ji} l'affinité de la protéine j avec le site inhibiteur du promoteur i. A_{ji} et I_{ji} ne dépendant que des séquences en amont et en aval du promoteur, ils sont constants au cours de la vie d'un individu et ne dépendent donc pas de t.

Le taux de transcription au temps t de l'ARN lié au promoteur i est alors calculé par une fonction de Hill dépendante du niveau basal, de l'activité activatrice et de l'activité inhibitrice :

$$e_i(t) = \beta_i \cdot \left(\frac{\theta^n}{I_i(t)^n + \theta^n} \right) \cdot \left(1 + \left(\frac{1}{\beta_i} - 1 \right) \cdot \left(\frac{A_i(t)}{A_i(t) + \theta^n} \right) \right) \tag{V.2}$$

avec θ et n des coefficients déterminant la forme de la fonction de Hill utilisée, constants durant les simulations.

La concentration des protéines varie alors selon une loi de synthèse-dégradation :

$$\begin{cases} c_i(0) = \beta_i \\ \frac{\partial c_i}{\partial t} = e_i(t) - \phi c_i(t) \end{cases} \tag{V.3}$$

avec ϕ le taux de dégradation constant dans le temps.

L'activité régulatrice de chaque protéine sur chaque promoteur est calculée à chaque pas de temps selon les équations V.1, V.2, V.3 avec un schéma d'intégration d'Euler explicite.

Dans *raevol*, la vie d'un individu est divisée en âges, eux-mêmes divisés en pas de temps (Figure V.2). Les concentrations sont actualisées à chaque pas de temps en fonction de

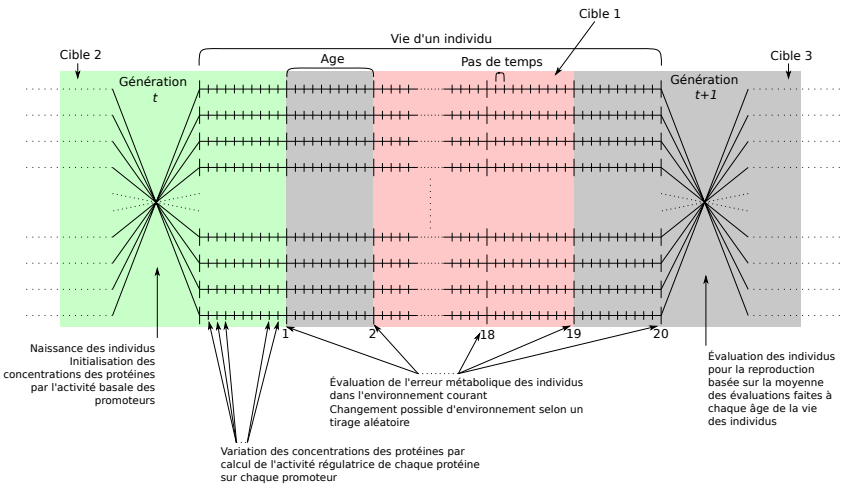


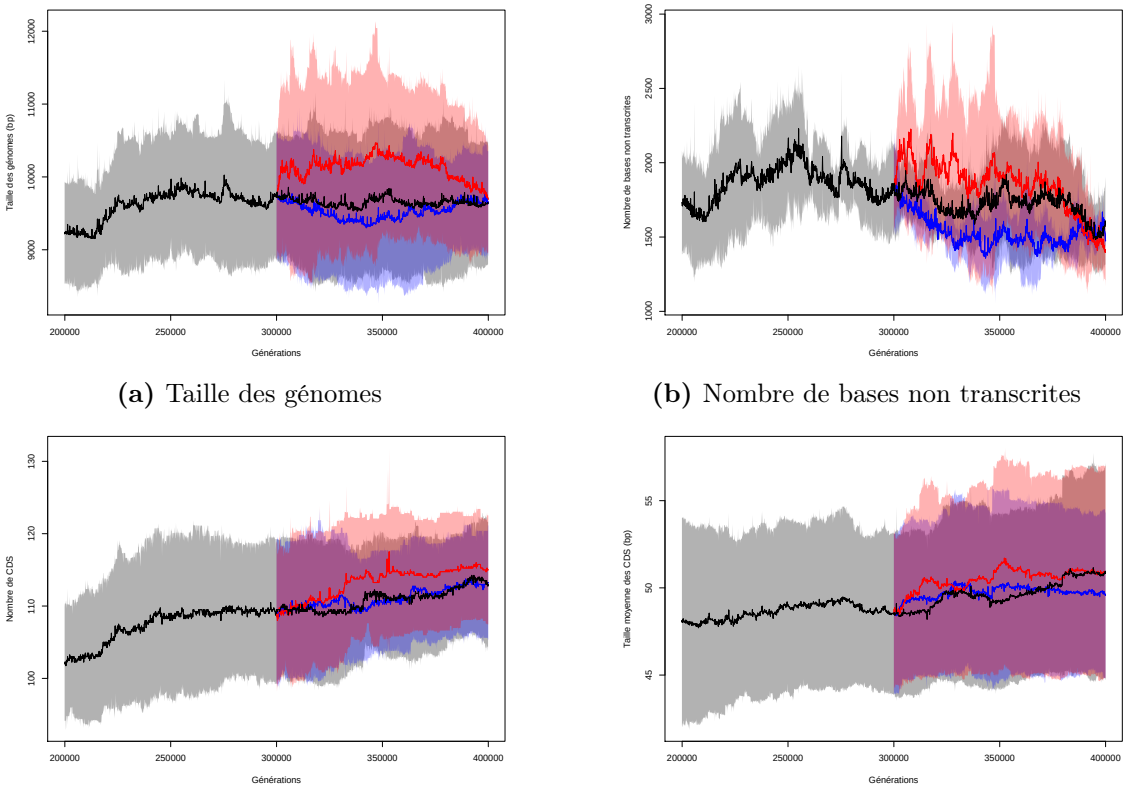
Figure V.2 – Notion de vie des individus dans *raevol*
Un individu naît au début d'une génération et meurt à la fin de cette génération. Sa vie est constituée de 20 âges, chacun divisé en 10 pas de temps. À chaque pas de temps, la concentration des protéines varie par calcul de l'activité régulatrice de chaque protéine sur chaque promoteur selon l'affinité de fixation. À chaque âge, l'individu est évalué en fonction de la cible courante, qui a une petite probabilité de changer à chaque âge. À la fin d'une génération, la probabilité de reproduction de l'individu est calculée en moyennant les évaluations faites aux différents âges.

l'équation V.3, ce qui permet de calculer le phénotype de l'individu à chaque "âge". Un individu est alors évalué à chaque "âge" par rapport à la cible courante et la moyenne des évaluations détermine sa capacité de reproduction à l'issue d'une génération. Comme les concentrations des protéines et le réseau de régulation sont mis à jour à chaque pas de temps (Figure V.2), les individus ont le temps d'adapter leurs réseaux métaboliques à la cible courante avant d'être évalués. De plus, à chaque "âge" des individus, la cible est susceptible de changer suivant des règles fixées pour chaque expérience.

Dans les simulations avec *aevol*, la cible fluctue à chaque génération selon un processus continu. Dans les simulations avec *raevol*, les variations ne sont pas continues : trois cibles différentes peuvent être rencontrées par les individus. L'apparition d'un environnement particulier est indiquée par l'activation d'une protéine signal extérieure aux génomes, qui peut avoir une activité régulatrice, si un individu acquiert un site de fixation pour cette molécule. Les changements de cible ne sont pas synchronisés avec les générations : la cible peut changer au cours de la vie des individus (Figure V.2). Quand elle change, elle est modifiée pour toute la population.

Les simulations avec *raevol* ont été lancées avec quatre graines du générateur aléatoire différentes pour les générateurs aléatoires, durant 300 000 générations, pour créer des souches selon la méthode proposée à la section III.2. La matrice de fixation ayant un impact fort sur les temps de construction des réseaux, la même matrice de fixation est utilisée pour les quatre simulations afin d'éliminer cet effet lors de l'analyse des scénarios.

Pour tester l'impact de la stabilisation et de la simplification de l'environnement, deux scénarios sont testés à partir des quatre simulations à la génération 300 000, pendant 100



(a) Taille des génomes

(b) Nombre de bases non transcrites

(c) Nombre de CDS

(d) Taille moyenne des CDS

Figure V.3 – Évolution de certaines caractéristiques génomiques des meilleurs individus pour les simulations de contrôle, du scénario de simplification de la variation de l'environnement et du scénario d'arrêt de la variation de l'environnement
Les lignes représentent les moyennes pour les 4 simulations de chaque scénario et la plage colorée à l'écart-type sur les 4 simulations.
Les données des simulations de contrôles sont en gris, celles du scénario de simplification de la variation de l'environnement en bleu et celles du scénario d'arrêt de la variation de l'environnement en rouge.

000 générations. Dans le premier scénario (simplification de la variation de l'environnement), seules deux des trois cibles sont présentées aux individus. Dans le second (arrêt de la variation de l'environnement), une seule cible est présentée (environnement constant).

V.2 Résultats

Aucun des deux scénarios testés avec *raevol* ne semble induire d'évolution réductive : la taille des génomes n'est pas réduite, ou alors seulement transitoirement (Figure V.3a).

Ainsi, aucune des caractéristiques génomiques analysées dans les scénarios précédents (taille des génomes, nombre de gènes, proportion de bases codantes, etc) n'est significativement différente dans les simulations des scénarios par rapport à celles de contrôle ($P > 0.05$, tests des rangs signés de Wilcoxon pour les 4 simulations sur des caracté-

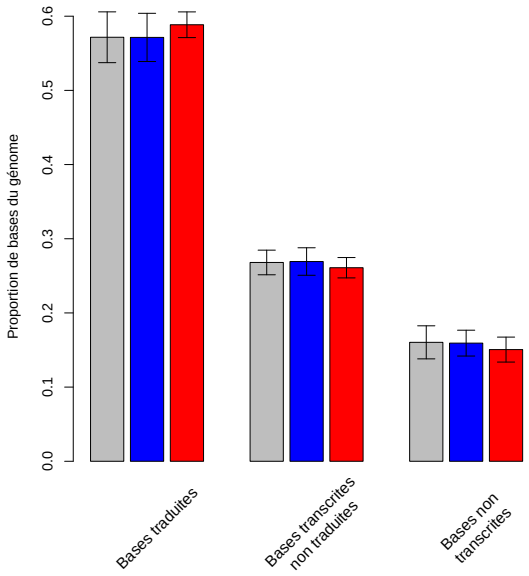
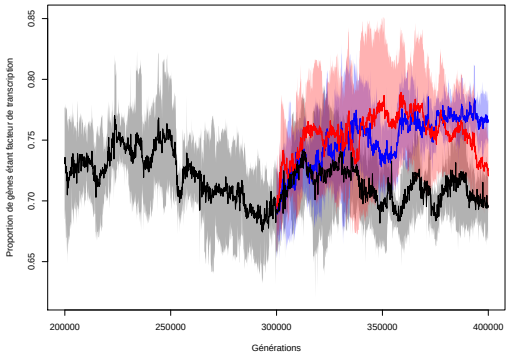


Figure V.4 – Répartition des bases des génomes, moyennée sur les meilleurs individus des 10 000 dernières générations pour les simulations de contrôle, du scénario de simplification de la variation de l'environnement et du scénario d'arrêt de la variation de l'environnement
Les valeurs représentées sont les moyennes sur les 4 simulations pour chacun des types de simulation avec l'écart-type sur les 4 simulations symbolisé par la barre d'erreur.
Les données des simulations de contrôles sont en gris, celles du scénario de simplification de la variation de l'environnement en bleu et celles du scénario d'arrêt de la variation de l'environnement en rouge.

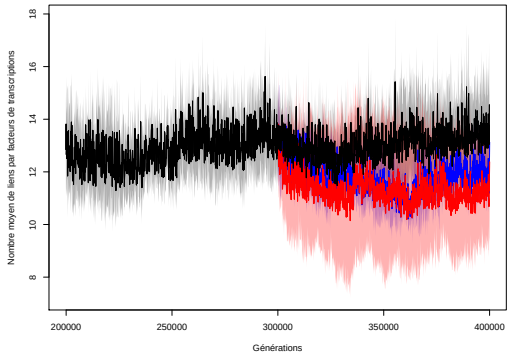
ristiques moyennées sur les 10 000 dernières générations pour les meilleurs individus). Contrairement à ce qui est observé dans les scénarios avec *aevol* (Tableau IV.2), la simplification et la stabilisation de la variation environnementale n'influent pas sur les caractéristiques génomiques lorsqu'un réseau de régulation génétique est présent. Ainsi, en présence d'organismes dotés de capacités régulatrices, c'est le réseau de régulation et non le génome qui semble absorber les variations dues à la modification de l'environnement (Figure V.5). Cependant, même sur les réseaux, les effets sont statistiquement faibles.

Dans les deux scénarios testés ici, les individus rencontrent au cours de leur existence moins de cibles (2 dans le cas de la simplification et 1 seule dans le cas de la stabilisation) que dans les simulations de contrôle (3 cibles). Ils devraient ainsi avoir moins besoin de réguler leurs gènes pour que ceux-ci s'adaptent aux différentes cibles rencontrées. Or, la proportion de gènes régulateurs (facteurs de transcription) augmente pour les deux scénarios (Figure V.5a), mais le nombre moyen de gènes régulés par un facteur de transcription diminue (Figure V.5b). Les deux effets s'équilibrent : le nombre de liens de régulation dans les réseaux est stable. Les réseaux sont donc composés de plus de nœuds pour un nombre total d'arêtes équivalent.

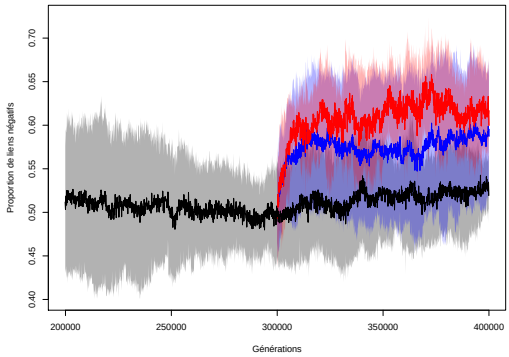
Dans les simulations de contrôle, les proportions de liens positifs et négatifs sont équivalentes alors que la proportion de liens négatifs augmente rapidement au moment du changement de variation environnementale pour les deux scénarios (Figure V.5c). Ainsi, dans les scénarios, les réseaux de régulation sont dominés par une inhibition des gènes.



(a) Proportion de gènes facteur de transcription



(b) Nombre moyen de liens par facteurs de transcription



(c) Proportion de liens négatifs

Figure V.5 – Évolution de certaines caractéristiques des réseaux des meilleurs individus pour les simulations de contrôle, du scénario de simplification de la variation de l'environnement et du scénario d'arrêt de la variation de l'environnement
Les lignes représentent les moyennes pour les 4 simulations de chaque scénario et la plage colorée à l'écart-type sur les 4 simulations.
Les données des simulations de contrôles sont gris, celles du scénario de simplification de la variation de l'environnement en bleu et celles du scénario d'arrêt de la variation de l'environnement en rouge.

Les scénarios ne donnent ainsi pas les résultats attendus de pertes de facteurs de transcription et des gènes correspondants quand les besoins de régulation diminuent. Ces observations peuvent être dues au fait que les simulations de contrôle ne sont pas vraiment à l'état stable malgré les 150 000 générations d'évolution supplémentaires par rapport aux simulations avec *aevol*. La construction de réseaux de régulation fins est bien plus longue que celle des génomes. Les scénarios ont donc peut-être été testés trop tôt dans l'évolution des simulations de contrôle et pas assez longtemps eux aussi. Ainsi, les tailles des génomes dans le scénario d'arrêt de la variation environnementale atteignent des valeurs proches de celles des simulations de contrôle en fin de simulation après une phase de croissance puis de réduction (Figure V.3a). 100 000 générations d'évolution pour les scénarios avec des réseaux de régulation est peut-être trop court pour observer des effets stables à la fois sur les caractéristiques des réseaux mais aussi sur les caractéristiques génomiques (les gènes liés à des facteurs de transcription devenus inutiles pourraient être perdus à terme). Une autre explication pourrait être la trop grande intrication du niveau "métabolique"

(triangles) et du réseau de régulation. En effet, dans *raevol*, toute protéine peut à la fois coder pour un triangle et avoir une activité de régulation. En pratique, dans les génomes simulés, la plupart des protéines ont effectivement les deux fonctions. Ainsi, perdre un facteur de transcription n'est en général pas neutre, même dans le scénario de suppression du besoin de régulation, car cela implique souvent aussi la perte d'un triangle nécessaire pour réaliser la cible. Simuler l'évolution sur une plus longue durée permettrait peut-être de voir émerger la séparation entre les deux fonctionnalités, avec l'apparition de facteurs de transcription "purs", sans activité métabolique (c'est le cas lorsque la hauteur ou la largeur du triangle codé est nulle). Bien que cela n'ait pas été possible ici en raison des temps de calcul mis en jeu, il aurait probablement été plus pertinent d'attendre l'émergence de cette séparation (si elle se produit effectivement) pour déclencher les scénarios.

Une autre limite importante de ces simulations est le peu de répétitions, seulement 4, et donc la faible puissance statistique. Cependant, ces scénarios montrent que la présence d'un réseau de régulation permet aux organismes de s'adapter à des changements d'environnement tout en conservant des structures génomiques similaires.

Chapitre VI

Discussion

Dans cette première partie du manuscrit, certaines hypothèses proposées dans la littérature pour l'évolution réductive chez *Prochlorococcus* ont été testées avec une méthodologie d'évolution expérimentale *in silico* développée dans le cadre de cette thèse (Batut *et al.*, 2013). Selon cette méthode, des populations de génomes artificiels sont construites par évolution pendant 150 000 générations : elles servent de bases pour les scénarios. Un scénario correspond au changement d'un paramètre de simulation (taille de population, taux de mutation spontané, ...) dans les populations souches puis à l'évolution des populations avec le nouveau paramètre. 11 scénarios ont été testés à partir de 10 populations souches, soit un total de 110 simulations.

Les paramètres des simulations sont différents de ceux couramment utilisées dans *aevol*. Les génomes obtenus après 150 000 générations d'évolution doivent avoir des structures génomiques qualitativement comparables à celles des génomes bactériens, un mode de sélection biologiquement réaliste et un nombre suffisant de gènes pour pouvoir simuler une évolution réductive dans les scénarios. Ainsi, à la génération 150 000, les individus ont des génomes d'environ 9 000 paires de bases, plus de 100 gènes et environ 75% de bases codantes, s'approchant ainsi de la compaction observée dans les génomes bactériens.

Pour favoriser la compaction des génomes et simuler le biais observé dans les génomes bactériens (Mira *et al.*, 2001; Kuo et Ochman, 2009), un biais favorisant les petites délétions par rapport aux petites insertions est présent, dans toutes les simulations, avec des taux spontanés de petites délétions deux fois supérieurs aux taux spontanés de petites insertions. Ce biais ne semble cependant pas avoir un effet important sur les simulations : les tailles des génomes des simulations de contrôle sont relativement stables. De plus, il ne pousse pas tous les scénarios testés vers une évolution réductive, mais peut favoriser la réduction des génomes quand elle intervient. Ce biais se reflète pourtant dans les mutations fixées : 1.95 ± 0.13 fois plus de petites délétions fixées que de petites insertions fixées dans les simulations de contrôles entre les générations 150 000 et 200 000[1]. L'impact du

[1]Ce biais semble être compensé par des duplications plus nombreuses ($P = 4.545 \cdot 10^{-3}$, test des rangs signés de Wilcoxon) et plus longue ($P = 9.766 \cdot 10^{-4}$, test des rangs signés de Wilcoxon) que les grandes

biais vers les petites délétions est donc limité mais sa présence dans les mutations fixées semble refléter les observations faites dans les génomes bactériens (Mira *et al.*, 2001; Kuo et Ochman, 2009).

Prochlorococcus a acquis de nombreux gènes au cours de l'évolution, principalement par transfert horizontal et semble donc avoir toute la machinerie liée à la recombinaison, contrairement aux endosymbiotes. L'arrêt de la recombinaison est l'une des causes évoquées pour l'évolution réductive des endosymbiotes et parfois aussi pour *Prochlorococcus*. Pour pouvoir tester cette hypothèse, nous avons doté les individus des populations simulées de la capacité de recombiner entre eux par transfert de portions homologues entre un individu donneur et un individu receveur. Ainsi, dans les simulations de contrôle, 80% des évènements mutationnels fixés entre les générations 150 000 et 200 000 sont des évènements de recombinaison.

Le choix des paramètres et les populations de génomes ainsi obtenues semblent donc convenir pour tester des scénarios d'évolution réductive. Dans les scénarios étudiés, les changements de paramètres impactent les structures génomiques de façon différente avec des dynamiques différentes. Ces changements semblent répondre à la combinaison de deux pressions de sélection : d'une part, la sélection immédiate qui dépend de la capacité des individus à effectuer la tâche demandée et d'autre part, une sélection indirecte, agissant en second lieu, pour un compromis entre la robustesse mutationnelle et l'évolvabilité des lignées. Ce compromis est représenté plus formellement par $F_\nu W \sim 1$ descendants neutres. En effet, dans tous les scénarios où le changement de paramètre entraîne un changement de $F_\nu W$, la valeur d'environ 1 descendant neutre est rétablie en fin de simulation (Figure IV.12). Les changements initiaux touchent surtout W, c'est-à-dire la structure de la population (distribution des fitness). L'ajustement se fait ensuite par des modifications de F_ν, principalement par la variation de la quantité de bases non codant. $F_\nu W \sim 1$ semble donc être un moteur important de l'évolution des caractéristiques dans *aevol* (Knibbe *et al.*, 2007a), quand les individus sont adaptés à la tâche qu'ils doivent accomplir. Les valeurs optimales de F_ν et W, telles que $F_\nu W \sim 1$, semblent cependant différentes selon les paramètres de simulation. Est-ce un biais du modèle ou existe-t-il une réalité évolutive derrière cette observation ? La présence de cette pression de sélection indirecte est cependant difficile à vérifier dans les populations réelles : il nous faudrait un accès aux données identique à celui dans les expériences d'évolution *in silico* avec la structure de la population, les évènements de reproduction détaillés, etc. De plus, dans *aevol*, un individu peut avoir plusieurs descendants à une génération donnée alors que les bactéries ont au maximum deux descendants. Une génération dans *aevol* correspond donc à plusieurs générations bactériennes. Pouvons-nous vraiment exporter la règle $F_\nu W \sim 1$ comme moteur de l'évolution hors du cadre d'*aevol*, pour des organismes réels ? L'idée d'une pression de sélection indirecte pour un compromis entre la robustesse mutationnelle et l'évolvabilité dirigeant l'évolution en plus de la sélection immédiate est cependant intéressante et devrait être approfondie, au moins avec des modèles théoriques dans un premier temps.

délétions.

L'évolution réductive pour les endosymbiotes est la conséquence du cliquet de Muller, processus affectant principalement les populations non recombinantes où N_e est faible (Felsenstein, 2005; Muller, 1964). Bien que ce processus soit une cause peu probable pour l'évolution réductive chez *Prochlorococcus*, étant donné les fortes tailles efficaces de population (Baumdicker *et al.*, 2012), nous avons testé cette hypothèse en diminuant la taille de population dans un scénario et en supprimant la recombinaison dans un autre. Aucun des deux ne semblent induire une évolution réductive, malgré la présence du biais vers les petites délétions. Il pourrait être intéressant d'étudier la combinaison de ces deux scénarios. Il faudrait aussi déterminer si les souches réduites de *Prochlorococcus* sont toujours capables de faire de la recombinaison. Les évènements de transfert de gènes documentés dans les souches de *Prochlorococcus* HL (souches réduites) semblent être moins nombreux que ceux pour les souches de *Prochlorococcus* LL (Zhaxybayeva *et al.*, 2009).

Comme sous l'effet du cliquet de Muller la sélection naturelle est dépassée par la dérive et est donc moins forte, nous avons aussi testé un scénario où la force de la sélection est réduite. Dans ce scénario, les génomes se réduisent par la perte de gènes et la compaction des génomes. Or, lorsque la pression de sélection est diminuée dans *aevol*, la taille efficace de population augmente et le contraire est aussi vrai. Cependant, les résultats obtenus avec le scénario de diminution de la taille de population et celui d'augmentation de la pression de sélection (diminution de N_e) sont différents, tout comme ceux avec le scénario d'augmentation de la taille de population et celui de diminution de la pression de sélection (diminution de N_e), à la fois en terme de structures génomiques et en terme de dynamiques mutationnelles. Ainsi, changer la taille de population n'est pas équivalent à changer la taille efficace de population, au moins dans les simulations avec *aevol*. Malgré ces différences, seul le scénario de diminution de la pression de sélection entraîne une évolution réductive. Or, d'après Hu et Blanchard (2009), les pressions de sélection seraient plus fortes chez *Prochlorococcus* que chez *Synechococcus*. Cependant, comme mentionné dans la section I.4, les estimations effectuées par Hu et Blanchard (2009) souffrent de quelques faiblesses méthodologiques et sont limitées à la comparaison entre *Prochlorococcus* et *Synechococcus* et pas entre les souches réduites et non réduites. Il faudrait déterminer de façon plus fiable les pressions de sélection et leurs changements potentiels le long de la phylogénie de *Prochlorococcus*.

La plupart des écotypes de souches réduites de *Prochlorococcus* se trouvent dans les eaux de surface, pauvres en nutriments toute l'année alors que les écotypes non réduits sont situés en bas de la colonne d'eau, environnement plus riche en nutriments. Pour l'hypothèse d'adaptation à une nouvelle niche écologique, nous avons testé un scénario de changement d'environnement (déplacement d'un lobe), deux scénarios de changement d'environnement liés à une simplification des tâches à accomplir (suppression et neutralisation d'un lobe), deux scénarios d'arrêt de la variation de l'environnement (en présence et en l'absence d'un réseau de régulation) et un scénario de simplification de la variation de l'environnement en présence d'un réseau de régulation.

Les génomes se réduisent pour 3 de ces 6 scénarios : suppression et neutralisation d'un lobe et arrêt de la variation de l'environnement en l'absence de réseau de régulation. Ce-

pendant, dans ce dernier, les génomes se réduisent seulement par l'augmentation de la compaction des génomes : le nombre de gènes augmente. Ces observations sont différentes de celles pour l'évolution réductive chez *Prochlorococcus*. Ce scénario pourrait cependant correspondre aux changements d'environnement entre *Prochlorococcus* et *Synechococcus*, *Synechococcus* vivant dans les eaux tempérées sujettes à des nombreuses variations annuelles, contrairement aux eaux tropicales et subtropicales où vivent *Prochlorococcus*. Or, d'après Sun et Blanchard (2014), l'évolution réductive aurait démarré à la divergence entre *Prochlorococcus* et *Synechococcus*, avec de nombreuses pertes dans l'ancêtre de *Prochlorococcus*, ce qui serait en contradiction avec les observations du scénario d'arrêt de la variation de l'environnement. Cependant, il pourrait être intéressant de déterminer si la proportion d'ADN non codant des génomes a évolué le long de cette branche, comme c'est le cas dans notre scénario. Dans l'autre scénario d'arrêt de la variation de l'environnement (en présence d'un réseau de régulation), tout comme dans le scénario de simplification de la variation de l'environnement, ce sont les réseaux de régulation présents, et non les structures génomiques, qui absorbent les changements, contrairement aux scénarios en l'absence de réseaux de régulation. Ces observations sont en contradiction avec l'idée qu'avec la stabilisation de l'environnement dans lequel évolue *Prochlorococcus*, des réseaux de régulation sophistiqués deviennent moins indispensables, voire même un fardeau, et sont perdus à moindre coût. Cette contradiction pourrait venir des limites du modèle de régulation utilisé, et notamment du fait qu'il n'impose pas une séparation fonctionnelle nette entre les enzymes et les facteurs de transcription : de nombreuses protéines dans les génomes simulés ont les deux fonctions. Toutefois, il conviendrait également de déterminer plus précisément quels sont les gènes de régulation perdus chez *Prochlorococcus* et dans quels mécanismes de régulation ils interviennent, afin d'identifier pourquoi ces voies de régulation ont été perdues (adaptation, dérive, etc).

Dans les scénarios de suppression et de neutralisation d'un lobe de l'environnement, les génomes se réduisent par la perte des gènes devenus inutiles, car codant pour des triangles remplissant le lobe éliminé. Malgré des dynamiques de changement des structures génomiques différentes, les résultats obtenus en fin de simulation pour ces deux scénarios sont similaires. Ainsi, la proportion de bases codantes et la taille des gènes sont constantes. Chez *Prochlorococcus*, la taille des gènes n'aurait pas changé (Marais *et al.*, 2008; Sun et Blanchard, 2014) alors que la proportion de bases codantes aurait augmenté (Rocap *et al.*, 2003). Cependant, les analyses effectuées sont sommaires (Marais *et al.*, 2008; Sun et Blanchard, 2014; Rocap *et al.*, 2003) et mériteraient d'être approfondies afin de déterminer si ces scénarios sont cohérents avec l'évolution réductive chez *Prochlorococcus*.

Le scénario de déplacement d'un lobe de l'environnement, quant à lui, n'induit pas une évolution réductive malgré des changements des répertoires de gènes (pertes/gains de gènes) et une réduction de la taille des gènes. Le déplacement du lobe a peut-être été trop rapide et un changement plus progressif pourrait permettre qu'au moins une partie des séquences des gènes codant des triangles dans le lobe déplacé s'adapte. Ce scénario pourrait cependant simuler les changements d'environnement liés à l'ascension de *Prochlorococcus* le long de la colonne d'eau (branche ancestrale aux souches de *Prochlorococcus* HL). D'après Kettler *et al.* (2007) et Sun et Blanchard (2014), de nombreux gains de gènes, accompagnés de pertes, auraient eu lieu le long de cette branche, comme observé

dans le scénario. Il serait ainsi intéressant d'avoir une annotation précise des familles de gènes afin de déterminer l'origine des gènes gagnés et perdus dans la branche ancestrale aux souches HL et si les pertes concernent des gènes devenus inutiles et les gains des gènes indispensables dans le nouvel environnement.

Au final, seuls 2 des 6 scénarios liés à des changements d'environnement induisent une évolution réductive qualitativement comparable à celle de *Prochlorococcus* (réduction du génome par perte de gènes et de non codant). En combinant ces scénarios, d'autres cas d'évolution réductive pourraient apparaître et apporter de nouvelles questions pour les analyses des génomes de *Prochlorococus*. Par exemple, l'environnement dans lequel évoluent les souches HL est supposé différent et plus "simple" que celui des souches LL. Nous pourrions ainsi tester un scénario de déplacement d'un lobe, combiné à la suppression/neutralisation d'un autre lobe. Les combinaisons ainsi possibles sont nombreuses.

Plusieurs travaux théoriques de génétique des populations ont montré que l'adaptation aux changements environnementaux peut être améliorée par l'augmentation des taux de mutation (Taddei *et al.*, 1997; Tenaillon *et al.*, 1999). Ainsi, dans le scénario de déplacement d'un lobe, les taux de mutation et de réarrangement fixés sont augmentés (Tableau IV.3). Mais l'augmentation des mutations fixées pourrait aussi provenir d'une augmentation des taux spontanés de mutation, favorisée par la perte des gènes de réparation de l'ADN. Nous avons donc testé un scénario d'augmentation des taux spontanés de mutation locale et un scénario d'augmentation des taux spontané de réarrangement. Dans les deux scénarios, la taille des génomes se réduit par la perte de gènes, la réduction de la taille de ceux-ci et une augmentation de la compaction des génomes. Cependant, les proportions et les dynamiques de changement sont différentes. Ainsi, dans le scénario d'augmentation des taux de mutation, le taux de réarrangement fixé diminue et dans le scénario d'augmentation des taux de réarrangement, le taux de mutation locale fixée diminue. Dans ces deux scénarios, $F_\nu W \sim 1$ descendant neutre est atteint en fin de simulation. Les changements effectués, après l'augmentation des taux de mutation et de réarrangement, sont similaires mais sont plus lents pour le scénario d'augmentation des taux de réarrangement, entraînant ainsi une différence dans la structure de population (moins d'individus reproducteurs et une plus faible proportion de descendants neutres que pour le scénario d'augmentation des taux de mutation). Cependant, dans le scénario d'augmentation des taux de réarrangement, les taux spontanés de réarrangement ont été multipliés par 2 et non par 5 comme dans le scénario d'augmentation des taux de mutation. En effet, la multiplication des taux de réarrangement par 5 entraîne une perte trop importante de l'adaptation de la population à la tâche à accomplir. Pour confirmer la différence entre les scénarios d'augmentation des taux de mutation et de réarrangement, il faudrait tester la multiplication par 3 et par 4 des taux de réarrangement.

L'augmentation des taux de mutation semble être un bon scénario pour l'évolution réductive. Il serait ainsi intéressant de déterminer si, quand et dans quelle proportion les taux de mutation ont augmenté le long de la phylogénie de *Prochlorococcus*. Pour cela, il faudrait étudier l'évolution des séquences, par les valeurs de d_N et d_S par exemple. Ces données se concentrent cependant sur les mutations ponctuelles. Il faudrait aussi étudier

les insertions et délétions, ainsi que les réarrangements. De plus, la principale différence de structure génomique entre le scénario d'augmentation des taux de mutation et celui d'augmentation des taux de réarrangement est la proportion de bases codantes, inchangée dans le second scénario mais augmentée dans le premier scénario, comme cela semble être le cas pour l'évolution réductive chez *Prochlorococcus* (Rocap *et al.*, 2003).

D'après les estimations d'Osburne *et al.* (2011), les taux de mutation actuels chez *Prochlorococcus* seraient équivalents à ceux des autres bactéries. L'augmentation serait ainsi transitoire. Se pose alors la question de la conservation de l'évolution réductive observée dans le scénario d'augmentation des taux de mutation si ceux-ci reprenaient leur valeur initiale. Un tel scénario serait aisément testable avec *aevol*. De plus, d'après la théorie sur les mutateurs (Taddei *et al.*, 1997; Tenaillon *et al.*, 1999), l'augmentation des taux de mutation serait due à un changement de niche. Il faudrait ainsi tester une combinaison de l'augmentation des taux de mutation et du changement d'environnement, avec plusieurs possibilités (déplacement d'un lobe seul, déplacement d'un lobe et stabilisation de l'environnement, déplacement d'un lobe et suppression d'un autre lobe et aussi les scénarios en présence d'un réseau de régulation), afin de couvrir les différentes phases possibles de l'évolution des génomes chez *Prochlorococcus*.

Avec ces 11 scénarios, nous avons testé plusieurs facettes des différentes hypothèses proposées pour l'évolution réductive, en particulier l'évolution réductive chez *Prochlorococcus* (cliquet de Muller, adaptation à une nouvelle niche écologique et forts taux de mutation). Une pression de sélection indirecte pour un compromis entre la robustesse mutationnelle et l'évolvabilité semble avoir un impact fort dans l'évolution des structures génomiques des scénarios avec *aevol*. Cependant, ces scénarios simples, seuls, ne permettent pas d'apporter des réponses claires sur l'évolution réductive chez *Prochlorococcus*. En effet, ils mériteraient d'être approfondis et combinés afin de répondre à certaines questions. De plus, des analyses plus poussées des génomes de *Prochlorococcus* sont indispensables dans un premier temps, pour mieux comprendre les mécanismes derrière l'évolution réductive chez *Prochlorococcus*. Ainsi, dans la seconde partie du manuscrit, nous présentons des analyses qui pourraient apporter des réponses : la proportion de bases non codantes, la recombinaison, l'évolution des contenus en gènes et de la longueur des gènes, l'évolution des séquences et des pressions de sélection.

Deuxième partie

Analyses de l'évolution réductive chez *Prochlorococcus*

Chapitre VII

Architecture des génomes et évolution réductive

Dans ce chapitre, nous présentons quelques caractéristiques génomiques de l'évolution réductive chez *Prochlorococcus*. Ces caractéristiques ne font pas l'objet d'études très approfondies mais sont utiles pour discriminer les différents scénarios d'évolution réductive présentés dans la première partie du manuscrit. Nous étudions ainsi l'évolution de caractéristiques, comme la proportion d'ADN non codant ou les structures opéroniques, en prenant en compte la phylogénie sous-jacente afin d'identifier quand les changements ont pu avoir eu lieu le long de la phylogénie. Nous utilisons pour cela la méthode des contrastes phylogénétiquement indépendants (Felsenstein, 1985), qui fournit pour chaque nœud d'un arbre phylogénétique le contraste, pour un caractère, entre les branches filles du nœud, en prenant en compte les longueurs des branches, c'est-à-dire les distances évolutives entre les nœuds fils (Figure VII.1). Nous cherchons en particulier à savoir si la proportion de bases non codantes a évolué avec la réduction des génomes, s'il en est de même des distances intergéniques et des opérons et si les souches réduites ont toujours la capacité de recombiner.

VII.1 Évolution de la proportion de bases non codantes et des distances intergéniques

Pour estimer la proportion de bases non codantes, nous utilisons tous les gènes des souches d'intérêt (Tableau C.1, en annexe) avec leurs positions obtenues à partir de la base de données du NCBI. Les gènes sont ensuite ordonnés par position croissante le long du génome. La quantité de bases entre les gènes est alors comptabilisée puis rapportée à la taille du génome. Pour les gènes chevauchants, la quantité de bases entre les gènes est nulle et comptabilisée comme telle. Nous nous intéressons aussi aux médianes des distances intergéniques.

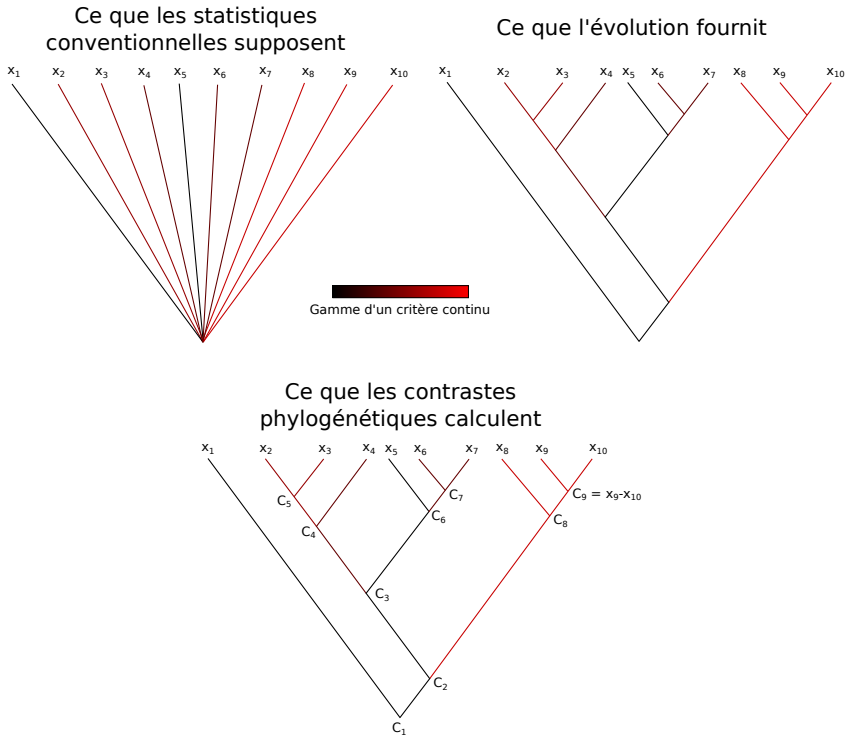


Figure VII.1 – Principe des contrastes phylogénétiquement indépendants
Les méthodes statistiques conventionnelles appliquées aux données de génomique comparative supposent que toutes les espèces sont totalement indépendantes. Les relations entre les espèces ressemblent alors à une étoile (Schéma en haut, à gauche). Or l'évolution se fait de façon hiérarchique depuis des ancêtres communs (Schéma en haut, à droite). Les méthodes de contrastes phylogénétiquement indépendants prennent en compte cette phylogénie en comparant les différences pour un caractère entre les branches filles d'un nœud (Schéma en bas, au centre). La logique de ces méthodes est d'utiliser l'information phylogénétique (et un modèle brownien d'évolution du trait) pour transformer les données brutes en données statistiquement indépendantes et identiquement distribuées.
La figure est inspirée de la figure 3 de Garland et Carter (1994).

Les contrastes phylogénétiques sur la proportion de bases non codantes et les médianes des distances intergéniques sont fortement corrélés positivement (Figure VII.2). Cela signifie que si pour deux branches filles d'un nœud, nous observons des différences de la proportion de bases non codantes, des différences similaires seront observées pour les médianes des distances intergéniques. Ainsi, ces deux caractéristiques génomiques semblent évoluer de façon similaire. Nous nous concentrons ainsi seulement sur la proportion de bases non codantes.

La taille des génomes et la proportion de bases non codantes ont évolué en opposition : les contrastes de ces caractéristiques sont corrélés négativement (Figure VII.3). Cependant, cette corrélation est peu soutenue statistiquement et est principalement due au nœud différenciant les deux souches de *Prochlorococcus* LLIV (nœud 4). Lorsque ce nœud est éliminé, la corrélation est positive et soutenue statistiquement (Figure VII.3). Les forts contrastes observés pour le nœud 4 ainsi que les faibles longueurs des branches filles à ce nœud signifie que des changements ont eu lieu dans un laps de temps court. Ainsi, MIT9303 a un

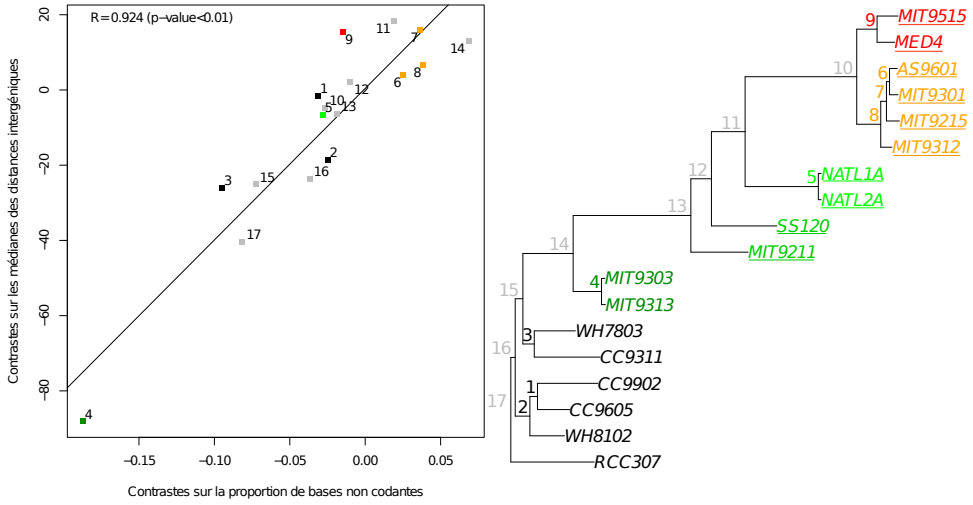


Figure VII.2 – Contrastes sur les médianes des distances intergéniques en fonction des contrastes sur la proportion de bases non codantes
Chaque point représente un nœud de l'arbre phylogénétique à droite. La droite en noir est la droite de régression linéaire. La valeur de corrélation est indiquée avec sa p-value.
Les couleurs des noms des souches symbolisent les différentes clades avec en noir les souches *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit. L'arbre phylogénétique est construit comme décrit dans la section C.4 (Annexe).

génome plus grand que MIT9313 mais une proportion plus faible de bases non codantes. Comment expliquer ces différences ? Dans la littérature, les souches MIT9313 et MIT9303 sont souvent étudiées ensemble, sous le nom des souches non réduites, sans distinction l'un de l'autre, étant donné les courtes branches les séparant. Cependant, ces souches, en particulier MIT9303, semblent évoluer dans une direction opposée à celle de l'évolution réductive, depuis leur différenciation avec les autres souches de *Prochlorococcus*. Il faut ainsi garder en tête pour les prochaines analyses le cas de MIT9303 et MIT9313, les souches de *Prochlorococcus* dites non réduites.

Sans prendre en compte le cas particulier du nœud 4, lorsque les génomes réduisent en taille, la proportion de bases non codantes diminue. Ainsi, les génomes réduits sont plus compacts avec environ 10% de bases non codantes (Figure VII.4a) et des médianes des distances intergéniques de l'ordre de 40 bases (Figure VII.4b). Notons que cette dernière valeur est cependant bien supérieure à la médiane des distances intergéniques de 3 bases observée chez *Pelagibacter ubique* (Giovannoni *et al.*, 2005). L'évolution réductive semble ainsi pousser à la réduction globale des génomes, en éliminant des gènes mais aussi de l'ADN intergénique.

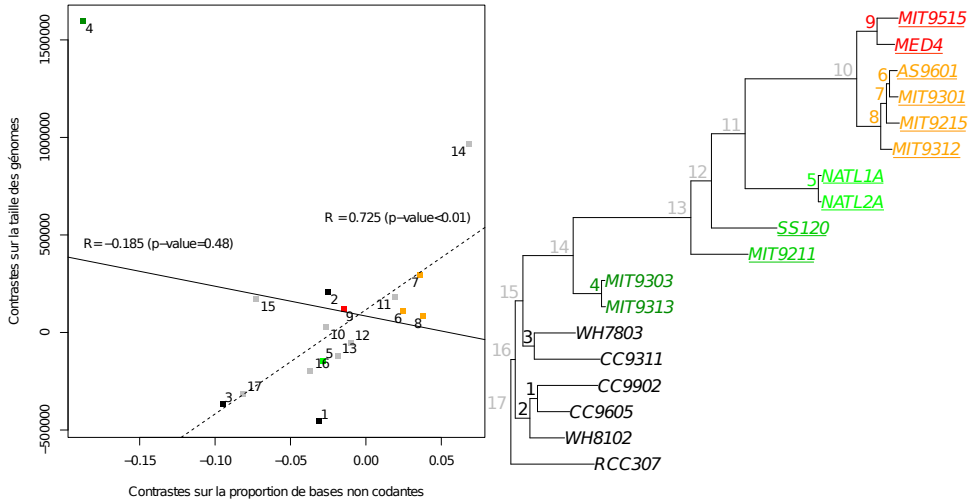
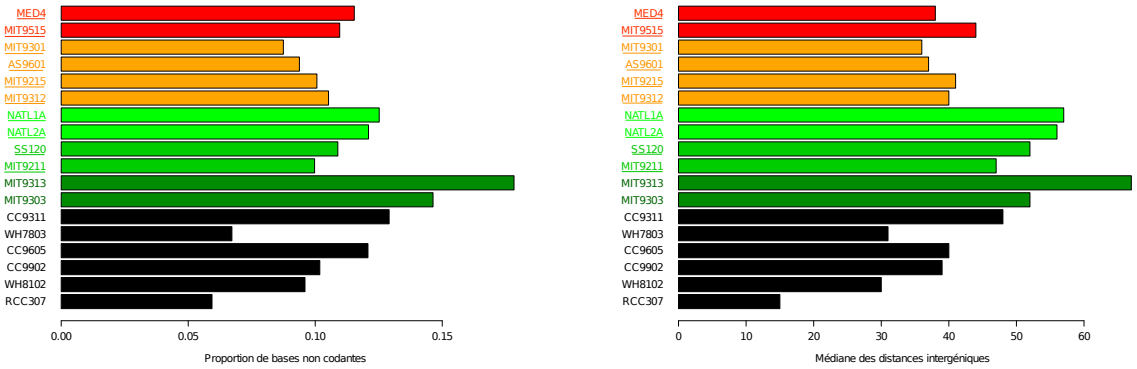


Figure VII.3 – Contrastes sur la taille des génomes en fonction des contrastes sur la proportion de bases non codantes
Chaque point représente un nœud de l'arbre phylogénétique à droite. La droite en trait plein est la droite de régression linéaire lorsque tous les points sont pris en compte. La droite en pointillés est celle obtenue en supprimant le nœud 4. La valeur de corrélation est indiquée avec sa p-value.
Les couleurs des noms des souches symbolisent les différentes clades avec en noir les souches *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit. L'arbre phylogénétique est construit comme décrit dans la section C.4 (Annexe).



(a) Proportion de bases non codantes **(b)** Médianes des distances intergéniques

Figure VII.4 – Proportion de bases non codantes et médianes des distances intergéniques pour les différentes souches de *Prochlorococcus* et de *Synechococcus*
Les couleurs des noms des souches symbolisent les différentes clades avec en noir les souches *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit.

VII.2 Évolution des structures opéroniques

Chez les bactéries, les opérons sont des portions d'ADN regroupant des gènes transcrits dans un même ARN messager. Les gènes au sein d'un même opéron, au moins au nombre de deux, sont souvent liés en terme de fonction et sont régulés de la même façon car ils sont sous le contrôle des mêmes régulateurs.

Chez *Prochlorococcus* MED4, les gènes situés au sein d'opérons évoluent plus lentement que les autres (Wang *et al.*, 2014). La proportion de gènes au sein d'opérons est plus forte pour le génome cœur que pour le génome flexible (Wang *et al.*, 2014), en accord avec des études montrant un enrichissement en gènes essentiels au sein des opérons (Price *et al.*, 2006, 2005). L'évolution des opérons serait ainsi un processus adaptatif en cours pour la co-régulation de gènes : les gènes au sein d'opérons relativement anciens seraient fortement corégulés (Memon *et al.*, 2013). Disposer de la carte opéronique pour *Prochlorococcus* et *Synechococcus* permettrait de situer les changements des gènes au sein des opérons dans un contexte adaptatif de co-régulation des gènes.

Wang *et al.* (2014) ont fourni la première carte opéronique basée sur des données expérimentales chez *Prochlorococcus* MED4. Ces données ne sont cependant disponibles que pour la souche *Prochlorococcus* MED4 alors qu'il serait intéressant de pouvoir comparer avec des souches de *Prochlorococcus* LL et de *Synechococcus*. Ainsi, Memon *et al.* (2013) ont prédit les opérons pour 41 cyanobactéries dont les souches qui nous intéressent. Leur méthode de prédiction des opérons repose sur la conservation de l'ordre des gènes et des espacements intergéniques.

Les données obtenues par prédiction pour *Prochlorococcus* MED4 sont un peu différentes de celles issues des données expérimentales (Tableau VII.1). Ainsi, bien que le pourcentage des bases du génome présentes dans un opéron soit significativement supérieur dans les données de Memon *et al.* (2013) ($P < 2.2 \cdot 10^{-16}$, test de χ^2 à un degré de liberté), le pourcentage de gènes dans un opéron est inférieur ($P = 3.527 \cdot 10^{-4}$, test de χ^2 à un degré de liberté). Même si les données de Wang *et al.* (2014) sont plus fiables, seules les données de Memon *et al.* (2013) seront utilisées par la suite car tous les génomes étudiés pourront être comparés par le biais de données opéroniques issus d'une même méthode.

De nombreux indicateurs des structures opéroniques sont accessibles avec ces données, comme le nombre d'opérons, la taille moyenne des opérons, le nombre moyen de gènes dans un opéron,... Afin de trouver les indicateurs les plus pertinents et faire ressortir des motifs intéressants, nous avons étudié la corrélation entre les contrastes phylogénétiques de ces indicateurs (Figure VII.5). Ainsi, le nombre d'opérons évolue parallèlement aux tailles des génomes. Lorsque la taille du génome se réduit, le nombre d'opérons diminue mais dans une proportion moindre, entrainant une augmentation de la proportion du génome couvert par des opérons par une diminution de la distance moyenne entre les opérons (Figure VII.5). A part pour les souches de *Prochlorococcus* LLIV (nœud 4), les changements de proportion de couverture du génome par les opérons s'accompagnent de changements du même ordre dans la proportion de gènes dans des opérons. Ainsi,

Caractéristiques	Wang *et al.* (2014)	Memon *et al.* (2013)
Nombre d'opérons	422	314
Taille moyenne des opérons (bp)	2094	3266
Pourcentage du génome couvert par les opérons	53	59
Nombre moyen de gènes par opéron	2.70	3.54
Nombre maximal de gènes dans un opéron	20	30
Pourcentage du gènes dans un opéron	66	65
Longueurs moyenne des séquences entre opérons (bp)	1839	2020
Nombre d'opérons identiques	142	

Table VII.1 – Comparaison des caractéristiques des opérons trouvés chez *Prochlorococcus* MED4 entre la méthodologie expérimentale de Wang *et al.* (2014) et la méthodologie de prédiction de Memon *et al.* (2013)

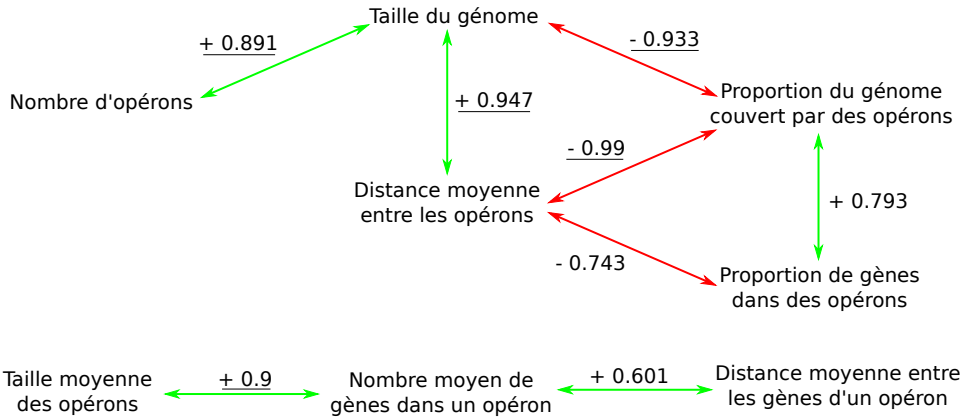


Figure VII.5 – Relation de corrélation entre les contrastes phylogénétiques de différents indicateurs des structures opéroniques
Seules les corrélations significatives d'après le test de corrélation de Spearman sont montrées dans ce schéma. Les valeurs soulignées correspondent aux cas où les corrélations sont significatives en prenant en compte tous les nœuds, les valeurs non soulignées lorsque la corrélation est significative avec l'élimination du nœud 4.
Les flèches rouges symbolisent les corrélations négatives et les flèches vertes les corrélations positives.

la réduction du génome entraîne une augmentation de la corégulation de gènes au sein d'opérons, via l'augmentation de la couverture du génome par les opérons.

VII.3 Recombinaison

De nombreux gènes ont été acquis le long de l'évolution de *Prochlorococcus*, principalement par transferts horizontaux. Ces gènes provenant d'autres organismes reflètent la présence d'échanges de matériel génétique entre espèces, dans des proportions différentes selon les souches. Les acquisitions résultantes de gènes sont facilitées par des éléments génétiques mobiles et des phages, au sein d'îlots génomiques. Les îlots génomiques sont des larges régions de plus de 8 kb contenant des gènes non conservés. Chez *Prochlorococcus* et *Synechococcus*, de nombreux îlots ont été caractérisés (Coleman *et al.*, 2006; Dufresne

et al., 2008; Avrani *et al.*, 2011), où les gènes non conservés sont surreprésentés (Kettler *et al.*, 2007; Luo *et al.*, 2011). Ils seraient ainsi des hotspots pour les transferts de gènes (Dufresne *et al.*, 2008; Luo *et al.*, 2011; Coleman *et al.*, 2006), mécanisme clé pour l'acquisition de nouvelles fonctions via le partage de gènes entre les souches. Ainsi, certaines familles auraient été transférées entre les souches HL et LL. Les échanges seraient fréquents entre *Synechococcus* et les souches LL (Zhaxybayeva *et al.*, 2009), potentiellement via des cyanophages (Zeidner *et al.*, 2005; Sullivan *et al.*, 2005, 2003; Lindell *et al.*, 2004).

Les évènements de transferts reflètent la présence de recombinaison au sein des souches de *Prochlorococcus*, contrairement aux endosymbiotes où le nombre de transferts de gènes recensés est faible (van Ham *et al.*, 2003; Toft et Andersson, 2010; Tamas *et al.*, 2002). Le cliquet de Muller, processus dégénératif affectant les petites populations non recombinantes (Muller, 1964; Felsenstein, 2005), semble donc ne pas pouvoir s'appliquer au cas de *Prochlorococcus*. Les évènements de transferts recensés concernent des acquisitions de nouveaux gènes entiers. Qu'en est-il des échanges d'allèles ? Ceux-ci permettent de favoriser la propagation des allèles avantageux et de limiter celle d'allèles délétères, en cassant des liens entre sites voisins et en augmentant ainsi l'efficacité de la sélection.

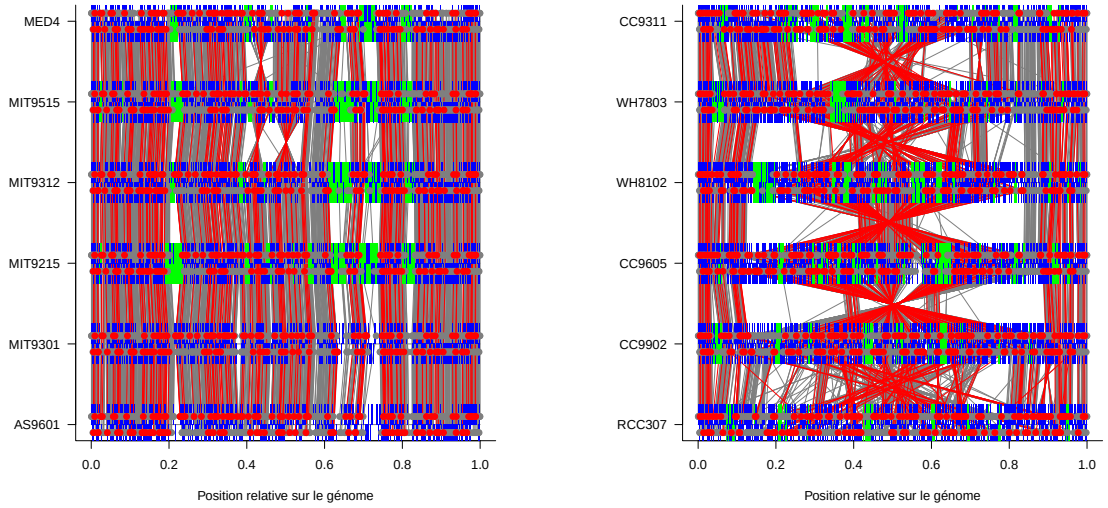
Nous avons estimé la recombinaison intragénique à l'aide du logiciel *PHI* (Bruen *et al.*, 2006), qui permet de détecter la présence de recombinaison dans des alignements multiples à l'échelle des gènes en identifiant des sites dont l'histoire est incompatible avec l'histoire de la famille de gènes. Cette méthode est l'une des plus robustes aux variations des taux de recombinaison, aux divergences des séquences et de dynamiques de population (Bruen *et al.*, 2006). L'objectif est de déterminer si une famille de gènes homologues a subi des évènements de conversion de gènes parmi les membres des taxons d'intérêt. Pour éviter un effet possible du nombre de gènes dans l'alignement sur l'estimation de la recombinaison, l'étude se concentre sur des familles de gènes homologues sans paralogues.

Nous souhaitons déterminer la présence et les caractéristiques de la recombinaison intragénique au sein des souches réduites de *Prochlorococcus*, par rapport aux souches non réduites. Les souches LL ne possèdent pas un contenu en bases GC homogène entre les souches, ce qui risque de biaiser la détection de recombinaison en la surestimant. De plus, pour éviter des estimations de recombinaisons biaisées par le nombre de séquences, il faudrait pouvoir comparer un même nombre de souches. C'est pourquoi nous avons utilisé les souches de *Prochlorococcus* HL et les souches de *Synechococcus*, ces deux groupes étant composés de six souches (Figure C.2a, en annexe). Parmi les familles de gènes homologues utilisées, présentées dans le tableau C.3 (en annexe) et alignées par codons avec *Prank* (Löytynoja et Goldman, 2005), 798 sont communes aux deux groupes, 287 présentes seulement pour les souches de *Synechococcus* et 335 seulement pour les souches de *Prochlorococcus* HL.

Les familles sont considérées comme recombinantes lorsque la p-value du test effectué par *PHI* (Bruen *et al.*, 2006) est significative avec 5% d'erreur (Tableau VII.2). Sur les 798 familles communes aux deux groupes, 2.8% sont détectées comme recombinantes à la fois pour *Synechococcus* et *Prochlorococcus*, 11.8% recombinantes pour *Prochlorococcus* et non pour *Synechococcus*, 11.7% recombinantes seulement pour *Synechococcus* et non

		Prochlorococcus HL				
		Non recomb.	Recomb.	Pas de signal	Pas présentes	Total
Syn.	Non recomb.	589	92	0	246	927
	Recomb.	96	21	0	41	158
	Pas de signal	0	0	0	0	0
	Pas présentes	308	27	0	-	335
	Total	993	140	0	287	-

Table VII.2 – Répartition des familles homologues des souches de *Synechococcus* et de *Prochlorococcus* HL selon leur état de recombinaison d'après les résultats de PHI (Bruen *et al.*, 2006)



(a) *Prochlorococcus* HL **(b)** *Synechococcus*

Figure VII.6 – Position relative des familles de gènes le long des génomes de *Prochlorococcus* HL et de *Synechococcus*
Les points en rouge correspondent aux familles recombinantes et les points en gris aux familles non recombinantes. Les lignes représentent les liens entre les familles homologues des différents génomes. Les rectangles verts correspondent aux ilôts génomiques issus de Avrani *et al.* (2011) et les rectangles bleus aux opérons inférés par Memon *et al.* (2013).

pour *Prochlorococcus* (Tableau VII.2). Pour les familles non communes, le pourcentage de familles recombinantes est de 14.6% pour les familles présentes seulement chez *Synechococcus* et 8.1% pour *Prochlorococcus* HL (Tableau VII.2). Les familles plus récentes présentes seulement chez *Prochlorococcus* HL semblent ainsi moins sujettes à la recombinaison intragénique que les familles plus anciennes présentes seulement chez *Synechococcus*.

Globalement, les familles de gènes chez *Synechococcus* sont plus recombinantes (14.6%) que chez *Prochlorocccus* HL (12.5%), mais cette différence n'est pas significative ($P = 0.2471$, test de χ^2 à un degré de liberté), même rapportée au nombre moyen de gènes dans les groupes de souches ($P = 0.09595$, test de χ^2 à un degré de liberté). Les souches réduites de *Prochlorococcus*, comme celles de *Prochlorococcus* HL semblent ainsi capables de recombiner, ce qui accrédite l'idée que le cliquet de Muller est une explication peu probable pour l'évolution réductive chez *Prochlorococcus*.

Au sein des souches de *Prochlorococcus* HL (Figure VII.6a), les positions des familles de gènes homologues sont relativement bien conservées malgré une inversion déjà recensée (Coleman *et al.*, 2006) autour du terminus de réplication entre les souches HLI (MED4, MIT9515) et HLII (MIT9301, AS9601, MIT9215, MIT9312). La structure des familles de gènes le long des génomes est moins conservée pour *Synechococcus* (Figure VII.6b), avec de nombreuses inversions autour du terminus de réplication. Les familles de gènes, qu'elles soient recombinantes ou non, sont réparties tout le long du génome de telle façon que les familles recombinantes et les familles non recombinantes sont à des distances similaires de l'origine de réplication (tests de Mann-Whitney pour la comparaison des distances relatives à l'origine entre familles recombinantes et non recombinantes, non significatifs avec une p-value de 5%).

Cette répartition quasi-homogène des familles de gènes est interrompue par des portions où les familles sont peu présentes : les îlots génomiques (Figure VII.6). En effet, alors que la surface totale des îlots représente 22% du génome chez MIT9215, seulement 2.9% de familles non recombinantes et 0.7% de familles recombinantes sont présentes dans les îlots. Cette observation est faite pour tous les génomes *Prochlorococcus* HL et *Synechococcus*. En effet, les îlots génomiques sont des portions des génomes où les gènes non conservés sont surreprésentés. Or, les familles étudiées ici sont des familles de gènes conservées dans au moins 6 souches, expliquant leur relative absence dans les îlots. Cependant, les familles recombinantes et les familles non recombinantes ne semblent pas égales face à leur présence au sein des îlots mais aussi des opérons pour les souches HL. Ainsi, dans ces dernières, les familles recombinantes sont moins présentes au sein des îlots génomiques mais plus présentes au sein des opérons que les familles non recombinantes ($P = 0.01429$ et $P = 0.002386$, tests de Mann-Whitney sur la proportion de familles recombinantes et non recombinantes au sein des îlots génomiques et au sein d'opérons, respectivement). Cette différence entre *Prochlorococcus* HL et *Synechococcus* vient du fait que les familles non recombinantes sont moins présentes au sein des opérons chez *Prochlorococcus* HL que chez *Synechococcus* ($P = 0.0010802$, test de Mann-Whitney sur la proportion de familles non recombinantes au sein d'opérons chez *Prochlorococcus* HL et *Synechococcus*), sans changement de proportion de familles recombinantes au sein des opérons entre *Prochlorococcus* HL et *Synechococcus*.

Synechococcus et *Prochlorococcus* HL semblent donc capables de recombinaison allélique, même entre les souches réduites. Cette recombinaison allélique ne semble pas limitée à certaines portions des génomes : les familles recombinantes sont réparties tout le long du génome, contrairement aux îlots génomiques où les principaux transferts de gènes ont lieu.

⋆ ⋆ ⋆

Les génomes des souches réduites de *Prochlorococcus* présentent ainsi une proportion d'ADN codant supérieure, des distances intergéniques réduites et une couverture par les opérons supérieure par rapport aux génomes des souches non réduites. L'évolution réductive au sein des souches de *Prochlorococcus* est donc accompagnée d'une compaction des génomes. Ce point différencie l'évolution réductive de *Prochlorococcus* de celle des endosymbiotes, mais la principale différence entre ces deux évolutions réductives est la

présence de recombinaison intragénique pour les souches de *Prochlorococcus* réduites, rejetant ainsi l'hypothèse de la perte de recombinaison comme mécanisme fondateur de l'évolution réductive.

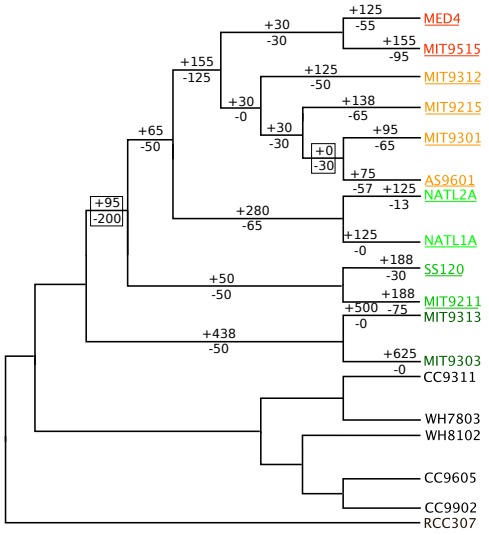
Chapitre VIII

Évolution des contenus en gènes

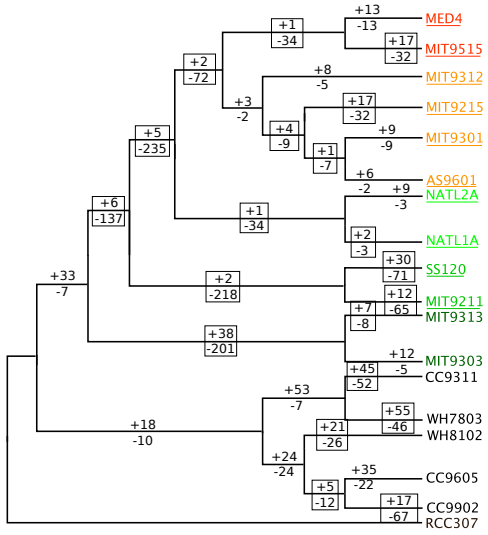
Les changements de répertoires de gènes peuvent donner des clés pour reconstruire l'histoire de l'évolution réductive. En effet, lors de l'évolution réductive, malgré des pertes de bases non codantes, les génomes se sont réduits par la perte de gènes. La principale hypothèse proposée dans la littérature est une hypothèse adaptative qui suppose que les changements du répertoire de gènes, principalement dus à la perte de gènes non essentiels, découlent de l'adaptation à un nouvel environnement. Cependant, toutes les souches réduites de *Prochlorococcus* ne se trouvent pas dans un environnement aussi différent des souches non réduites que les souches de *Prochlorococcus* HL. Ainsi, les souches LLII/LIII aux génomes réduits et les souches non réduites LLIV se situent en bas de la colonne d'eau. Quand ont eu lieu les pertes de gènes ? Est-ce que ce fut progressif ? Y a-t-il eu des évènements de compensation de pertes par des gains ? Pour répondre à ces questions, les gains et les pertes de gènes sont comparés dans un contexte phylogénétique par la construction d'un arbre de gains et pertes de gènes.

Trois études du répertoire génique au cours de l'histoire de l'évolution réductive de *Prochlorococcus* (Kettler *et al.*, 2007; Luo *et al.*, 2011; Sun et Blanchard, 2014) ont émis des conclusions différentes (Figure VIII.1). Dans ces études, la reconstruction des gains et des pertes a été effectuée par maximum de parcimonie. Afin de mieux comprendre l'évolution du répertoire génique, nous utilisons une modélisation explicite des processus d'évolution des familles de gènes basée sur une approche de vraisemblance (Hahn *et al.*, 2005; Csűrös et Miklós, 2006; Cohen et Pupko, 2010).

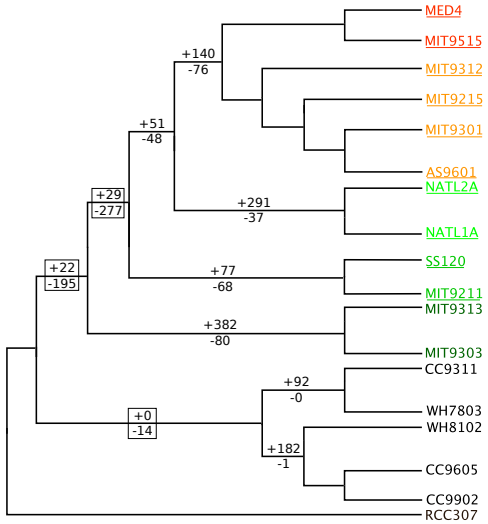
Avec cette méthode plus rigoureuse statistiquement, nous pouvons effectuer une nouvelle analyse plus complète. L'objectif principal est de déterminer quand ont eu lieu les pertes et les gains de gènes le long de la phylogénie de *Prochlorococcus* et si l'évolution réductive a été progressive (Luo *et al.*, 2011) ou due à un évènement temporaire (Sun et Blanchard, 2014). Nous utilisons ici le modèle phylogénétique de naissance et mort des gènes (Csűrös et Miklós, 2006), implémenté dans le logiciel *Count* (Csűrös, 2010). Celui-ci prend en compte les transferts, les duplications et les pertes de gènes. Il est basé sur un processus stochastique déterminant l'évolution de la taille des familles pour chaque branche avec



(a) Kettler *et al.* (2007)



(b) Luo *et al.* (2011)



(c) Sun et Blanchard (2014)

Figure VIII.1 – Nombre de familles de gènes gagnées et perdues le long de l'arbre phylogénétique de *Prochlorococcus* et *Synechococcus* pour trois analyses des gains et pertes de gènes présentes dans la littérature (Luo *et al.*, 2011; Kettler *et al.*, 2007; Sun et Blanchard, 2014)
Pour les arbres de Luo *et al.* (2011) et Sun et Blanchard (2014), les valeurs correspondent exactement à celles fournies dans les articles. Pour l'arbre de Kettler *et al.* (2007), comme les valeurs brutes ne sont pas fournies, des approximations faites sur la base de la figure 3 de Kettler *et al.* (2007) ont été reportées sur l'arbre.
Les conclusions de ces trois analyses sont différentes car les branches où les pertes sont les plus importantes changent selon les analyses, les rectangles symbolisant les cas où le nombre de pertes est supérieur au nombre de gains.
Les couleurs des noms des souches symbolisent les différentes clades avec en noir les souches *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit. L'arbre phylogénétique est construit comme décrit dans la section C.4 (Annexe) mais les longueurs des branches sont arbitraires et ne reflètent pas les taux d'évolution.

		Souche	CDS	Familles	CDS liés à une famille	CDS orphelins
Prochlorococcus	HLI	MED4	1 717	1 523	1 636	49
		MIT9515	1 905	1 610	1 720	154
	HLII	MIT9312	1 810	1 578	1 693	82
		MIT9215	1 982	1 687	1 806	143
		MIT9301	1 906	1 672	1 775	198
		AS9601	1 920	1 680	1 787	100
	LLI	NATL1A	2 193	1 804	1 971	189
		NATL2A	2 162	1 797	1 965	164
	LLII/LLIII	SS120	1 883	1 562	1 673	182
		MIT9211	1 853	1 513	1 616	209
	LLIV	MIT9313	2 269	1 847	2 024	207
		MIT9303	2 997	1 959	2 154	804
Synechococcus		CC9311	2 832	2 026	2 236	615
		WH7803	2 533	2 057	2 220	273
		WH8102	2 519	1 978	2 171	306
		CC9605	2 645	2 038	2 240	368
		CC9902	2 306	1 959	2 110	159
		RCC307	2 231	1 843	2 000	498

Table VIII.1 – Nombre de CDS, familles de gènes, CDS dans une famille de gènes et CDS orphelins (sans familles de gènes, CDS uniques à la souche) dans les souches étudiées
Le nombre de CDS correspond aux données du Tableau C.2 (Annexe). Le nombre de familles de gènes est le nombre de famille Hogenom différentes correspondant aux CDS des souches.

trois paramètres : le taux de perte, le taux de duplication et le taux de gains (transfert ou innovation). La reconstruction des états ancestraux se fait par calcul des probabilités *a posteriori* des tailles des familles aux nœuds internes de l'arbre. Contrairement aux autres méthodes de vraisemblance, l'information associée aux nœuds n'est pas seulement la présence ou l'absence d'une famille, mais aussi si une famille a plusieurs représentants (paralogues).

L'arbre de gains et de pertes pour *Prochlorococcus* est reconstruit en incluant les 6 souches de *Synechococcus* et en prenant comme arbre guide l'arbre inféré selon la méthode décrite dans la section C.4 (en annexe), avec les familles de gènes présentes dans toutes les souches de *Prochlorococcus* et de *Synechococcus* (Figure C.2a, en annexe). Nous considérons ensuite les gains et les pertes le long de cet arbre guide pour les familles *Hogenom* présentes au sein d'au moins un génome étudié (Tableau VIII.1), soit 3 778 familles. Pour l'inférence probabiliste des taux de gains, de pertes et de duplications de gènes le long de l'arbre phylogénétique, nous utilisons le modèle phylogénétique de naissance-mort (Csűrös et Miklós, 2006) avec des taux différents pour chacune des branches. Lors de la reconstruction des états ancestraux à l'aide des probabilités *a posteriori*, nous considérons qu'une famille est présente sur un nœud si une des trois conditions suivantes est vérifiée (Figure VIII.2) :

- Elle est présente dans les deux nœuds fils et les probabilités de gains le long des branches filles sont inférieures à 0.8 ;
- Elle est présente dans au moins un des nœuds fils et la probabilité de perte dans la

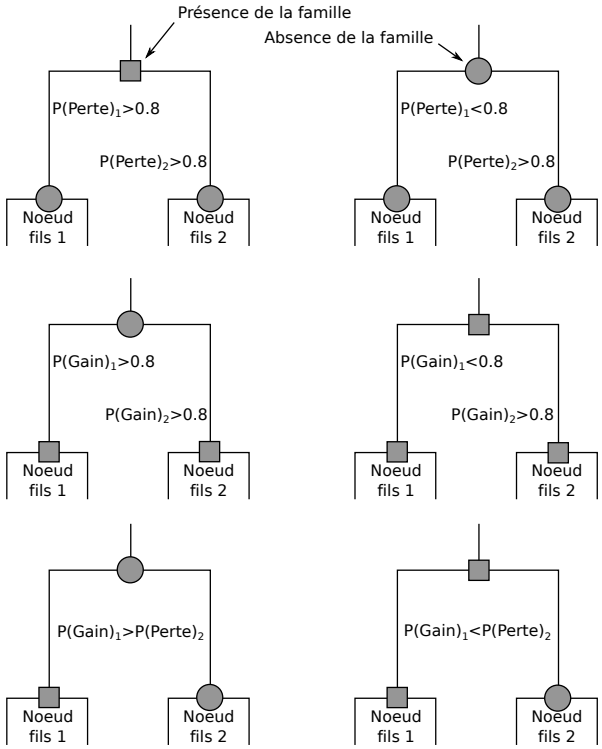


Figure VIII.2 – Principe d'inférence de la présence et l'absence d'une famille de gènes à un nœud lors de la reconstruction des états ancestraux à l'aide des probabilités *a posteriori* inférées avec Count La présence d'une famille à un noeud est symbolisée par un carré et l'absence par un rond. Elles dépendent de la présence ou l'absence de cette famille aux nœuds fils et des probabilités de gains et pertes de la famille dans les branches conduisant aux nœuds fils.

branche fille conduisant au nœud sans la famille est supérieure à la probabilité de gain dans la branche avec la famille ;

- Elle est absente dans les deux nœuds fils et les probabilités de perte dans les deux branches filles sont supérieures à 0.8.

Une famille peut être présente en plusieurs exemplaires sur un nœud. Cette information est récupérable comme la présence et l'absence pour chaque nœuds de l'arbre phylogénétique mais le nombre exact de paralogues n'est pas disponible. Nous nous concentrons ainsi seulement sur la présence et l'absence des familles de long de la phylogénie. *Count* ne prenant pas en compte les gènes présents seulement dans un génome (regroupés dans une famille *Hogenom* de gènes dits orphelins), ceux-ci sont ensuite ajoutés aux branches terminales de l'arbre comme des gains. La souche *Synechococcus* RCC307 est utilisée comme groupe externe afin d'estimer les familles présentes au sein de l'ancêtre commun aux autres souches. Les évènements ayant lieu dans la branche différenciant RCC307 des autres souches ne sont donc pas étudiés.

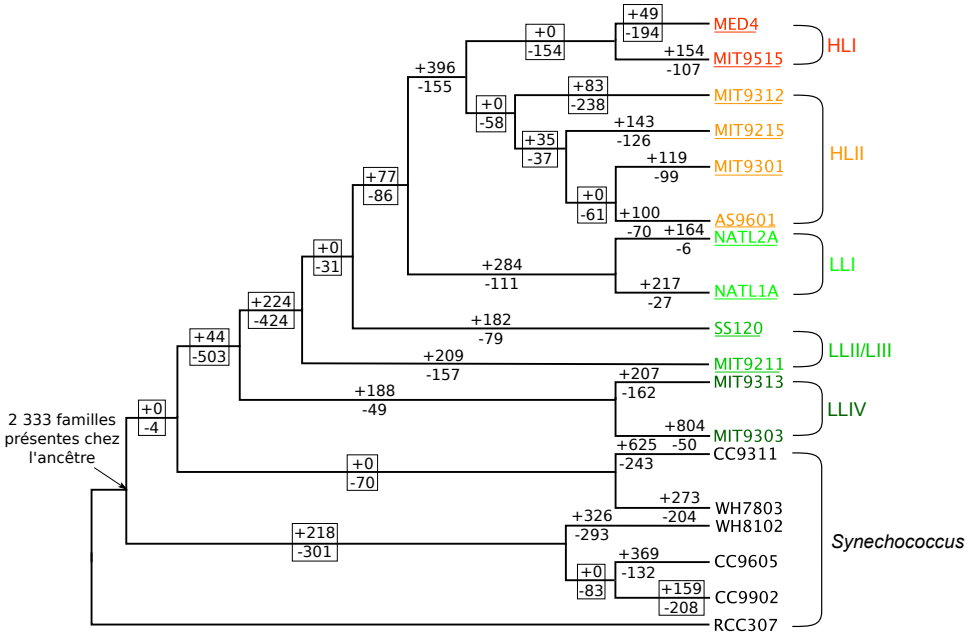


Figure VIII.3 – Nombre de familles de gènes gagnées et perdues le long de l'arbre phylogénétique de *Prochlorococcus* et *Synechococcus* depuis les 2 333 familles de l'ancêtre commun
La reconstruction des nombres de pertes et gains pour chaque branche se fait en utilisant les probabilités *a posteriori* de présence et absence des 3 778 familles de gènes présentes dans au moins une souche actuelle (Figure VIII.2).
Les rectangles symbolisent les cas où le nombre de pertes est supérieur au nombre de gains.
Les couleurs des noms des souches symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit. L'arbre phylogénétique est construit comme décrit dans la section C.4 (Annexe) mais les longueurs de branches sont arbitraires et ne reflètent pas les taux d'évolution.

VIII.1 Arbre de gains et pertes de familles de gènes

De nombreux gains et pertes de familles de gènes ont lieu tout le long de l'arbre phylogénétique (Figure VIII.3). Sur 12 des 16 branches internes, les pertes de familles de gènes dépassent les gains. Dans les branches terminales, à l'exception des souches CC9902, MIT9312 et MED4, le nombre de gains est plus élevé que le nombre de pertes, reflétant ainsi l'acquisition des gènes orphelins, spécifiques aux souches.

Dans notre étude, les nombres de pertes et gains entre la racine de l'arbre et les branches terminales sont supérieurs à ceux de Luo *et al.* (2011) et Sun et Blanchard (2014) (Tableau VIII.2, tests des rangs signés de Wilcoxon unilatéraux significatifs avec 5% d'erreur). Les différences entre les résultats obtenus par Luo *et al.* (2011) et Sun et Blanchard (2014) et les nôtres peuvent venir de la définition des familles utilisées dans les analyses, de l'ajout des gènes orphelins pour les branches terminales ou des méthodes utilisées pour

Souches	Nombre total de pertes de familles			Nombre total de gains de familles			Ratio pertes-gains		
	Luo *et al.* (2011)	Sun et Blanchard (2014)	Notre étude	Luo *et al.* (2011)	Sun et Blanchard (2014)	Notre étude	Luo *et al.* (2011)	Sun et Blanchard (2014)	Notre étude
MED4	466	608	1 551	70	364	790	6.66	1.67	1.96
MIT9515	485	631	1 464	74	372	895	6.55	1.70	1.64
MIT9312	458	594	1 499	64	365	824	7.16	1.63	1.82
MIT9215	494	602	1 424	77	400	919	6.42	1.51	1.55
MIT9301	478	605	1 458	70	343	895	6.83	1.76	1.63
AS9601	471	594	1 429	67	343	876	7.03	1.74	1.63
NATL1A	416	555	1 186	54	478	846	7.70	1.16	1.40
NATL2A	416	564	1 165	47	456	793	8.85	1.24	1.47
SS120	433	581	1 041	71	327	450	6.10	1.78	2.31
MIT9211	427	576	1 088	53	297	477	8.06	1.94	2.28
MIT9313	216	298	718	78	750	439	2.77	0.40	1.63
MIT9303	213	293	606	83	857	1 036	2.57	0.34	0.58
CC9311	69	133	317	116	685	625	0.59	0.19	0.51
WH7803	63	120	278	126	426	273	0.5	0.28	1.02
WH8102	60	125	594	63	546	544	0.95	0.23	1.09
CC9605	68	129	516	82	708	587	0.83	0.18	0.88
CC9902	113	189	592	64	407	377	1.77	0.46	0.64

Table VIII.2 – Flux des familles de gènes pour les 17 souches de *Prochlorococcus* et de *Synechococcus* depuis l'ancêtre commun pour les données de Luo *et al.* (2011), Sun et Blanchard (2014) et notre étude.

reconstruire les gains et les pertes. Ainsi, même si les reconstructions des gains et des pertes de gènes de Luo *et al.* (2011) et Sun et Blanchard (2014) sont basées dans les deux cas sur la parcimonie, les méthodes diffèrent. Luo *et al.* (2011) ont défini d'abord les familles présentes chez l'ancêtre commun, soit 1 872 familles présentes dans au moins un génome de *Prochlorococcus* ou de *Synechococcus* et un génome d'un groupe externe. La reconstruction des gains et pertes démarre à partir de ces familles. Dans leur analyse, Sun et Blanchard (2014) reconstruisent le contenu de gènes ancestraux, 2 028 familles de gènes, après celui des nœuds intermédiaires. Notre méthode, bien que n'utilisant pas la parcimonie, repose sur le même principe que Sun et Blanchard (2014). Nous obtenons ainsi 2 333 familles ancestrales.

Pour l'ensemble des branches de *Prochlorococcus*, avec 3 679 familles gagnées et 2 984 familles perdues, le ratio de 0.811 entre le nombre de pertes et le nombre de gains (nommé ratio perte-gain dans la suite) est largement inférieur à celui estimé par Luo *et al.* (2011). Ce ratio descend à 0.218 pour les branches des souches non réduites de *Prochlorococcus* LLIV, significativement inférieur à celui des branches des souches réduites de *Prochlorococcus*, pour lesquelles le ratio est de 2.27 ($P < 2.2 \cdot 10^{-16}$, test de χ^2 à un degré de liberté). Ainsi, même pour l'ensemble des branches conduisant aux souches réduites de *Prochlorococcus*, le nombre de gains excède le nombre de pertes, reflétant les nombreux gains de gènes orphelins dans les branches terminales. En effet, lorsque ces gains ne sont pas pris en compte, le ratio perte-gain est de 2.30 pour l'ensemble des branches conduisant

aux souches de *Prochlorococcus*, 1.39 pour celles conduisant aux souches non réduites de *Prochlorococcus* LLIV et 2.08 pour celles conduisant aux souches réduites de *Prochlorococcus*.

Pour toutes les souches réduites de *Prochlorococcus*, le nombre de pertes depuis la racine jusqu'aux feuilles de l'arbre est supérieur au nombre de gains (Tableau VIII.2, $P = 0.002531$, test des rangs signés de Wilcoxon unilatéral). Lorsque les gains de gènes orphelins ne sont pas pris en compte, les pertes excèdent les gains même pour les souches non réduites de *Prochlorocccus* LLIV. En effet, dans la branche ancestrale aux souches de *Prochlorococcus*, le ratio perte-gain dépasse 11. Ainsi, le long de cette branche, en accord avec les observations de Sun et Blanchard (2014), de nombreux gènes sont perdus, initiant l'évolution réductive avant la diversification des souches de *Prochlorococcus* (Figure VIII.3). Cette évolution réductive continue jusqu'à la divergence de *Prochlorocccus* HL (Figure VIII.3). Dans la branche ancestrale de ces souches, le ratio perte-gain est inférieur à 1 à cause de nombreux gains de familles de gènes. L'évolution réductive reprend après la diversification des différentes souches de *Prochlorococcus* HL, entrainant un ratio perte-gain de 1.37, supérieur à celui de l'ensemble des souches réduites de *Prochlorococcus* (ratio = 0.911). Les gènes perdus dans les branches conduisant aux souches de *Prochlorococcus* HL correspondent principalement à des gènes gagnés dans la branche ancestrale à ces souches alors que dans les autres branches, les gènes perdus étaient principalement issus des familles de gènes de l'ancêtre (Figure VIII.4).

VIII.2 Annotations des gènes gagnés et perdus

Les gènes perdus et gagnés identifiés par Sun et Blanchard (2014) ne semblent pas affecter des catégories particulières : la répartition des gènes des génomes actuels dans les différentes catégories semble relativement conservée parmi les différentes souches (Sun et Blanchard, 2014). Est-ce aussi le cas avec nos données ? Les répertoires de gènes potentiellement impliqués dans la réparation de l'ADN diffèrent entre les souches (Partensky et Garczarek, 2010). Quand les gènes de réparation ont-ils été perdus ou gagnés le long de la phylogénie de *Prochlorococcus* ?

Pour étudier l'évolution des catégories des familles gagnées et perdues, nous avons assigné aux 3 778 familles utilisées des catégories COG selon la méthode présentée à la section C.5 (en annexe). Cependant, seulement 44% des familles ont pu être catégorisées, ce qui est susceptible d'affecter nos résultats.

Alors que les familles gagnées sont uniformément réparties entre les catégories (information, processus cellulaires, métabolisme et catégories mal caractérisées ; $P > 0.05$, tests de Student avec correction de Bonferroni pour les comparaisons multiples), la répartition des familles perdues est plus disparate (Figure VIII.5), contrairement aux observations de Sun et Blanchard (2014). Ainsi, les familles perdues sont principalement liées au métabolisme, mais seulement faiblement aux processus de stockage et gestion de l'information,

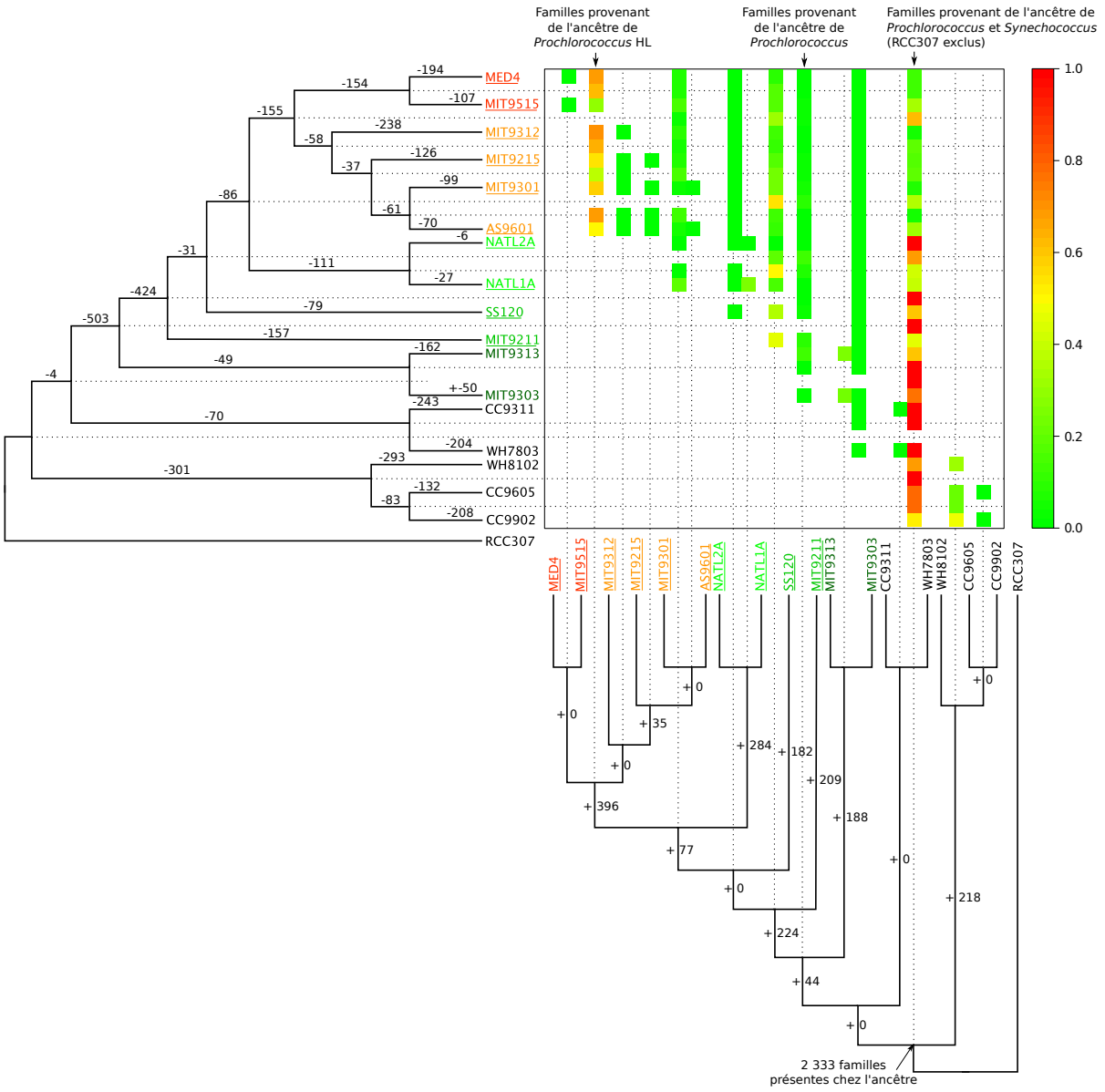


Figure VIII.4 – Origine des familles de gènes perdues le long de l'arbre phylogénétique de *Prochlorococcus* et *Synechococcus*
Chaque valeur du graphique correspond à la proportion des familles de gènes gagnées dans la branche en colonne puis perdues dans la branche en ligne.
Les couleurs des noms des souches symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit. L'arbre phylogénétique est construit comme décrit dans la section C.4 (Annexe) mais les longueurs des branches sont arbitraires et ne reflètent pas les taux d'évolution.

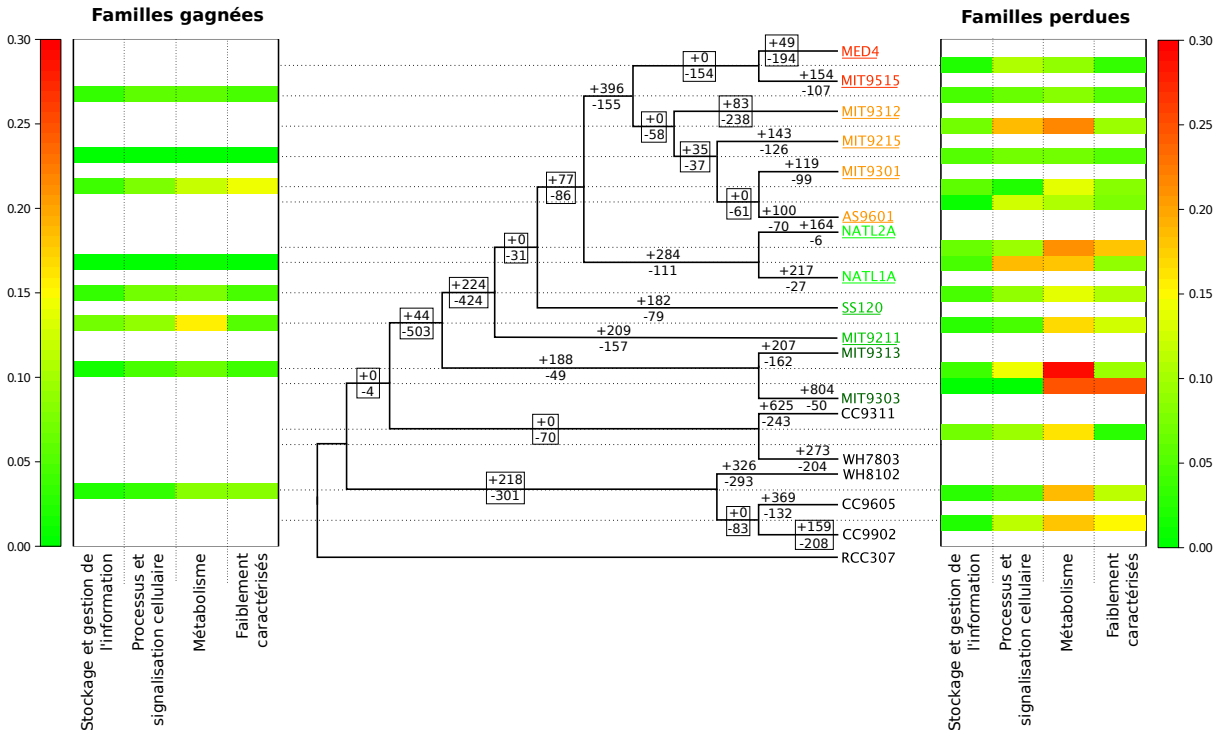


Figure VIII.5 – Proportion par branche des catégories COG des familles de gènes gagnées et perdues le long de l'arbre phylogénétique de *Prochlorococcus* et *Synechococcus*
L'assignation des catégories COG aux familles de gènes perdues ou gagnées est faite selon la méthode présentée à la section C.5 (en annexe). Seules 44% des familles ont pu être catégorisées.
Chaque valeur est associée à la branche traversée par la ligne en pointillé à la même hauteur.
Les couleurs des noms des souches symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit. L'arbre phylogénétique est construit comme décrit dans la section C.4 (Annexe) mais les longueurs de branches sont arbitraires et ne reflètent pas les taux d'évolution.

et de façon similaire aux processus cellulaires et aux catégories mal caractérisées (tests de Student avec correction de Bonferroni pour les comparaisons multiples). En effet, les processus de stockage et gestion de l'information génomique sont sujets à une forte conservation entre les espèces et sont peu sujets aux pertes, ou alors de façon très délétères pour les organismes. Cependant, le contenu de 28 familles potentiellement liées à la réplication, la recombinaison et la réparation de l'ADN, familles identifiées par Partensky et Garczarek (2010), a évolué le long de la phylogénie (Figure VIII.6) par des pertes et des gains. Au cours de l'évolution réductive, de nombreuses pertes de familles liées principalement à la réparation de l'ADN ont donc eu lieu, même si certaines sont seulement temporaires, car la famille a été réacquise par la suite. Ainsi, toutes les familles perdues dans la branche ancestrale de *Prochlorococcus* ont été réacquises soit le long de la branche ancestrale aux souches de *Prochlorococcus* HL, soit le long de la branche ancestrale aux souches de *Prochlorococcus* HL et LLI (Figure VIII.6). C'est le cas des gènes liés à la production d'enzymes de réparation des dommages causés par les UV, utiles pour la vie

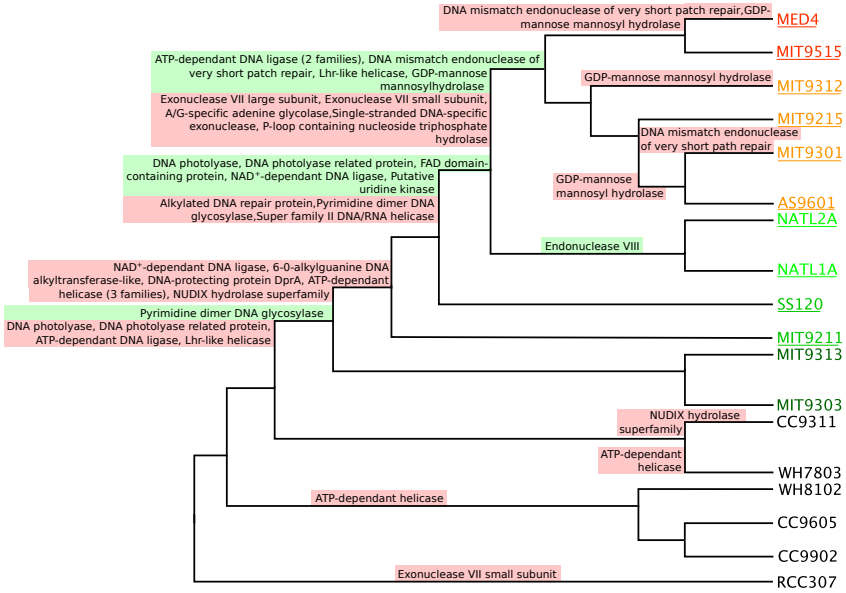


Figure VIII.6 – Gains et pertes de familles de gènes potentiellement impliquées dans la réplication, la recombinaison et la réparation de l'ADN
Les gènes potentiellement impliqués dans la réplication, la recombinaison et la réparation de l'ADN qui sont présents seulement dans un sous-ensemble de souches de *Prochlorococcus* sont issus du tableau 3 de Partensky et Garczarek (2010). Les gains et pertes des familles correspondant à ces gènes ont été reconstruits avec la méthode décrite précédemment. Les produits de ces gènes sont indiqués dans cette figure. Les rectangles rouges correspondent aux familles perdues et les rectangles verts aux familles gagnées le long d'une branche.
Les couleurs des noms des souches symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit. L'arbre phylogénétique est construit comme décrit dans la section C.4 (Annexe) mais les longueurs des branches sont arbitraires et ne reflètent pas les taux d'évolution.

dans un environnement proche de la surface de l'océan. Cependant, en dehors de pertes le long de la branche ancestrale de *Prochlorococcus*, la plupart des pertes de familles liées à la réparation de l'ADN semblent définitives.

VIII.3 Discussion

Les gains dans l'histoire des souches de *Synechococcus* et non réduites de *Prochlorococcus* étant inférieurs aux pertes dans la phylogénie des souches réduites de *Prochlorococcus* ($P = 5.142 \cdot 10^{-5}$, test des rangs signés de Wilcoxon unilatéral), les différences de taille de génomes entre *Prochlorococcus* et *Synechococcus* ne sont pas simplement dues aux gains pour les souches de *Synechococcus* et non réduites de *Prochlorococcus* comme l'avaient suggéré Kettler *et al.* (2007). La phase initiale de réduction des génomes est en accord

avec les observations de Sun et Blanchard (2014), avec de nombreuses pertes de familles de gènes peu après la séparation entre *Prochlorocccus* et *Synechococcus* et avant la diversification de *Prochlorococcus*, selon les auteurs, à cause d'une forte sélection pour la simplification des génomes entrainant la perte de gènes ayant de petits effets sur la fitness (Sun et Blanchard, 2014). Une autre phase de réduction des génomes semble avoir eu lieu après le changement d'environnement, au moment de la diversification des souches *Prochlorocccus* HL, en accord avec les observations de Luo *et al.* (2011). La présence de deux phases n'est cependant pas si évidente. En effet, des pertes de familles ont lieu tout le long de la phylogénie dans des proportions similaires comme pour un processus continu.

L'évolution réductive chez *Prochlorococcus* aurait donc démarré peu après la divergence avec *Synechococcus*, avec une réduction du nombre de gènes accompagnée de l'acquisition de quelques gènes. L'initiation de cette réduction peut être due aux changements d'environnements sous-jacents à la divergence entre *Synechococcus* et *Prochlorococcus*. La réduction du nombre de gènes s'arrête au sein du clade de *Prochlorococcus* LLIV, avec même une forte expansion du génome pour *Prochlorococcus* MIT9303 par l'acquisition de nombreux gènes orphelins. Cet arrêt de la réduction est attendu sous l'hypothèse adaptative : une fois adaptées à leur nouvel environnement, ces souches ne devraient plus changer de taille et de répertoire génique. Or, la réduction des génomes continue dans les branches conduisant aux souches réduites de *Prochlorococcus*, avec des pertes plus nombreuses que les gains.

Cependant, dans la branche ancestrale aux souches de *Prochlorococcus* HL, l'évolution réductive semble arrêtée puisque le nombre d'acquisitions de gènes est supérieur au nombre de pertes (même si celles-ci restent cependant nombreuses). Ces gains pourraient être dus à un changement de niche, initié avant la divergence des souches de *Prochlorococcus* LLI mais ayant surtout lieu le long de la branche ancestrale aux souches de *Prochlorococcus* HL. De fait, les souches de *Prochlorococcus* HL vivent dans un environnement proche de la surface de l'océan, riche en lumière. Les besoins sont donc différents du bas de la colonne d'eau où vivent les écotypes de *Prochlorococcus* LL. Cette explication adaptative des gains peut en particulier expliquer la réacquisition de gènes de réparation des dommages causés par les UV, gènes perdus à la divergence entre *Prochlorococcus* et *Synechococcus*. Cependant, l'explication adaptative ne permet pas de comprendre les nombreuses pertes définitives de familles liées à la réparation de l'ADN. En particulier, la perte des gènes impliqués dans la formation des exonucléases VII pourrait entrainer une augmentation des taux de réarrangement, comme observé pour des mutants d'*E. coli* (Chase et Richardson, 1977). L'augmentation des taux de mutation peut temporairement favoriser l'adaptation à un nouvel environnement, mais l'effet à long terme est délétère.

Après la divergence des différentes souches de *Prochlorococcus* HL, les nombres de gains redeviennent inférieurs aux nombres de pertes. Alors que dans les autres branches les pertes touchent principalement les 2 333 familles ancestrales, les pertes dans les branches conduisant aux souches de *Prochlorococcus* HL correspondent essentiellement aux familles gagnées récemment dans la branche ancestrale aux souches de *Prochlorococcus* HL.

Ces pertes ainsi que celles dans les autres branches sont-elles dues à un relâchement des

pressions de sélection qui ferait que seuls les gènes indispensables sont conservés et les autres éliminés, rongés par les biais vers les délétions, comme chez les endosymbiotes ? Ce relâchement des pressions de sélection pourrait avoir été initié après la divergence des *Prochlorococcus* LLIV, expliquant que ces dernières n'aient pas subi d'évolution réductive. Cette hypothèse est discutée dans le chapitre XII, tout comme la cause des pertes ininterrompues de gènes tout le long de la phylogénie des souches réduites de *Prochlorococcus*.

Chapitre IX

Évolution de la longueur des gènes

Lors de l'évolution réductive, la taille des génomes décroît au sein de certaines lignées principalement par la perte de gènes et de séquences intergéniques. Qu'en est-il pour la taille des gènes ? D'après Wang *et al.* (2011), lorsque le nombre de gènes codant pour des protéines décroît, la taille des protéines et donc des gènes décroît aussi. Chez les endosymbiotes, la taille des protéines des endosymbiotes ne semble pas se réduire de façon uniforme avec la réduction du génome, mais présente une grande variabilité (Kenyon et Sabree, 2014). Pour *Prochlorococcus*, l'évolution de la longueur des gènes a fait l'objet de deux analyses montrant l'absence de différence de longueur des gènes entre *Synechococcus* et *Prochlorococcus* (Sun et Blanchard, 2014) et entre les souches réduites et non réduites (Marais *et al.*, 2008). Cependant, ces analyses sont sommaires : comparaison des longueurs de gènes orthologues d'une souche réduite et d'une souche non réduite (Marais *et al.*, 2008) et comparaison statistique simple de la taille moyenne des gènes de *Synechococcus* et de *Prochlorococcus* (Sun et Blanchard, 2014). Il serait ainsi intéressant d'étudier plus en détail l'évolution des longueurs des gènes. L'évolution réductive ne change-t-elle pas les longueurs des gènes, comme cela semble être le cas d'après les deux analyses effectuées (Marais *et al.*, 2008; Sun et Blanchard, 2014) ? Les potentiels changements sont-ils du même ordre entre les endosymbiotes et *Prochlorococcus* ? Y a-t-il un gradient de réduction des gènes le long de l'arbre phylogénétique ? S'il y a des réductions de la taille des gènes, les pertes ont-elles lieu plutôt au milieu des gènes ou aux extrémités, où elles pourraient être moins délétères ?

Pour répondre à ces différentes questions, les longueurs des gènes sont comparées entre les souches, gène à gène, en utilisant les 693 familles de gènes orthologues aux 12 souches de *Prochlorococcus* et aux 6 souches de *Synechococcus*, récupérées selon la méthode décrite dans la section C.1.2 (en annexe). Nous souhaitons comparer l'évolution des longueurs de gènes chez *Prochlorococcus* à celle chez les endosymbiotes, en prenant en compte la phylogénie. Nous utilisons *Buchnera aphidicola* comme exemple d'endosymbiote et comparons à *Escherichia coli*, une bactérie libre proche phylogénétiquement. Avec la même méthodologie que celle utilisée pour récupérer les séquences d'intérêt chez *Prochlorococcus* (Section C.1.2), 226 familles de gènes orthologues à 5 souches de *Buchnera* et 26 souches d'*E. coli*

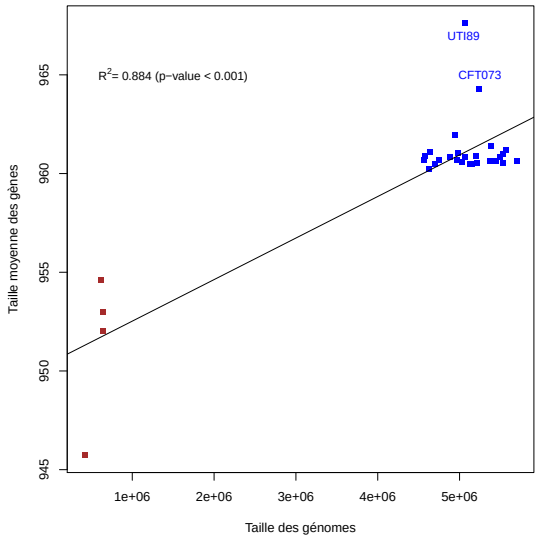
sont utilisées. Comme ces gènes sont conservés entre les différentes souches, les pressions de sélection pour le maintien de l'information des séquences sont probablement fortes. L'absence de changement de la longueur de ces gènes pourrait être due à de fortes pressions de sélection pour la conservation de l'information, au contraire des gènes spécifiques aux souches où des changements de longueurs pourraient être plus visibles. Cependant, la comparaison de longueurs pour des gènes non orthologues serait plus délicate à interpréter. Seuls les changements de longueur constatés sur des gènes orthologues ont la capacité de révéler directement une éventuelle pression (mutationnelle ou sélective) sur la longueur des gènes.

IX.1 Différence de longueur des gènes : cas de *Buchnera* et de *Prochlorococcus*

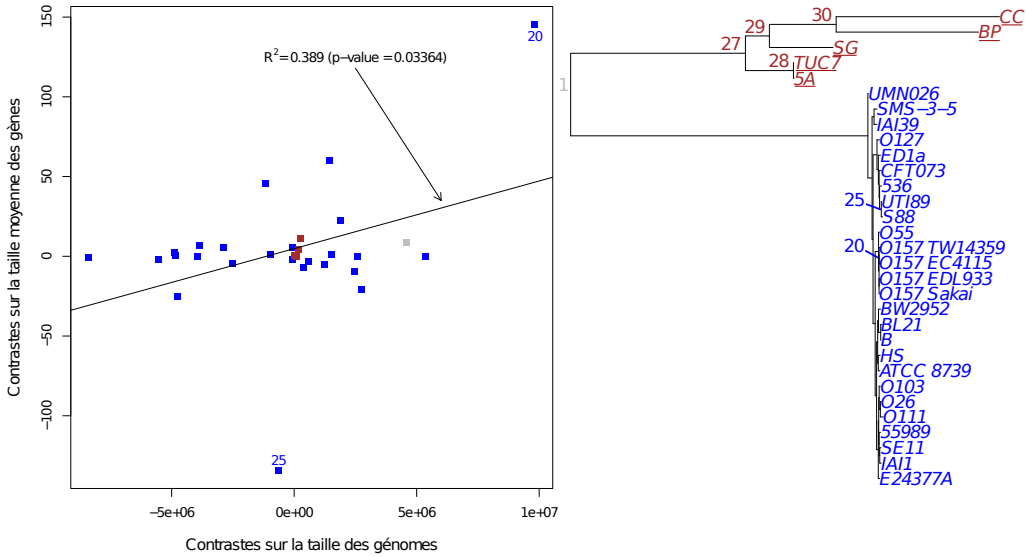
Avec la réduction des génomes et du nombre de gènes, la taille des protéines devrait décroitre (Wang *et al.*, 2011), entraînant une corrélation positive entre taille du génome et taille moyenne des gènes. Cette corrélation s'observe à la fois pour les endosymbiotes (Figure IX.1a) et pour *Prochlorococcus* (Figure IX.2a).

Les souches de *Buchnera* présentent cependant une grande variabilité de longueur des gènes avec des valeurs moyennes relativement proches de celles d'*E. coli* (Figure IX.1a), comme ce qui était déjà observé par Kenyon et Sabree (2014). Chez *Prochlorococcus*, les souches réduites LL ont des tailles moyennes de gènes proches de celles des souches de *Synechococcus* (Figure IX.2a), remettant en cause la relation entre réduction de la taille des génomes et réduction de la taille des gènes. De plus, au sein des différents clades, il n'y a pas de signes d'une corrélation entre la taille des génomes et la taille des gènes. Toute la corrélation observée semble provenir de la différence entre les clades, c'est-à-dire des changements des deux caractères étudiés entre les clades. Dans changements de la longueur des génomes et de gènes dans les séquences des organismes ancestraux affectent les génomes de tous les descendants. Il est alors important de prendre en compte la phylogénie sous-jacente pour étudier la corrélation entre des caractères comme la taille des génomes et la taille des gènes. Dans ce but, nous utilisons la méthode des contrastes phylogénétiquement indépendants (Felsenstein, 1985). Cette méthode renvoie pour chaque nœud d'un arbre phylogénétique le contraste pour un caractère entre les branches filles du nœud, la différence pour le caractère entre les branches filles en fonction des longueurs des branches, c'est-à-dire les distances évolutives entre les nœuds fils (Figure VII.1).

Pour les endosymbiotes et *E. coli*, la corrélation entre la taille des génomes et la taille moyenne des gènes est bien moindre lorsque la phylogénie est prise en compte (Figure IX.1b), et elle est principalement due à l'alignement du nuage de points avec deux points extrêmes. Le point en bas, le numéro 25, correspond à la divergence entre *E.coli* UTI89 et S88 (Figure IX.1b). Or UTI89 a la valeur moyenne des longueurs de gènes maximale et S88 une valeur au sein du nuage de points d'*E.coli* (Figure IX.1a). Ainsi, le contraste entre ces deux souches sur la longueur moyenne des gènes est importante.

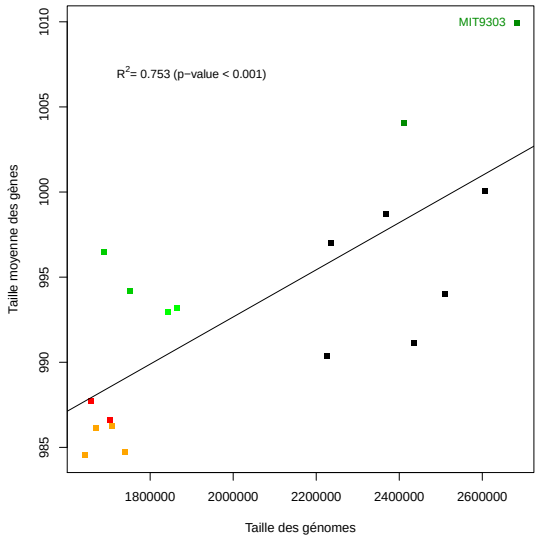


(a) Taille moyenne des gènes en fonction de la taille des génomes (valeurs brutes)

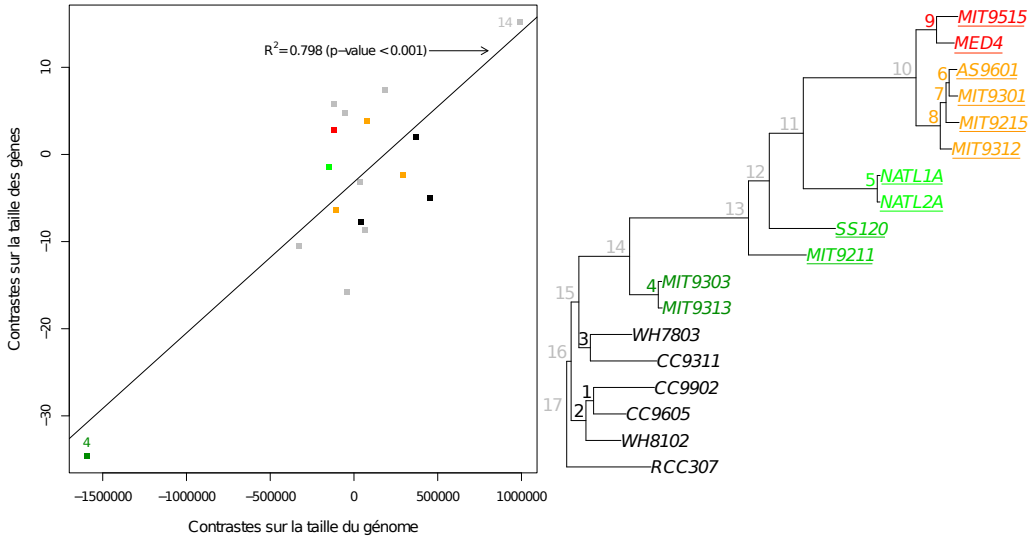


(b) Contrastes phylogénétiques sur la taille moyenne des gènes orthologues en fonction des contrastes phylogénétiques sur de la taille des génomes

Figure IX.1 – Taille moyenne des 226 gènes orthologues en fonction de la taille des génomes pour *Buchnera* et *E. coli*, avec et sans prise en compte de la phylogénie sous-jacente
Les couleurs symbolisent les différentes espèces avec en bleu les souches *E. coli* et en rouge foncé les souches *Buchnera*. Les droites en noir sont les droites de régression linéaire. Les valeurs de corrélation sont indiquées avec leur p-value.
Pour les valeurs brutes, les points correspondent aux caractéristiques des différentes souches *E. coli* et *Buchnera*. Pour les contrastes phylogénétique, chaque point représente un nœud de l'arbre phylogénétique à droite, avec en gris les nœuds ancestraux à des souches de plusieurs espèces. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit. L'arbre phylogénétique est construit avec la même méthode que celle utilisée pour la construction de l'arbre phylogénétique de *Prochlorococcus* (Section C.4, en annexe).



(a) Taille moyenne des gènes en fonction de la taille des génomes (valeurs brutes)



(b) Contraste phylogénétique sur la taille moyenne des gènes orthologues en fonction du contraste phylogénétique sur de la taille des génomes

Figure IX.2 – Taille moyenne des 693 gènes orthologues en fonction de la taille des génomes pour *Prochlorococcus* et *Synechococcus*, avec et sans prise en compte de la phylogénie sous-jacente
Les couleurs symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les droites en noir sont les droites de régression linéaire. Les valeurs de corrélation sont indiquées avec leur p-value.
Pour les valeurs brutes, les points correspondent aux caractéristiques des différentes souches de *Synechococcus* et de *Prochlorococcus*. Pour les contrastes phylogénétiques, chaque point représente un nœud de l'arbre phylogénétique à droite, avec en gris les nœuds ancestraux à des souches de plusieurs écotypes. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit. L'arbre phylogénétique est construit comme décrit dans la section C.4 (en annexe).

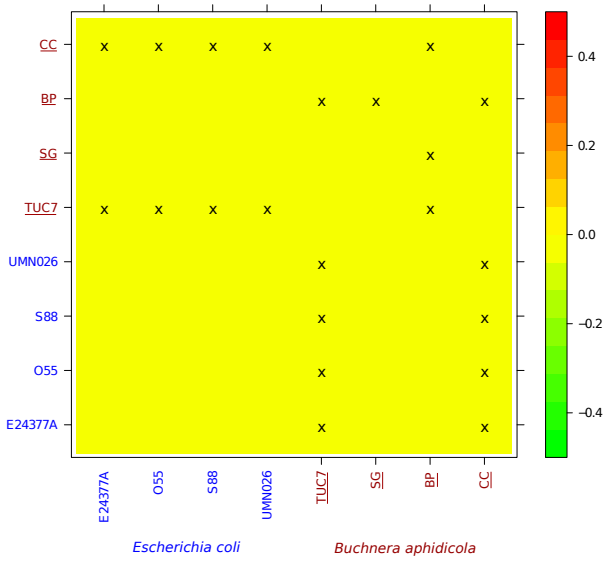


Figure IX.3 – Différences de longueurs des gènes entre *Buchnera* et *E. coli* pour 226 familles de gènes orthologues
Les couleurs sur le graphique symbolisent la médiane de la différence de longueurs gène à gène entre une souche en ligne et une souche en colonne. Les croix correspondent aux cas où le test des rangs signés de Wilcoxon de la différence des longueurs de gènes entre les souches est significatif avec une erreur de 5%.
Les souches utilisées pour la comparaison ont été choisies pour représenter des portions des arbres phylogénétiques, les tendances étant conservées au sein de ces portions. Les couleurs de noms symbolisent les différentes espèces avec en bleu les souches *E. coli*, en rouge foncé les souches *Buchnera*. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit.

Au sein des *Buchnera*, les contrastes sur la taille des génomes et la taille des gènes sont quasiment nuls (Figure IX.1b). Les différences de ces deux caractères au sein des *Buchnera* sont alors minimes par rapport à celles pour *E.coli*. À la racine de l'arbre, à la différenciation entre *Buchnera* et *E.coli*, le contraste est nul pour la longueur des gènes mais relativement important pour la taille des génomes (Figure IX.1b). Ainsi les changements de tailles des génomes entre *E. coli* et *Buchnera* ne s'accompagnent de réduction ou d'augmentation globale de la taille des gènes. De plus, les différences de longueur gène à gène entre *Buchnera* et *E. coli* ont des médianes nulles (Figure IX.3), malgré des différences parfois significatives. Certaines familles sont plus longues pour les souches d'*E. coli* que pour les souches de *Buchnera* et vice-versa. Le clustering hiérarchique appliqué sur les longueurs de gènes rassemble les souches d'*E. coli* d'un côté et les souches de *Buchnera* de l'autre. Les gènes des familles ont des longueurs similaires pour les souches au sein de chacun des deux groupes, mais les longueurs sont différentes entre les groupes.

Si la relation entre la taille des génomes et la taille des protéines émise par Wang *et al.* (2011) ne semble pas évidente pour les endosymbiotes et *E.coli*, elle semble tenir pour *Prochlorococcus* et *Synechococcus* (Figure IX.2b). En effet, les contrastes sur la taille des gènes orthologues semblent corrélés avec les contrastes sur la taille des génomes avec un coefficient de corrélation proche de 0.8 (Figure IX.2b). Cette relation est cependant due à un alignement entre deux points extrêmes et le nuage de points. Une fois ces deux points

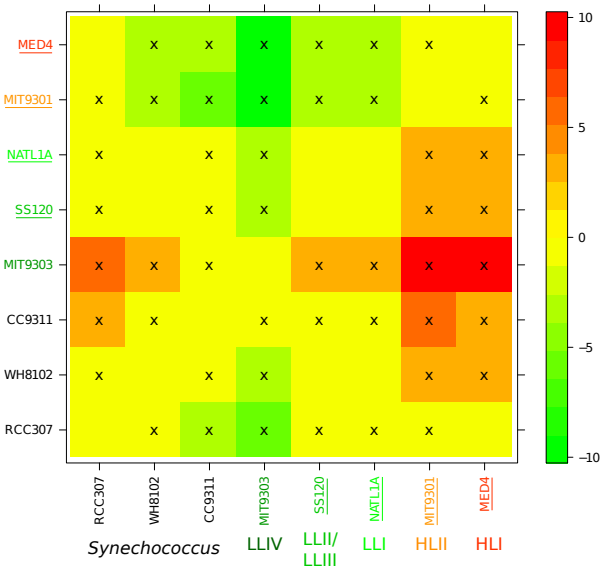


Figure IX.4 – Différences de longueurs des gènes entre les différentes souches de *Prochlorococcus* et *Synechococcus* pour 697 familles de gènes
Les couleurs sur le graphique symbolisent la médiane de la différence de longueurs gène à gène entre une souche en ligne et une souche en colonne. Les croix correspondent aux cas où le test des rangs signés de Wilcoxon de la différence des longueurs de gènes entre les souches est significatif avec une erreur de 5%.
Les souches utilisées pour la comparaison ont été choisies pour représenter des portions des arbres phylogénétiques, les tendances étant conservées au sein de ces portions. Les couleurs de noms des souches symbolisent les différents écotypes avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit.

éliminés, la corrélation ne tient plus. Néanmoins, ces deux points ont du sens et fournissent des informations importantes sur l'évolution de la longueur des génomes et des gènes au sein de *Prochlorococcus*.

- Le point en bas à gauche correspond au nœud différenciant MIT9303 et MIT9313, les deux souches non réduites de *Prochlorococcus*. Alors que MIT9313 est proche des souches de *Synechococcus*, MIT9303 possède des gènes plus grands au sein du plus grand génome (Figure IX.2a), où de nombreux gènes orphelins ont été acquis (Figure VIII.3). La différence de la taille des gènes entre ces deux souches est-elle due à des insertions au sein des gènes de MIT9303 dans un contexte d'expansion du génome, comme pourraient le suggérer les nombreuses acquisitions récentes de gènes ?

- L'autre point extrême, en haut à droite, est le nœud différenciant les souches réduites des souches non réduites de *Prochlorococcus* (Figure IX.2b). La différence de taille de génome entre ces groupes (fort contraste) serait ainsi accompagnée d'une différence de taille des gènes orthologues (dans le même sens).

Les différences de longueurs gènes à gènes sont significatives pour la plupart des paires de souches (Figure IX.4), avec des médianes nettement supérieures à celles observées chez les endosymbiotes (Figure IX.3). Ainsi, les différences de longueurs entre les souches non réduites (MIT9303, MIT9313) et les souches réduites de *Prochlorococcus* HL (MED4, MIT9515, AS9601, MIT9312, MIT9301, MIT9215) atteignent 10 bases (Figure IX.4). Les gènes des souches réduites de *Prochlorococcus*, en particulier les souches HL, sont globalement plus courts que ceux des souches non réduites de *Prochlorococcus*.

Ainsi, au cours de l'évolution réductive chez *Prochlorococcus*, génomes et gènes se seraient réduits. Les évènements de réduction des gènes correspondent-ils aux évènements de pertes de gènes ? Sont-ils des évènements convergents au sein de chacune des souches ou des évènements dans les séquences ancestrales ? Des changements dans les autres branches ont-ils lieu ? Les réductions des gènes touchent-elles les extrémités ou l'intérieur des gènes, c'est-à-dire des portions subissant des pressions différentes ?

IX.2 Étude des insertions et des délétions et de leur impact sur la longueur des gènes : cas de *Buchnera* et de *Prochlorococcus*

La reconstruction des évènements d'insertions et de délétions au sein des différentes familles de gènes orthologues pourrait répondre aux questions précédentes. Lors de l'alignement des familles de gènes (section C.2, en annexe) avec l'outil *Prank* (Löytynoja et Goldman, 2005), les évènements de changement de séquences, en particulier les insertions et les délétions, sont fournis pour chacune des branches de l'arbre phylogénétique de la famille de gènes étudiée et peuvent être utilisés pour reconstruire l'évolution de la longueur des gènes. Cependant, ces évènements sont ainsi liés aux branches de l'arbre de la famille de gènes. Or, ce dernier est souvent différent de l'arbre des espèces. Se pose ainsi la question de la réconciliation des évènements ayant lieu sur un arbre de gènes avec un arbre des espèces, afin de construire un arbre recensant les évènements d'insertions et de délétions au sein des gènes (Figure IX.5).

La réconciliation s'effectue en élaguant, dans les arbres de familles de gènes, les portions qui sont en conflit avec l'arbre des espèces. Pour cela, la fiabilité des branches des arbres des familles est évaluée à l'aide des rapports de vraisemblance approximée (aLRT) (Anisimova et Gascuel, 2006) calculés par *PhyML* (Guindon et Gascuel, 2003). L'outil *Prunier* (Abby *et al.*, 2010) est ensuite utilisé pour la recherche de transferts au sein des arbres de gènes par comparaison avec l'arbre des espèces. En effet, des évènements de transfert au sein des familles de gènes sont la cause la plus probable d'incohérence entre un arbre de gènes et l'arbre des espèces. Trouver ces évènements de transfert permet de déterminer des portions des arbres des familles de gènes qui sont en accord avec l'arbre des espèces. Nous utilisons ensuite ces portions d'arbres comme guide pour un nouvel alignement avec *Prank* (Löytynoja et Goldman, 2005) des séquences concernées. Cette étape permet d'obtenir

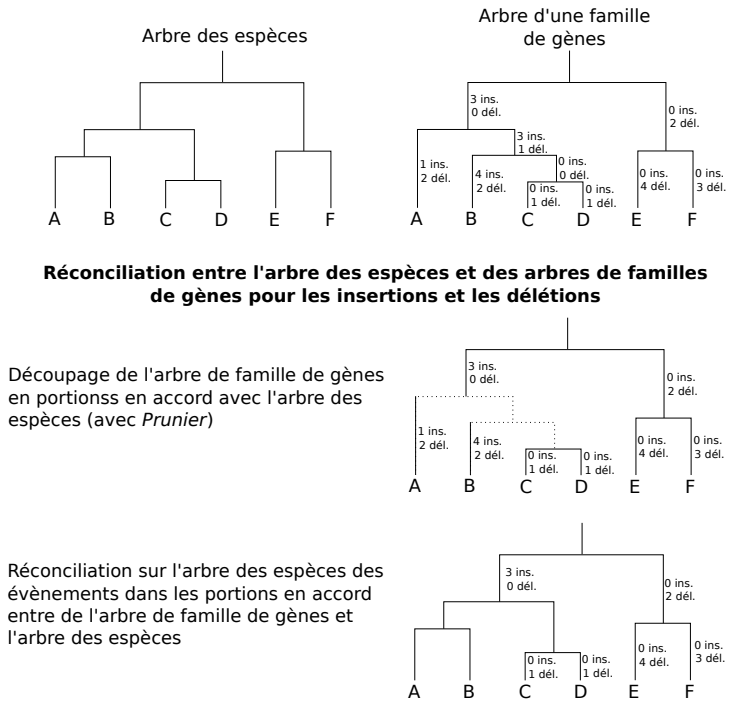


Figure IX.5 – Méthodologie de réconciliation des évènements d'insertion et délétion inférés sur un arbre de famille de gènes avec l'arbre des espèces

des alignements et des évènements d'insertions et délétions plus fiables puisque *Prank* n'a plus à estimer un arbre guide (source potentielle d'erreurs). Il suffit ensuite de reporter les évènements d'insertions et de délétions détectés par *Prank* sur l'arbre des espèces.

IX.2.1 Endosymbiotes

Pour *Buchnera* et *E. coli*, les insertions sont moins nombreuses que les délétions ($P = 2.432 \cdot 10^{-3}$, test des rangs signés de Wilcoxon unilatéral, Figure IX.6), mais de taille moyenne similaire ($P = 0.06545$, test des rangs signés de Wilcoxon en éliminant les comparaisons pour lesquelles le nombre d'insertions et/ou délétions est nul, soit 32 cas sur 61), avec des évènements plutôt intra-génique qu'aux extrémités des gènes ($P = 0.02929$, test des rangs signés de Wilcoxon), ces évènements étant de taille moyenne comparable ($P = 0.2941$, test des rangs signés de Wilcoxon en éliminant les comparaisons pour lesquelles le nombre d'évènements terminaux et/ou non terminaux est nul, soit 39 cas sur 61). Globalement, la taille des gènes semble s'être réduite : la taille totale des insertions est ainsi inférieure à celle des délétions ($P = 2.72 \cdot 10^{-3}$, test des rangs signés de Wilcoxon unilatéral), les délétions étant plus nombreuses. Cette observation est en contradiction avec les observations précédentes sur les différences de longueur des gènes. Les réductions ont principalement lieu le long des branches ancestrales aux souches de *Buchnera* et d'*E. coli* (Figure IX.7), avec des réductions jusqu'à 7.5 bases dans la branche ancestrale aux souches de *Buchnera*. Cependant, ces pertes dans les branches ancestrales pourraient être un biais méthodologique. Pour les cas d'incohérences entre les séquences de *Buchnera*

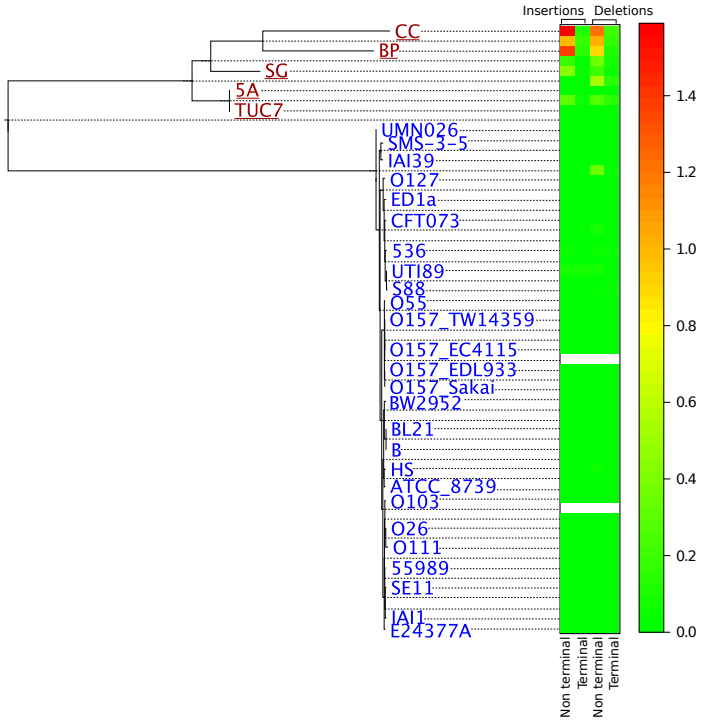


Figure IX.6 – Nombre d'insertions et délétions rapportés au nombre de familles par branche le long de l'arbre phylogénétique de *Buchnera* et *E. coli*.
Les évènements d'insertions et délétions ont été déterminés avec *Prank* (Löytynoja et Goldman, 2005) sur 226 familles de gènes orthologues, après réconciliation des arbres des gènes et de l'arbre des espèces avec *Prunier* (Abby *et al.*, 2010) (Figure IX.5). Les évènements dits terminaux correspondent aux évènements ayant lieu en début ou fin de gène et les évènements non terminaux aux autres.
Les couleurs de noms des souches symbolisent les différentes espèces avec en bleu les souches *E. coli* et en rouge foncé les souches *Buchnera*. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit. L'arbre phylogénétique est construit avec la même méthode que celle utilisée pour la construction de l'arbre phylogénétique de *Prochlorococcus* (Section C.4, Annexe).

et les séquences d'*E. coli*, *Prank* estime que les différentes possibilités trouvées dans les séquences sont présentes dans la séquence ancestrale et que des évènements de délétions ont conduit aux séquences observées. Cependant, ces incohérences pourraient aussi être dues à des évènements d'insertions et de délétions dans une seule des branches, la branche conduisant aux endosymbiotes, par exemple, comme suggéré par la forte variabilité des différences de longueur des gènes entre *E. coli* et *Buchnera*. Pour plus de fiabilité sur les évènements le long des branches ancestrales aux souches de *Buchnera* et d'*E. coli*, il faudrait ajouter un groupe externe de souches, comme *Vibrio cholerae* ou *Haemophilus influenzae*, afin de connaître l'état ancestral. Cette analyse n'a pas pu être faite dans le cadre de ce travail de thèse par manque de temps. En comparant les taux de délétions dans les deux branches, le taux de délétions total est légèrement supérieur dans la branche ancestrale aux *Buchnera* (1.8 fois plus fort). Ainsi, les souches le long de cette branche auraient subi légèrement plus de délétions au sein des gènes que le long de la branche conduisant aux souches d'*E. coli*.

Malgré une réduction globale des gènes (Figure IX.8), les tailles augmentent dans cer-

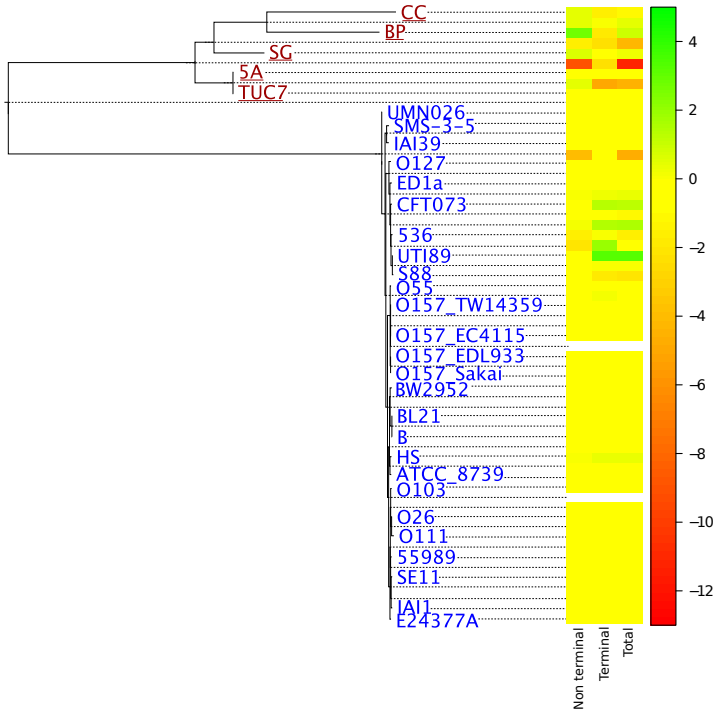


Figure IX.7 – Changement moyen de la longueur des gènes le long des branches de l'arbre phylogénétique de *Buchnera* et *E. coli*
Les évènements d'insertions et délétions ont été déterminés avec *Prank* (Löytynoja et Goldman, 2005) sur 226 familles de gènes orthologues, après réconciliation des arbres des gènes et de l'arbre des espèces avec *Prunier* (Abby *et al.*, 2010) (Figure IX.5). Les évènements dits terminaux correspondent aux évènements ayant lieu en début ou fin de gène et les évènements non terminaux aux autres.
Les couleurs de noms des souches symbolisent les différentes espèces avec en bleu les souches *E. coli* et en rouge foncé les souches *Buchnera*. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit. L'arbre phylogénétique est construit avec la même méthode que celle utilisée pour la construction de l'arbre phylogénétique de *Prochlorococcus* (Section C.4, Annexe).

taines branches. En particulier, l'augmentation de la taille des gènes par des évènements terminaux le long de la branche terminale de UTI89 (Figure IX.7) peut expliquer le fort contraste de longueur des gènes entre UTI89 et S88 (Figure IX.1b). De même, le long de la branche terminale de *Buchnera* BP, la taille des gènes a augmenté, principalement par des évènements non terminaux (Figure IX.7), entraînant ainsi une réduction moins importante des gènes par rapport aux autres souches de *Buchnera* (Figure IX.8).

Pour toutes les souches de *Buchnera* (Figure IX.8), la taille des gènes semble avoir été réduite depuis la racine, ce qui est en contradiction avec l'absence de différences de longueur des gènes avec *E. coli* (Figure IX.3). Or, comme les longueurs des gènes se réduisent aussi pour *E. coli* (Figure IX.8), il pourrait y avoir un phénomène de compensation : réduction pour certains gènes pour *Buchnera* et pour d'autres gènes pour *E. coli*, expliquant ainsi la forte variabilité des longueurs des gènes. Mais ces observations peuvent aussi être dues à l'absence de groupe externe pour déterminer de façon fiable les états ancestraux. Cependant, la réduction des gènes est plus importante chez *Buchnera*, à l'exception de la souche BP (Figure IX.8). La taille des gènes pourrait ainsi s'être réduite chez les endosymbiotes.

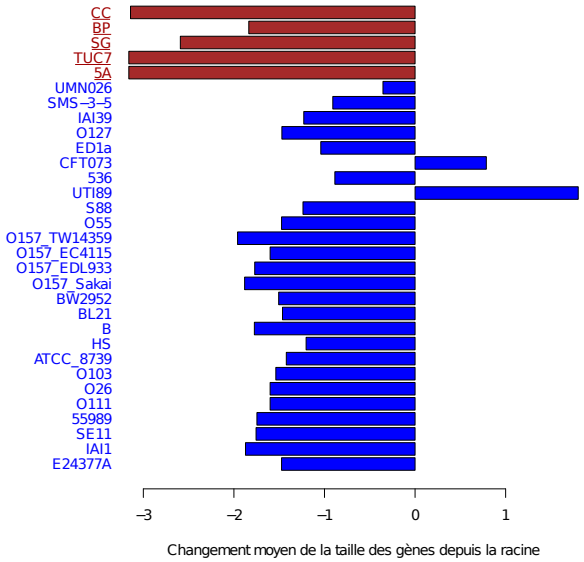


Figure IX.8 – Changement de la longueurs des gènes depuis la racine jusqu'aux branches terminales pour *Buchnera* et *E. coli*
Les couleurs des barres et des souches symbolisent les différentes espèces avec en bleu les souches *E. coli* et en rouge foncé les souches *Buchnera*. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit.
Les évènements d'insertions et délétions ont été déterminés avec *Prank* (Löytynoja et Goldman, 2005) sur 226 familles de gènes orthologues, après réconciliation des arbres des gènes et de l'arbre des espèces avec *Prunier* (Abby *et al.*, 2010) (Figure IX.5).

L'analyse avec un groupe externe est donc indispensable pour confirmer ces observations et résoudre le conflit sur les branches ancestrales aux souches d'*E. coli* et de *Buchnera*.

IX.2.2 *Prochlorococcus*

Les insertions le long de la phylogénie de *Prochlorococccus* et *Synechococcus* sont comparables en nombre aux délétions (Figure IX.9, $P = 0.7891$, test des rangs signés de Wilcoxon), mais sont en moyenne plus petites ($P = 1.975 \cdot 10^{-5}$, test des rangs signés de Wilcoxon unilatéral). La plupart des évènements sont intra-géniques (Figure IX.9, $P = 4.657 \cdot 10^{-10}$, test des rangs signés de Wilcoxon unilatéral) mais ces derniers sont en moyenne plus petits que les évènements aux extrémités des gènes ($P = 1.041 \cdot 10^{-7}$, test des rangs signés de Wilcoxon unilatéral). Ces résultats contrastent avec ceux obtenus pour *Buchnera* et *E. coli*. Les mécanismes impliqués pourraient donc être différents.

Le fort contraste de la taille des gènes observé entre MIT9303 et MIT9313 (Figure IX.2b) est dû à des insertions terminales longues. Les gènes ont subi des insertions aux extrémités des gènes de 102.95 bases en moyenne le long de la branche terminale conduisant à MIT9303 et 36.65 le long de la branche terminale conduisant à MIT9313, valeurs significativement différentes ($P = 1.845 \cdot 10^{-3}$, test de Mann-Whitney). Au contraire, pour

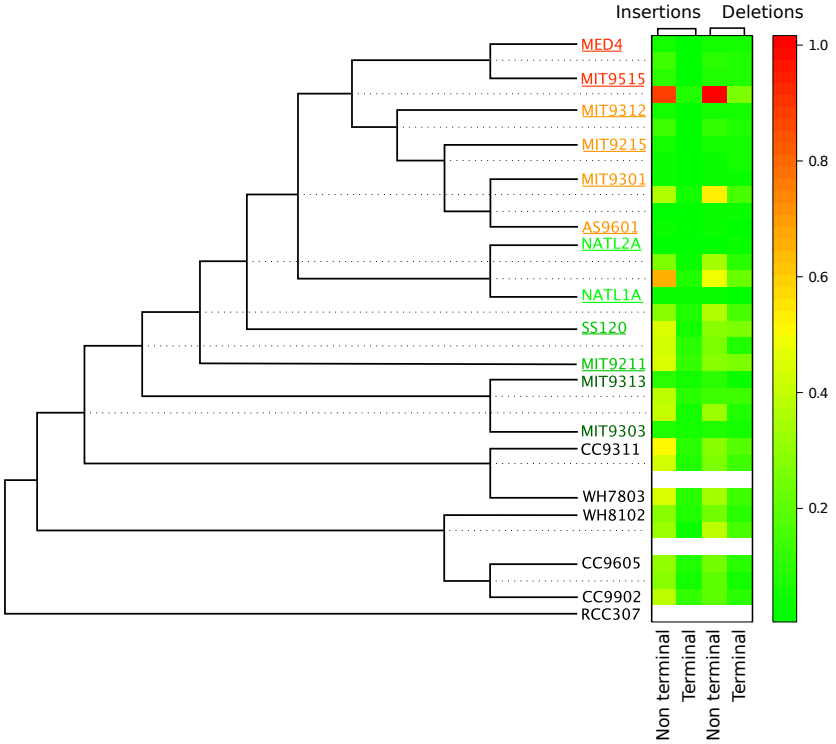


Figure IX.9 – Nombre d'insertion et délétion rapporté au nombre de familles par branche le long de l'arbre phylogénétique de *Prochlorococcus* et *Synechococcus*
Les évènements d'insertions et délétions ont été déterminés avec *Prank* (Löytynoja et Goldman, 2005) sur 693 familles de gènes orthologues, après réconciliation des arbres des gènes et de l'arbre des espèces avec *Prunier* (Abby *et al.*, 2010) (Figure IX.5). Les évènements dits terminaux correspondent aux évènements ayant lieu en début ou fin de gènes et les évènements non terminaux aux autres.
Les couleurs de noms des souches symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit. L'arbre phylogénétique est construit comme décrit dans la section C.4 (Annexe) mais les longueurs des branches sont arbitraires et ne reflètent pas les taux d'évolution.

les insertions intra-géniques, les évènements sont en moyenne de même taille le long des deux branches ($P = 0.1182$, test de Mann-Whitney). Ainsi, en plus de l'acquisition de nombreux gènes orphelins (Figure VIII.3), le génome de MIT9303 a subi une croissance de ces gènes, depuis sa divergence avec MIT9313. Cette croissance a eu lieu par l'acquisition de bases en fin de gènes alors que la taille des gènes de MIT9313 n'a pas changé (Figure IX.10). La souche MIT9303 semble ainsi en phase d'expansion de son génome, contrastant avec la réduction observée dans d'autres souches. Pourquoi ce génome est-il en expansion alors que celui de MIT9313 semble stable ? Les pressions de sélection ont-elles changé depuis la divergence entre MIT9313 et MIT9303 ? Des erreurs d'annotations des gènes chez MIT9303 pourraient aussi expliquer les différences observées.

Alors que de nombreuses pertes de gènes ont eu lieu peu après la divergence entre *Prochlo-*

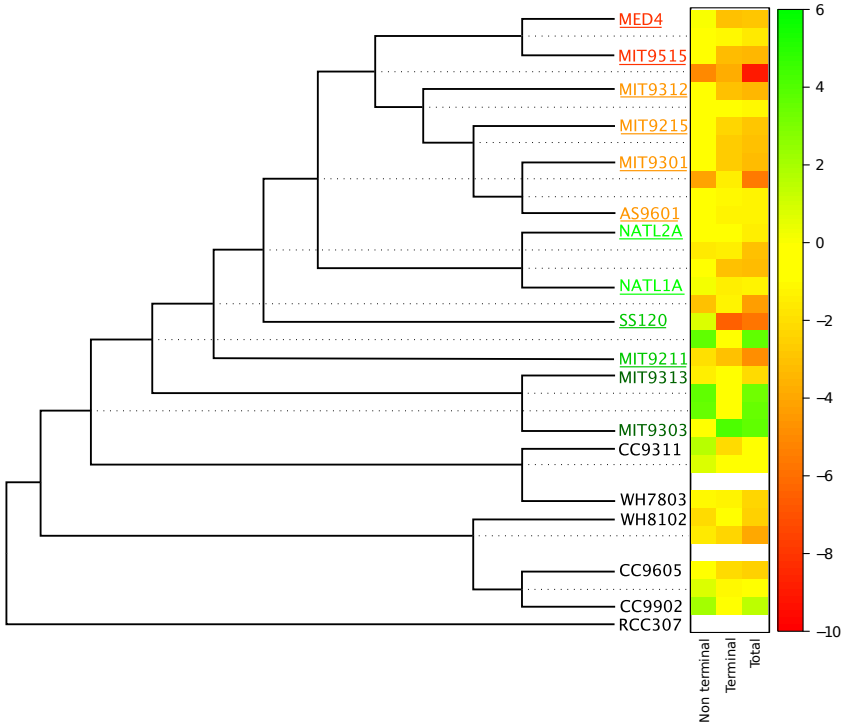
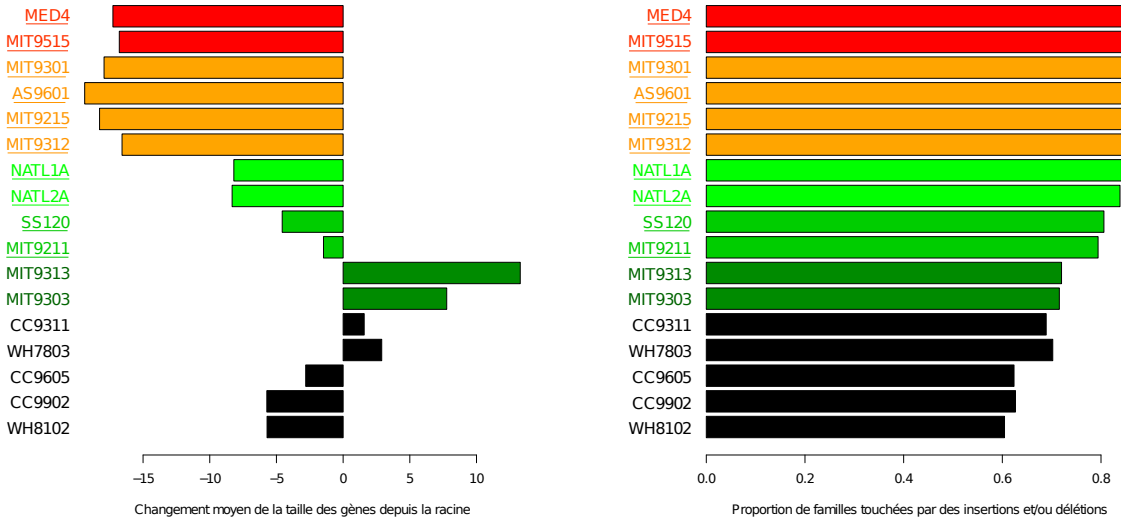


Figure IX.10 – Changement moyen de la longueur des gènes le long des branches de l'arbre phylogénétique de *Prochlorococcus* et *Synechococcus*
Les évènements d'insertions et délétions ont été déterminés avec *Prank* (Löytynoja et Goldman, 2005) sur 693 familles de gènes orthologues, après réconciliation des arbres des gènes et de l'arbre des espèces avec *Prunier* (Abby *et al.*, 2010) (Figure IX.5). Les évènements dits terminaux correspondent aux évènements ayant lieu en début ou fin de gènes et les évènements non terminaux aux autres.
Les couleurs de noms des souches symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit. L'arbre phylogénétique est construit comme décrit dans la section C.4 (Annexe) mais les longueurs des branches sont arbitraires et ne reflètent pas les taux d'évolution.

rococcus et *Synechococcus* (Figure VIII.3), les longueurs des gènes semblent augmenter dans la branche précédent la divergence du clade MIT9313-MIT9303 puis diminuer après, dans la branche ancestrale et dans toutes les branches conduisant aux souches réduites (Figure IX.10). Ainsi la taille des gènes a augmenté de la racine jusqu'aux souches de *Prochlorococcus* LLIV et a diminué de la racine jusqu'aux souches réduites de *Prochlorococcus* (Figure IX.11a), expliquant ainsi le fort contraste observé au noeud différenciant les souches réduites de *Prochlorococcus* et les souches non réduites de *Prochlorococcus* (Figure IX.2b).

En accord avec les différences observées de longueurs des gènes orthologues (Figure IX.4), le nombre d'évènements est plus important le long de la branche conduisant aux souches de *Prochlorococcus* HL (Figure IX.9). Ainsi, les gènes orthologues ont été réduits de 8 bases en moyenne dans la branche ancestrale aux souches HL (Figure IX.10). Depuis la



(a) Changement de la longueur des gènes

(b) Proportion de familles touchées par au moins un évènement d'insertion et/ou délétion

Figure IX.11 – Changement de la longueur des gènes et proportion de familles touchées depuis la racine jusqu'aux différentes souches de *Prochlorococcus* et de *Synechococcus*
Les couleurs symbolisent les différentes clades avec en noir les souches *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit.
Les évènements d'insertions et délétions ont été déterminés avec *Prank* (Löytynoja et Goldman, 2005) sur 693 familles de gènes orthologues, après réconciliation des arbres des gènes et de l'arbre des espèces avec *Prunier* (Abby *et al.*, 2010) (Figure IX.5).

racine jusqu'aux souches HL, les gènes ont perdu plus de 16 bases en moyenne (Figure IX.11a), des valeurs cinq fois supérieures à celles observées chez *Buchnera* ; de plus ces changements de taille des gènes touchent plus de 85% des familles (Figure IX.11b).

Contrairement aux endosymbiotes, l'évolution réductive chez *Prochlorococcus* touche donc les gènes à la fois dans leur nombre et dans leur taille, malgré les pressions pour le maintien de l'information.

IX.3 Discussion

Chez les endosymbiotes, l'évolution réductive a un impact sur les séquences des gènes conservés entraînant une divergence de la longueur des gènes entre les endosymbiotes et les bactéries libres. Cependant, contrairement à ce qui était attendu, les différences ne reflètent pas un biais mutationel vers la délétion mais une forte variabilité des longueurs liée à un grand nombre d'insertions et de délétions au sein des gènes, en accord avec les observations de Kenyon et Sabree (2014). Ces évènements ont principalement eu lieu

chez *Buchnera*, probablement en raison du relâchement des pressions de sélection sous l'impulsion du cliquet de Muller. Des délétions dans les branches ancestrales compensées par des délétions dans les branches d'*E. coli* peuvent expliquer le peu de différences de longueur des séquences lorsqu'elles sont étudiées globalement (Figure IX.3), mais ces observations doivent être confirmées en ajoutant un groupe externe.

Au sein des souches d'*E. coli*, les changements de longueur de gènes ont surtout lieu dans les parties terminales des gènes (Figure IX.7). Au contraire, chez *Buchnera*, les changements affectent à la fois les zones terminales et les zones intra-géniques (Figure IX.7), contrairement à ce qui est observé par Kenyon et Sabree (2014). Or, dans notre étude, les mêmes familles de gènes sont étudiées pour *E. coli* et *Buchnera*. Ces différences dans les zones affectées par les changements peuvent refléter des pressions de sélection différentes chez *Buchnera* et *E. coli*. Les pressions pourraient donc être moins fortes chez *Buchena* pour conserver les séquences, en dépit du fait que les séquences étudiées sont des séquences conservées. Le cliquet de Muller pourrait toucher toutes les séquences, même les plus conservées, à des timings potentiellement différents selon les pressions de sélection exercées sur les séquences.

Chez *Prochlorococcus*, l'évolution réductive s'initie par la perte de gènes peu après la divergence entre *Prochlorococcus* et *Synechococcus* (Figure VIII.3). Cependant, le long de la branche différenciant *Prochlorococcus* et *Synechococcus*, les gènes augmentent en taille tout comme le long des branches du clade de *Prochlorococcus* LLIV (Figure IX.10). Au contraire, le long des branches conduisant aux souches réduites, une réduction des gènes orthologues par délétion a eu lieu (Figure IX.10) entraînant ainsi une réduction de la taille des gènes, supérieure aux réductions observées pour certaines souches de *Buchnera*. Ainsi, un évènement, peut-être la perte de certains gènes de réparation, après la divergence des souches de *Prochlorococcus* LLIV, semble avoir enclenché, en même temps que l'enrichissement en AT, une réduction de la taille des gènes orthologues. Cette réduction de la taille des gènes vient donc contredire les conclusions des analyses précédemment effectuées sur la taille des gènes chez *Prochlorococcus* (Sun et Blanchard, 2014; Marais *et al.*, 2008).

Les codons stop étant riches en bases AT, la densité en codons stop pourrait être plus élevée dans les organismes riches en AT comme les souches réduites de *Prochlorococcus* que dans les organismes riches en bases GC. Même si la plupart des évènements observés n'ont pas lieu en fin de gènes (Figure IX.9), les évènements aux extrémités des gènes expliquent une part importante des réductions de longueur des gènes au sein des branches terminales des souches réduites de *Prochlorococcus* (Figure IX.10). Ainsi, l'enrichissement en bases AT et l'augmentation de la densité de codons stop pourraient expliquer une partie de la réduction de la longueur des gènes.

Cependant, dans la branche ancestrale aux souches de *Prochlorococcus* HL, là où les réductions sont les plus importantes (Figure IX.10), la réduction a principalement lieu par de petits évènements non terminaux. Le long de cette branche, les répertoires géniques (Figure VIII.3) changent avec l'acquisition de gènes, des pertes de gènes et des réductions de la taille d'une large gamme de gènes orthologues (Figure IX.11b). Le changement d'environnement qu'ont subi ces souches semble accélérer la réduction des génomes. Cette

réduction peut être due soit à une pression issue de l'environnement pauvre en nutriments pour une économie des ressources, soit à un relâchement des pressions de sélection faisant ainsi ressortir les biais mutationnels vers les délétions, potentiellement favorisés par la perte de gènes de réparation de l'ADN le long de cette branche (Figure VIII.6).

Chapitre X

Contenu en bases GC, usage des codons, ARNt et codons optimaux

Les endosymbiotes et *Prochlorococcus* ont des contenus en bases GC extrêmement faibles avec 26% pour *Buchnera* et 30.8-38% pour les souches réduites de *Prochlorococcus*, contrairement à *E. coli* et aux souches non réduites de *Prochlorococcus*. Comme la composition moyenne en acides aminés des protéines est corrélée au contenu en bases GC (Sueoka, 1961), le contenu des protéines a dû changer au cours de l'évolution réductive et de l'appauvrissement en bases GC. Chez *Buchnera*, le biais de composition nucléotidique a eu un impact négatif fort sur la stabilité des protéines, un handicap en partie compensé par une surexpression des protéines chaperonnes comme GroEL (van Ham *et al.*, 2003). Au contraire, pour les souches réduites de *Prochlorococcus*, les changements majeurs dans la constitution des protéines sont liées à une optimisation balancée entre la stabilité protéique et la flexibilité (Paul *et al.*, 2010). Le biais de composition nucléotidique semble donc moins nocif chez *Prochlorococcus* que chez les endosymbiotes.

Les changements de GC génomique peuvent aussi impacter les codons utilisés pour coder les acides aminés. En effet, le code génétique qui fait le lien entre les codons et les acides aminés est dégénéré. Avec 61 codons codant pour des acides aminés et seulement 20 acides aminés, plusieurs codons codent pour le même acide aminé. Ces codons dits synonymes sont toujours très peu différents les uns des autres, avec rarement plus d'un changement de nucléotide, généralement en troisième base des codons.

Pour de nombreux organismes, certains codons synonymes sont utilisés à des fréquences supérieures à celles attendues par hasard. Par exemple, chez *Synechococcus* RCC307, sur les six codons codant pour l'arginine, trois seulement (CGC, CGG et CGT) totalisent plus de 82% des codons utilisés. Ce phénomène d'usage différencié des codons synonymes, appelé biais d'usage des codons, se retrouve dans pratiquement toutes les espèces bactériennes, mais les codons synonymes préférés peuvent être différents selon les espèces. Par exemple, la fréquence des deux codons de la lysine est différente au sein des clades de *Prochlorococcus* et *Synechococcus* (Tableau X.1).

	AAA	AAG
Synechococcus CC9605	36.5%	63.5%
Prochlorococcus MIT9313	48.3%	51.7%
Prochlorococcus MIT9211	68.6%	31.4%
Prochlorococcus NATL2A	76.7%	23.3%
Prochlorococcus MIT9215	81.9%	18.1%
Prochlorococcus MED4	81.2%	18.8%

Table X.1 – Usage des codons synonymes de la lysine chez *Prochlorococcus* et *Synechococcus*
Les couleurs symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Une seule souche par clade est représentée, les autres souches ayant des tendances similaires.

Un biais d'usage des codons, propre à chaque organisme, pourrait s'expliquer par la composition globale des génomes. Dans le cas de *Synechococcus* et de *Prochlorococcus*, l'utilisation différenciée des codons synonymes de la lysine semble refléter une différence de contenu en bases GC. En effet, les génomes riches en bases AT (*Prochlorococcus* MIT9211, NATL2A, MIT9215 et MED4) utilisent préférentiellement le codon finissant par A au contraire des autres génomes (Tableau X.1). Chez *Buchnera*, l'usage des codons a été perturbé par un fort biais mutationnel vers AT à l'échelle du génome (Charles *et al.*, 2006). Est-ce aussi le cas pour les souches de *Prochlorococcus* qui ont subi un enrichissement en bases AT ?

Ajouté à la composition en bases GC, un autre biais de composition peut influencer le biais d'usage des codons : la composition des deux brins d'ADN. En l'absence de biais mutationnel, la fréquence d'emploi de la base A doit être égale à celle de T, idem pour C et G (Sueoka, 1995), au sein de chacun des brins. Cependant, un enrichissement en bases G et T sur le brin précoce est observé (Kano-Sueoka *et al.*, 1999; Rocha *et al.*, 1999; McLean *et al.*, 1998; Sueoka, 1995). Dans les génomes bactériens caractérisés par une asymétrie marquée entre les brins pour l'usage des codons (Das *et al.*, 2005, 2006; Lafay *et al.*, 1999; McInerney, 1998), la sélection réplicationnelle et la sélection transcriptionnelle jouent souvent un rôle majeur dans ce phénomène. Les brins précoces de ces organismes, répliqués et transcrits plus rapidement que les brins retardés, contiennent généralement un nombre plus élevé de gènes à cause de la sélection réplicationnelle et sont aussi enrichis en gènes fortement exprimés comme un effet de la sélection transcriptionnelle. En effet, lors de la réplication, l'ADN polymérase synthétise l'ADN dans le sens 5'-3'. Lors de la transcription d'un gène sur le brin retardé, l'ARN polymérase se déplace dans le sens 3'-5' et peut donc entrer en collision avec l'ADN polymérase, interrompant ainsi la transcription et ralentissant la réplication (French, 1992). Conserver la plupart des gènes sur le brin précoce permet d'éviter ces chocs et procure ainsi un avantage sélectif par une réplication rapide (sélection réplicationnelle). La transcription des gènes sur le brin retardé est aussi fortement impactée par le déplacement de l'ADN polymérase. Cependant, pour des gènes faiblement exprimés, l'interruption de la transcription par le choc avec l'ADN polymérase est moins délétère que pour des gènes fortement exprimés. Ces derniers, transcrits plus régulièrement, ont ainsi un avantage à être sur le brin précoce pour une transcription

optimale (McInerney, 1998) (sélection transcriptionnelle). Les forces réplicationnelles et transcriptionnelles peuvent donc être responsables de l'orientation des gènes et influencer indirectement l'usage des codons (codons potentiellement enrichis en bases GT pour les gènes fortement exprimés présents sur le brin précoce). Paul *et al.* (2010) ont observé une asymétrie d'usage des codons entre le brin précoce et le brin retardé pour les souches de *Prochlorococcus* LL mais pas pour les souches de *Prochlorococcus* HL. Cependant, pour les souches de *Prochlorococcus* LL, les gènes sont équitablement répartis entre les deux brins. Pour les souches MIT9303 et MIT9313, le brin précoce est appauvri en gènes codant pour les protéines ribosomales, considérés comme des gènes fortement exprimés. Pour les autres souches de *Prochlorococcus* LL, les gènes fortement exprimés sont présents en plus grand nombre dans le brin précoce. Il est ainsi difficile de conclure quant à l'implication de la sélection transcriptionnelle dans l'usage des codons des souches LL réduites, mais elle ne semble pas pouvoir expliquer le biais d'usage des codons pour les souches HL.

Les autres hypothèses sur le biais d'usage des codons sont principalement basées sur l'influence du choix des codons dans le processus de traduction. En effet, les détails du processus d'insertion d'un acide aminé dans le peptide ne sont pas forcément équivalents pour tous les codons synonymes. Chez les procaryotes, le contenu en ARN_t est biaisé de telle sorte que les ARN_t majoritaires puissent s'apparier avec les codons les plus fréquents (Gouy et Gautier, 1982; Gouy et Grantham, 1980; Dong *et al.*, 1996; Kanaya *et al.*, 1999). L'utilisation de ces codons majeurs permettrait une traduction rapide et efficace. La sélection traductionnelle résultante est alors supposée plus forte dans les gènes fortement exprimés qui doivent être traduits rapidement. Le facteur principal de la variabilité intra-espèces, c'est-à-dire entre les gènes d'un organisme, de l'usage des codons chez les bactéries serait donc la sélection pour l'optimisation de la traduction (Ikemura, 1981; Gouy et Gautier, 1982; Kanaya *et al.*, 1999; Gautier, 2000).

Le biais d'usage des codons chez les bactéries est ainsi soumis à une balance entre des biais mutationnels et des formes variées de la sélection naturelle. Pour *Synechococcus*, la sélection traductionnelle semble être à l'origine du biais d'usage des codons, alors que chez *Prochlorococcus*, le facteur principal semble être le biais de composition GC (Yu *et al.*, 2012). Ce dernier ne serait pas le résultat d'une dérive génétique continue mais plutôt d'une diversification de niche vers une plus grande stabilité et fidélité des protéines dans des environnements divers selon Paul *et al.* (2010).

Cependant, les deux analyses effectuées sur le biais d'usage des codons chez *Prochlorococcus* (Paul *et al.*, 2010; Yu *et al.*, 2012) présentent des problèmes méthodologiques et ne vont pas assez loin dans l'analyse des biais. Paul *et al.* (2010) utilisent une analyse des correspondances sur des fréquences relatives de codons, introduisant ainsi des biais et des erreurs possibles d'interprétations (Perrière et Thioulouse, 2002) (décrites plus en détail dans la section X.3). Dans l'analyse de Yu *et al.* (2012), différents indicateurs du biais d'usage des codons à l'échelle des génomes sont comparés entre les souches. Cependant, cette analyse ne s'intéresse pas aux différences d'usage des codons au sein des génomes, pourtant nécessaires pour comprendre les causes du biais d'usage des codons des différents génomes. De plus, cette analyse se concentre sur la comparaison entre *Prochlorococcus* et *Synechococcus*, sans réelle distinction entre les différents clades de *Prochlorococcus* et les

différentes étapes de l'évolution réductive. L'analyse de Paul *et al.* (2010) tente d'identifier les déterminants moléculaires associés au partitionnement vertical des niches et se concentre ainsi seulement sur les souches de *Prochlorococcus*. Or, la perte de gènes et l'évolution réductive se sont initiées à la divergence entre *Synechococcus* et *Prochlorococcus* (Figure VIII.3). Elle s'est arrêtée dans les souches de *Prochlorococcus* LLIV mais a continué dans les autres. Il serait donc intéressant d'explorer plus en détail l'évolution de l'usage des codons chez *Synechococcus* et *Prochlorococcus* en mettant en relation avec les changements de composition des génomes et de répertoires de gènes ARN_t, afin d'identifier les forces évolutives en action et de mieux comprendre qui de la sélection ou de la dérive dirige principalement la réduction des génomes chez *Prochlorococcus*.

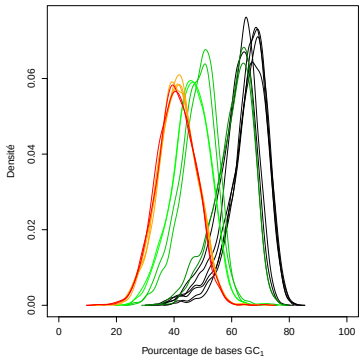
X.1 Biais de composition

Les biais de composition des génomes peuvent expliquer les différences de biais d'usage des codons entre les espèces. Chez *Prochlorococcus*, le principal biais identifié est l'enrichissement en bases AT au cours de l'évolution réductive.

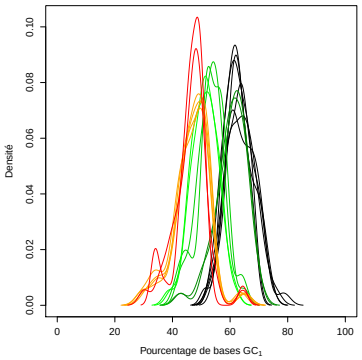
L'enrichissement en bases AT touche l'ADN intergénique mais aussi les séquences des gènes, principalement les troisièmes bases des codons (Figure X.1e). Étant donné la dégénérescence du code génétique, de nombreux changements en troisième base des codons sont synonymes et sont donc considérés comme quasi-neutres. Ces bases sont ainsi, au sein des gènes, les premières touchées par un biais mutationnel, comme le biais vers un enrichissement en bases AT.

Certains changements au niveau des premières bases des codons sont synonymes comme pour certains codons de la leucine et l'arginine, alors que tout changement en seconde base des codons entraîne une modification d'acide aminé. L'enrichissement en bases AT touche donc aussi les premières bases (Figure X.1a) mais dans une proportion moindre par rapport aux changements observés pour les troisièmes bases (Figure X.1e). En revanche, il ne change que marginalement la composition des secondes bases des codons (Figure X.1c). Ces observations sont exacerbées pour les gènes fortement exprimés, comme les gènes codant pour les protéines ribosomales (Figures X.1b, X.1d). En effet, dans les gènes fortement exprimés, les pressions de sélection sont fortes pour le maintien des séquences et les biais mutationnels sont contrés. Ainsi, chez *Prochlorococcus* MED4, le pourcentage des bases GC des gènes corrèle avec le niveau d'expression (Figure X.2), les gènes fortement exprimés ayant des taux de GC plus élevés, proches de ceux des souches non réduites de *Prochlorococcus*.

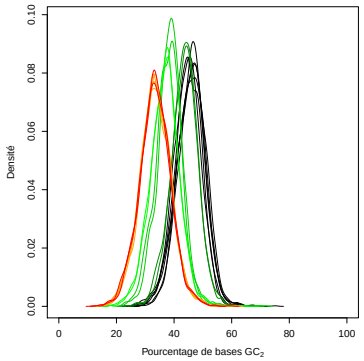
Pour les différentes mesures de la composition en bases GC, l'enrichissement en bases AT suit une sorte de gradient le long de la phylogénie correspondant à l'évolution réductive (Figure X.1). Ainsi, les souches de *Synechococcus* ont les taux de GC les plus élevés. Ces taux se réduisent avec la divergence *Synechococcus*-*Prochlorococcus* : les souches non réduites de *Prochlorococcus* ont des taux de GC inférieurs à ceux de *Synechococcus* mais



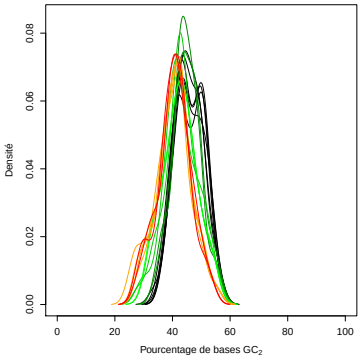
(a) GC_1 pour tous les gènes



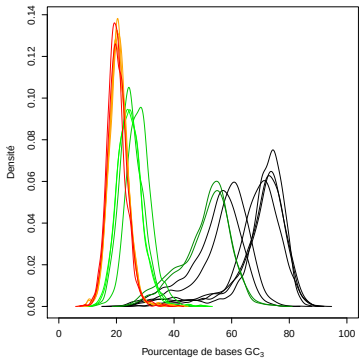
(b) GC_1 pour les gènes ribosomaux



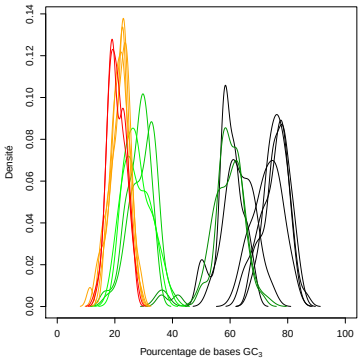
(c) GC_2 pour tous les gènes



(d) GC_2 pour les gènes ribosomaux



(e) GC_3 pour tous les gènes



(f) GC_3 pour les gènes ribosomaux

Figure X.1 – Densité des pourcentages de bases GC aux trois positions des codons chez *Prochlorococcus* et *Synechococcus*
Les pourcentages de bases GC ont été calculés avec des scripts Biopython, sur toutes les familles de gènes des souches de *Prochlorococcus* et de *Synechococcus*. GC_1 correspond au pourcentage de bases GC en première position des codons, GC_2 au pourcentage en seconde position des codons, GC_3 au pourcentage en troisième position des codons.
Les couleurs symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI.

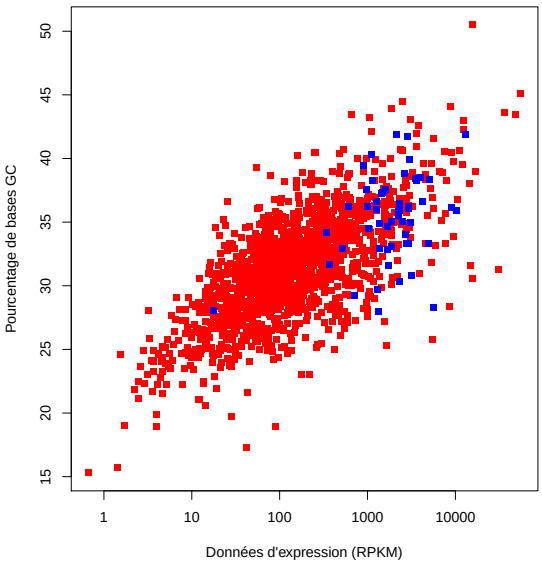


Figure X.2 – Pourcentage de bases GC en fonction des données d'expression chez *Prochlorococcus* MED4
Les données d'expression sont issues des données de Wang *et al.* (2014). Le pourcentage de bases GC est calculé avec des scripts Biopython, sur tous les gènes de *Prochlorococcus* MED4.
Les points en bleu correspondent aux gènes codant pour les protéines ribosomales.

supérieurs à ceux des autres souches de *Prochlorococcus*. La réduction des contenu en GC suit ensuite la divergence des souches réduites de *Prochlorococcus* : LLII/LIII puis LLI puis HL (Figure X.1). Ainsi, l'évolution réductive s'accompagne tout le long, depuis la divergence entre *Prochlorococcus* et *Synechococcus*, d'un enrichissement en bases AT. Comment cet enrichissement impacte-t-il les biais d'usage des codons ? Le biais mutationnel à l'origine de l'enrichissement est-il lié à un relâchement de toutes les pressions de sélection ou alors est-il favorisé car les nucléotides A et T sont moins coûteux à produire ? La pression traductionnelle semble toujours présente car les gènes fortement exprimés sont sujets au biais mutationnel, mais dans une moindre mesure.

Selon la théorie développée par Sueoka (1962, 1988), la présence d'un biais de composition en GC dans les génomes entraîne des changements plus importants dans les parties neutres que dans les parties fonctionnelles. Le contenu en GC des séquences est alors soumis à un équilibre entre les pressions mutationnelles et les contraintes sélectives. Il est alors possible d'estimer le niveau de sélection subi par les séquences en comparant leur contenu en GC aux positions neutres (GC_3) avec le contenu en GC aux positions non neutres (GC_{12}) à l'aide d'un "neutrality plot" (Figure X.3). Si les corrélations sont significatives et la pente de la régression est proche de 1, les gènes subissent un impact similaire aux différentes positions des codons et l'usage des codons est principalement causé par les pressions mutationnelles (Sueoka, 1988). Au contraire, si la sélection traductionnelle est le facteur dominant, la sélection agit contre le biais mutationnel. Ce dernier touche alors principalement les troisièmes bases des codons et peu les deux premières bases. La corrélation entre entre GC_{12} et GC_3 est ainsi limitée et la pente de la droite de régression

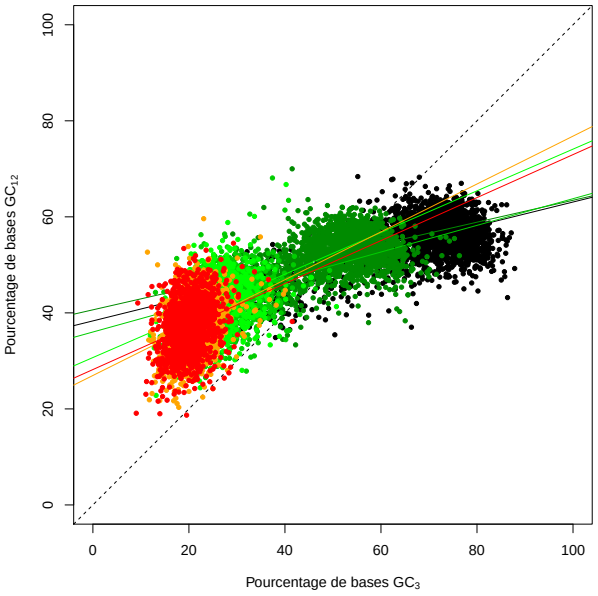
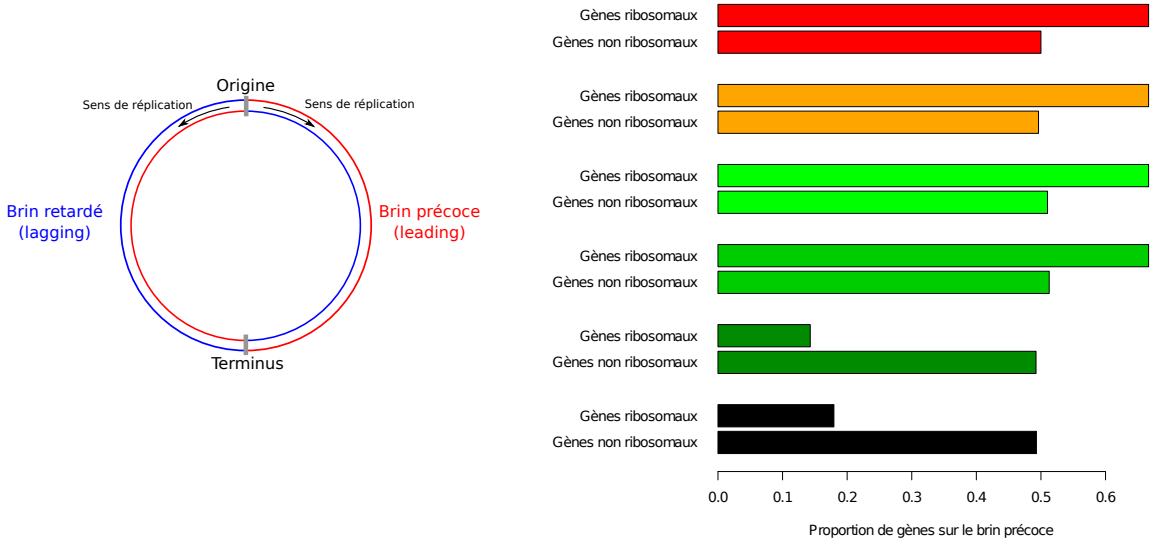


Figure X.3 – "Neutrality plot" (GC_{12} en fonction de GC_3) chez *Prochlorococcus* et *Synechococcus* Les couleurs symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Une seule souche par clade est représentée, les autres souches ayant des tendances similaires.
La droite en pointillés représente la droite $y = x$. Les droites en trait plein sont les droites de régression pour les différentes souches : en noir, *Synechococcus* CC9605, $y = 0.25x + 38.33$, $r = 0.52$ ($p < 0.01$) ; en vert foncé, *Prochlorococcus* MIT9313, $y = 0.23x + 10.64$, $r = 0.42$ ($p < 0.01$) ; en vert, *Prochlorococcus* MIT9211, $y = 0.28x + 36.01$, $r = 0.27$ ($p < 0.01$) ; en vert clair, *Prochlorococcus* NATL2A, $y = 0.43x + 28.19$, $r = 0.37$ ($p < 0.01$) ; en orange, *Prochlorococcus* MIT9215, $y = 0.50x + 26.96$, $r = 0.32$ ($p < 0.01$) ; en rouge, *Prochlorococcus* MED4, $y = 0.45x + 38.33$, $r = 0.27$ ($p < 0.01$).

est proche de 0 (Kawabe et Miyashita, 2003).

Le "neutrality plot" exploré par Yu *et al.* (2012) avait montré que chez *Prochlorococcus*, en particulier les souches réduites, la pression mutationnelle de composition génomique est plus forte que la sélection traductionnelle, au contraire des souches de *Synechococcus* et non réduites de *Prochlorococcus*. Cependant, dans cette étude, les comparaisons entre les différents groupes ne sont pas faites avec le même nombre de souches (10 souches pour *Synechococcus*, 2 souches non réduites de *Prochlorococcus* et 10 souches réduites de *Prochlorococcus*) et les régressions sont effectuées sur l'ensemble des gènes des souches de ces groupes sans distinction entre les souches. Or, les souches réduites de *Prochlorococcus* LLII/LLIII semblent avoir un motif différent de celui des autres souches réduites de *Prochlorococcus* (Figure X.3), plus proche de celui des souches non réduites de *Prochlorococcus* et de *Synechococcus*, c'est-à-dire une sélection traductionnelle supérieure à la pression mutationnelle. Ces souches réduites sont les plus proches phylogénétiquement des souches non réduites, signifiant que le passage d'une domination de la sélection traductionnelle à un équilibre entre la sélection et la pression mutationnelle a eu lieu après



(a) Explication des brins précoces et retardés chez les bactéries

(b) Proportion de gènes ribosomaux et non ribosomaux au sein du brin précoce

Figure X.4 – Répartition des gènes entre les brins précoces et retardés chez *Prochlorococcus* et *Synechococcus*
Pour la Figure X.4b, les couleurs symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Une seule souche par clade est représentée, les autres souches ayant des tendances similaires.

l'initialisation de l'évolution réductive et de l'enrichissement en bases AT. Les droites de régression ont toutes des pentes inférieures à 0.5 (Figure X.3). La pression mutationnelle, bien qu'ayant un impact plus fort dans les souches réduites de *Prochlorococcus*, ne domine donc pas la sélection traductionnelle, contrairement à ce qui est observé par Yu *et al.* (2012).

L'autre biais de composition des génomes est celui qui différencie la composition des deux brins d'ADN. Dans les génomes bactériens, le brin précoce, c'est-à-dire le brin orienté de 5' vers 3' dans la direction de réplication, est enrichi en bases G et T (Kano-Sueoka *et al.*, 1999; Rocha *et al.*, 1999; McLean *et al.*, 1998; Sueoka, 1995). Cet enrichissement est suffisamment fort et cohérent tout le long du chromosome pour permettre l'identification de l'origine de réplication (Lobry, 1996). En effet, la valeur d'un indicateur comme $GC\_skew = {}^{(G-C)}/_{(G+C)}$ avec G et C la fréquence d'utilisation des bases G et C respectivement, subit un changement de signe au niveau de l'origine de réplication. En traversant l'origine de réplication et en continuant à lire le chromosome dans le même sens, le brin précoce devient le brin tardif et inversement (Figure X.4a) et donc l'excès de GT d'un côté devient un déficit de l'autre. En utilisant ce principe (Frank et Lobry, 2000), les origines de réplication des souches de *Prochlorococcus* et de *Synechococcus* ont été estimées, l'asymétrie de composition des deux brins ayant été conservée avec l'évolution réductive.

Pour toutes les souches étudiées ici, les gènes non ribosomaux sont équirépartis entre

les deux brins (Figure X.4b). La sélection réplicationnelle, liée à une surreprésentation des gènes sur le brin précoce pour faciliter la réplication, est donc peu probable chez *Prochlorococcus* et *Synechococcus*. La tendance est différente pour les gènes ribosomaux, fortement exprimés. De façon surprenante, ils sont sous-représentés sur le brin précoce pour les souches de *Synechococcus* et de *Prochlorococcus* LLIV mais surreprésentés sur le brin précoce pour les souches réduites de *Prochlorococcus*, conformément aux observations de Paul *et al.* (2010). Pourrait-il y avoir une sélection transcriptionnelle pour les gènes fortement exprimés dans les souches réduites de *Prochlorococcus* mais pas chez les souches non réduites avec les gènes fortement exprimés sur le brin précoce ? Cependant, pour *Prochlorococcus* MED4, les gènes fortement exprimés sont équirépartis entre les deux brins, signifiant que la sélection transcriptionnelle est peu probable pour cette souche.

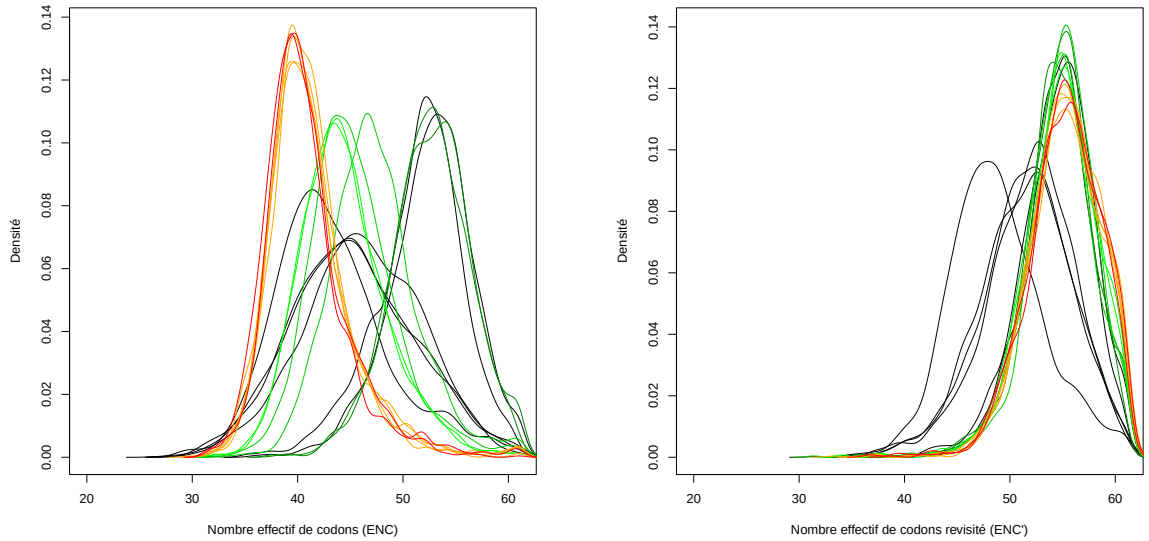
X.2 Nombre effectif de codons

Les méthodes d'estimation du biais d'usage des codons à l'échelle des gènes ou du génome sont nombreuses. La plupart permettent d'obtenir un simple indicateur chiffré afin de comparer les gènes ou les génomes entre eux et estimer les facteurs de sélection.

Basé sur l'idée de comparer l'usage des codons à un usage non biaisé, le nombre effectif de codons (ENC) (Wright, 1990) calcule combien de codons sont "réellement" utilisés par un gène, en donnant un poids à chaque codon en fonction de sa fréquence d'emploi par rapport à ses synonymes. Le défaut principal de cet indicateur est sa variation avec le contenu en bases GC de la séquence. Une version normalisée par rapport au contenu en bases GC a été développée : ENC' (Novembre, 2002). Cet indicateur est moins sensible que d'autres indicateurs comme CAI[1] aux petites longueurs de gènes et permet d'avoir une mesure incluant les biais de composition du génome. Nous utilisons cet indicateur pour comparer les génomes de *Synechococcus* et de *Prochlorococcus* et chercher une éventuelle sélection traductionnelle.

Plus l'ENC (ou ENC') est faible, moins de codons sont utilisés au sein de la séquence et plus l'usage des codons est biaisé. Ainsi, d'après les distributions d'ENC des gènes des souches d'intérêt (Figure X.5a), il semble y avoir une sorte de gradient d'usage des codons avec un usage très biaisé pour les souches réduites de *Prochlorococcus* HL, suivi des souches réduites de *Prochlorococcus* LL et de *Synechococcus* puis des souches non réduites de *Prochlorococcus*. Cependant, ce gradient est similaire à celui observé pour les biais de composition en GC (Figure X.1). En effet, avec les distributions d'ENC' des gènes (Figure X.5b), toutes les souches de *Prochlorococcus* semblent avoir des biais d'usage similaires, faibles, alors que les souches de *Synechococcus* ont un biais d'usage des codons plus important. Cependant, deux souches de *Synechococcus* (CC9902 et CC9311) ont des biais similaires à ceux de *Prochlorococcus*. Ces souches présentent aussi des taux de bases

[1]CAI ou *Codon Adaptation Index* (Sharp et Li, 1987) utilise un ensemble de gènes fortement exprimés prédéfini, pour mesurer le biais d'usage des codons en comparant directement les fréquences des codons d'un groupe de gènes fortement exprimés par rapport aux fréquences des codons des autres gènes.



(a) ENC **(b)** ENC'

Figure X.5 – Densité de nombre effectif de codons (ENC et ENC') chez *Prochlorococcus* et *Synechococcus*
Le nombre effectif de codons (ENC) correspond au nombre de codons utilisés dans une séquence (Wright, 1990). Il prend ainsi la valeur de 61 quand tous les codons sont utilisés à fréquence égale et il décroit quand l'usage des codons devient moins uniforme. La valeur brute ne prend pas en compte des biais nucléotidiques, au contraire de la valeur revisitée, ENC' (Novembre, 2002).
Les valeurs d'ENC et ENC' ont été calculées sur les tous gènes des souches, à l'aide de l'outil ENCprime
Les couleurs symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI.

GC similaires à ceux des souches de *Prochlorococcus* LLIV (Figure X.1). Il est difficile de comprendre pourquoi ces deux souches sont différentes des autres sachant qu'elles ne sont pas monophylétiques et ne sont pas les souches les plus proches de *Prochlorococcus*. Les changements de biais d'usage des codons pour ces souches pourraient correspondre à une convergence évolutive avec *Prochlorococcus*, dont les causes restent mystérieuses.

Le biais d'usage des codons semble ainsi similaire et relativement faible pour toutes les souches de *Prochlorococcus*, la sélection traductionnelle serait donc plus faible que pour *Synechococcus*. Ainsi, pour *Prochlorococcus* MED4, ENC' et niveau d'expression ne sont pas significativement corrélés malgré une légère décroissance de l'ENC' avec l'augmentation des niveaux d'expression (Figure X.6). De plus, les gènes ribosomaux n'ont pas des valeurs d'ENC' différentes de celle des autres gènes. Ainsi les gènes fortement exprimés n'utilisent pas moins de codons synonymes que les autres gènes. Le biais d'usage des codons observé pour ces gènes ne serait donc pas dû à une sélection pour une traduction plus rapide. Ces observations semblent en contradiction avec l'idée d'une forte corrélation entre le biais d'usage des codons d'un gène et son niveau d'expression (Ikemura, 1985; Gouy et Gautier, 1982), attendu dans les cas de sélection traductionnelle.

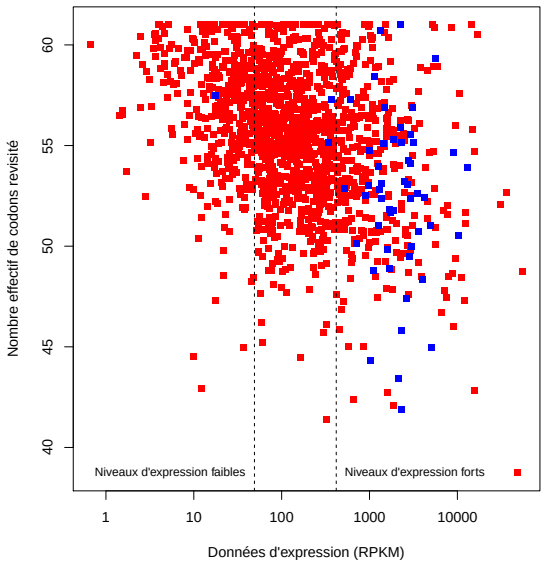
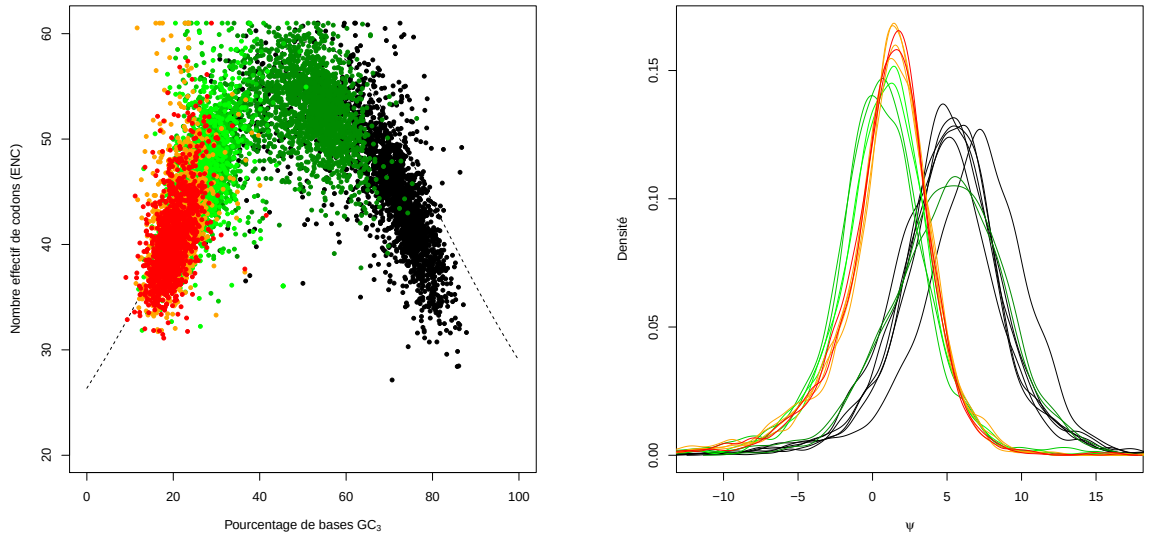


Figure X.6 – ENC' en fonction des données d'expression chez *Prochlorococcus* MED4
Les données d'expression sont issues des données de Wang *et al.* (2014). Les valeurs d'ENC' ont été calculées à l'aide de l'outil ENCprime sur tous les gènes de *Prochlorococcus* MED4.
Les points en bleu correspondent aux gènes codant pour les protéines ribosomales.

Pour déterminer si le biais d'usage des codons chez *Prochlorococcus* et *Synechococcus* est principalement dû à la différence de composition du génome, comme le suggère la différence entre ENC et ENC', nous utilisons le "Nc plot", c'est-à-dire ENC en fonction du GC_3 (Wright, 1990), que nous comparons à la courbe attendue entre GC_3 et ENC en l'absence de sélection (Figure X.7). Pour les gènes proches de la courbe, le biais d'usage des codons n'est ainsi pas dû à une sélection traductionnelle mais aux contraintes de compositions GC ou à une sélection pour des codons terminant par G/C ou A/T. Alors que pour les gènes en dessous de la courbe, le biais d'usage des codons est principalement dirigé par à une sélection des codons optimaux pour la traduction, comme c'est le cas pour les souches de *Synechococcus* et de *Prochlorococcus* LLIV non réduite (Figure X.7). Afin de quantifier la sélection sur les codons optimaux, nous étudions la différence ψ entre les valeurs attendues en l'absence de sélection (f_1) et les valeurs observées d'ENC (Figure X.7b). Les distributions de ψ sont centrées autour de 0 pour les souches réduites de *Prochlorococcus* et autour de valeurs positives pour les souches non réduites de *Prochlorococcus* et de *Synechococcus*. Les gènes des souches réduites de *Prochlorococcus* ont donc des valeurs d'ENC attendues en l'absence de sélection traductionnelle. Pour les souches non réduites de *Prochlorococcus* et de *Synechococcus*, les valeurs observées d'ENC sont inférieures à celles attendues en l'absence de sélection traductionnelle. Comme la sélection traductionnelle semble inefficace pour les souches réduites de *Prochlorococcus*, le biais d'usage des codons dans ces souches est probablement principalement impacté par la contrainte de composition des génomes, contrairement aux souches non réduites de *Prochlorococcus* et de *Synechococcus*.



(a) ENC en fonction de GC_3 **(b)** Densité de ψ

Figure X.7 – "Nc plot" (ENC en fonction de GC_3) et densité de la quantité de sélection traductionnelle chez *Prochlorococcus* et *Synechococcus*
La courbe en pointillés de la Figure X.7a correspond à $f_1(x) = -6+s+\frac{34}{s^2+(1.025-s)^2}$ avec $s = {}^{GC_3}/_{100}$, relation théorique entre ENC et GC_3 en l'absence de sélection traductionnelle (Wright, 1990) avec les coefficients revisités par Reis *et al.* (2004). Dans la Figure X.7b, $\psi = f_1(x) - ENC$ correspond à la quantité de sélection agissant sur l'usage des codons.
Les couleurs symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Une seule souche par clade est représentée dans la Figure X.7a, les autres souches ayant des tendances similaires.

X.3 Analyses inter- et intra-acides aminés de l'usage des codons

En résumant le biais d'usage des codons à une valeur unique, les indicateurs précédents entraînent des pertes d'information et sont limités pour l'identification des causes des biais d'usage des codons, car ils ne reflètent que partiellement la distribution des fréquences des codons. L'analyse factorielle des correspondances (AFC) de l'usage des codons permet d'évaluer le biais d'usage des codons à l'intérieur des génomes en conservant le maximum d'informations. En effet, l'AFC est une méthode multivariée dont le but est de résumer les structures de données dans un espace à grande dimension par projection dans des sous-espaces de faible dimension en perdant aussi peu d'information que possible, c'est-à-dire dans des espaces où les axes expliquent le maximum de variabilité des données. Pour l'analyse de l'usage des codons dans un génome, chaque gène est considéré comme un ensemble de 61 valeurs, les fréquences des codons qu'il contient. Ainsi, un gène correspond à un point dans un espace à 61 dimensions. Les gènes sont plus ou moins proches dans cet espace selon leur usage des codons. L'AFC dans les études d'usage des codons a été

utilisée pour un certain nombre d'espèces bactériennes (Médigue *et al.*, 1991; Perrière et Thioulouse, 2002). Cependant, l'AFC sur des tables contenant les effectifs des codons peut masquer les effets directement liés aux préférences des codons car la composition en acides aminés n'est pas prise en compte. C'est pourquoi de nombreuses analyses de l'usage des codons basées sur l'AFC ont utilisé des fréquences relatives des codons synonymes comme RSCU[1] plutôt que les données brutes de comptage des codons, par exemple McInerney (1998, 1997); Lafay *et al.* (1999, 2000); Romero *et al.* (2000); Gupta et Ghosh (2001). Ainsi, la seule analyse globale de l'usage des codons chez *Prochlorococcus* a utilisé l'AFC sur les données de RSCU (Paul *et al.*, 2010). Cependant, cette méthode n'est pas sans problèmes méthodologiques (Perrière et Thioulouse, 2002). En particulier, elle introduit des poids statistiques injustifiés conduisant à des résultats biaisés, spécialement pour l'usage des codons dans les acides aminés rares.

Une solution à ce problème est d'utiliser des analyses des correspondances inter- et intra-acides aminés sur les données de comptages des codons (Lobry et Chessel, 2003; Charif *et al.*, 2005). Ces méthodes décomposent la variabilité totale des données en une partie non synonyme due à l'usage différencié des acides aminés et en une partie synonyme due à l'usage des codons synonymes. Ces analyses permettent ainsi d'analyser d'un côté la variabilité inter-acides aminés et d'un autre la variabilité intra-acides aminés pour chacune des souches de *Prochlorococcus* et de *Synechococcus* mais aussi de comparer les différentes souches entre elles.

Les analyses pour les souches d'intérêt sont effectuées sur les effectifs des différents codons pour l'ensemble des gènes. Les gènes contenant plus d'un codon stop et moins de 100 codons ne sont pas pris en compte pour limiter la quantité d'ELF[2]. Nous obtenons ainsi, pour chaque souche, une table de comptage des codons pour chaque gène. A l'aide des paquets *ade4* (Thioulouse *et al.*, 1997) et *seqinR* (Charif et Lobry, 2007), des analyses des correspondances sont effectuées sur les effectifs des codons sans tenir compte de la structuration en groupe des codons. Celles-ci servent ensuite de point de départ pour les analyses inter-acides et intra-acides aminés qui nous intéressent.

X.3.1 Analyse de l'usage des acides aminés

La variabilité de l'usage des acides aminés représente entre 35% et 45% de la variabilité totale des données de l'usage des codons au sein des souches. Les coordonnées des gènes sur le premier axe (entre 15 et 24% de la variabilité intra-acides aminés) corrèlent fortement

[1]RSCU ou *Relative synonymous codon usage* (Sharp *et al.*, 1986) est une simple mesure de l'usage non uniforme de l'usage des codons dans une séquence codante. Les valeurs RSCU, une pour chacun des 61 codons possibles, correspondent au nombre de fois qu'un codon particulier est observé, rapporté au nombre de fois où ce codon serait observé si tous les codons d'un acide aminé ont la même probabilité d'être utilisé. En l'absence de tout biais d'usage des codons, la valeur RSCU d'un codon est 1.0. Un codon moins utilisé qu'attendu a une valeur RSCU inférieure à 1.0 et vice-versa pour un codon plus utilisé qu'attendu.

[2]ELF ou *Evil little f...ellows* (Ochman, 2002) sont des gènes dont la petite taille suggère que ce ne sont pas des gènes mais des portions aléatoires d'ADN détectés, à tort, comme des gènes.

avec le GC_{12} des gènes ($|r|$ entre 0.74 et 0.96), et sur le second axe (entre 11 et 15% de la variabilité intra-acides aminés) avec le score de gravy (Kyte et Doolittle, 1982), c'est-à-dire avec l'hydrophobicité moyenne des protéines ($|r|$ entre 0.76 et 0.91) pour toutes les souches. Ainsi, les principales causes de la variabilité de l'usage des acides aminés entre les gènes au sein des différentes souches de *Prochlorococcus* et de *Synechococcus* est la composition en GC des acides aminés puis l'hydrophobicité des acides aminés. Ces observations sont quelques peu différentes de celles de Paul *et al.* (2010) où l'hydrophobicité et l'aromaticité moyennes étaient les principaux contributeurs de la variation d'usage des acides aminés au sein des génomes de *Prochlorococcus* (sans suprématie de l'un par rapport à l'autre). Cette différence vient peut-être de l'utilisation dans l'analyse de Paul *et al.* (2010) d'une AFC sur l'usage relatif des acides aminés, pour laquelle des biais similaires à ceux observés pour l'AFC sur le RSCU ont pu être introduits.

Il n'y a pas de ségrégation claire des gènes dans le plan formé par les deux axes entre les gènes sur le brin précoce et ceux sur le brin tardif, alors que les gènes codant pour les protéines ribosomales se détachent sur le premier plan factoriel pour toutes les souches de *Prochlorococcus* et de *Synechococcus*. Ainsi, les gènes ribosomaux sont systématiquement moins hydrophobes que les autres gènes ($P < 0.01$, tests de Mann-Whitney pour toutes les souches). Pour le premier axe (GC_{12}), la tendance est moins claire : les gènes ribosomaux sont moins riches en GC_{12} que les autres gènes pour les souches de *Synechococcus* et de *Prochlorococcus* LLIV, ils ont des taux de GC_{12} équivalents aux autres gènes pour les souches de *Prochlorococcus* LLII/LLIII et LLI, et des taux supérieurs de GC_{12} pour les souches de *Prochlorococcus* HL. Ces différences s'expliquent par les changements de composition en bases GC des génomes entre *Synechococcus* et *Prochlorococcus* HL (Figure X.1).

D'après Paul *et al.* (2010), l'usage des acides aminés a changé au cours de l'évolution réductive, avec une augmentation de l'aromaticité et une diminution de l'hydrophobicité des protéines. Nous observons bien une différence d'utilisation des acides aminés entre les souches de *Prochlorococcus* lors de l'analyse inter-espèces de l'usage des acides aminés effectuée sur une table de comptage des acides aminés (Figure X.8). Cependant, le premier axe différenciant les clades les uns des autres et expliquant plus de 98% de la variabilité ne corrèle pas avec l'hydrophobicité des acides aminés, mais plutôt avec le contenu en GC des deux premières bases des codons (Figure X.8). Ainsi, les souches de *Synechococcus* et non réduites de *Prochlorococcus* LLIV utilisent principalement des acides aminés dont les codons commencent par G et/ou C. Les souches réduites de *Prochlorococcus* LL semblent utiliser un répertoire mixte d'acides aminés alors que les souches de *Prochlorococcus* HL utilisent principalement des acides aminés dont les codons commencent par A et/ou T. Ainsi, le changement d'usage des acides aminés au cours de l'évolution réductive semble principalement dû au changement de composition des génomes avec l'enrichissement en bases AT et non à une augmentation de la stabilité des protéines comme suggéré par Paul *et al.* (2010). Les changements pourraient donc ne pas être aussi adaptatifs qu'envisagé, même si l'utilisation d'acides aminés riches en AT pourrait être une adaptation à un environnement pauvre en nutriments, où les bases A et T sont moins coûteuses à produire.

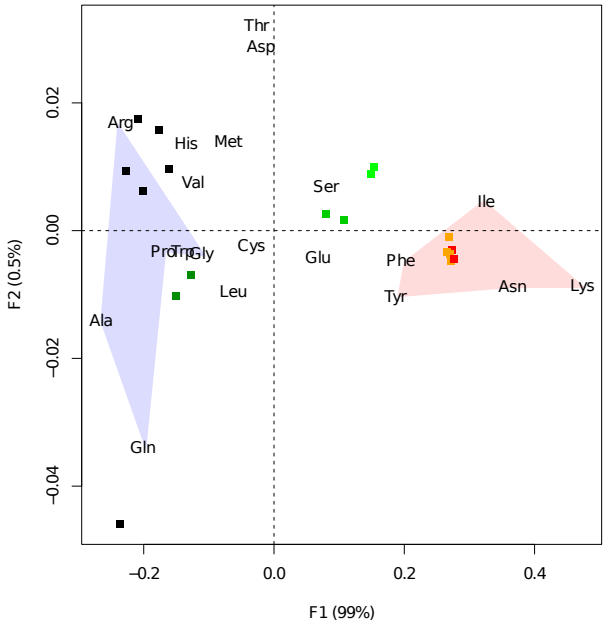


Figure X.8 – Première carte factorielle de l'analyse inter-espèces de l'usage des acides aminés pour *Prochlorococcus* et *Synechococcus*
Les couleurs symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI.
Les acides aminés dans la zone rouge pâle correspondent aux acides aminés dont les deux premières bases des codons sont A ou T alors que les acides aminés dans la zone bleu pâle correspondent aux acides aminés dont les deux premières bases des codons sont C ou G.

X.3.2 Usage des codons synonymes

La variabilité intra-acides aminés, c'est-à-dire la variabilité d'usage des codons synonymes, représente la part la plus importante de la variabilité totale des données de comptage des codons au sein des souches de *Prochlorococcus* et de *Synechococcus* (55 à 65%). Le premier axe est celui qui explique principalement la variabilité de l'usage des codons synonymes. Pour les souches *Synechococcus* et les souches non réduites de *Prochlorococcus*, cet axe explique entre 21% et 36% de la variabilité (Figure X.9) alors que pour les souches réduites, la variabilité expliquée par le premier axe chute jusqu'à 6% signifiant que la variabilité de l'usage des codons synonymes s'explique moins facilement pour les souches réduites, peut-être parce qu'elle est limitée.

La principale cause de la variabilité des codons synonymes entre les gènes, premier axe de l'analyse, est différente selon les clades : GC_3 pour les souches *Synechococcus*, GT_3 pour les souches de *Prochlorococcus* LL et C_3 pour les souches de *Prochlorococcus* HL (Figure X.9). Le deuxième axe corrèle principalement avec GT_3 pour les souches de *Synechococcus* et de *Prochlorococcus* HLI, avec GC_3 pour les souches de *Prochlorococcus* LL et A_3 pour les souches de *Prochlorococcus* HLII (Figure X.9).

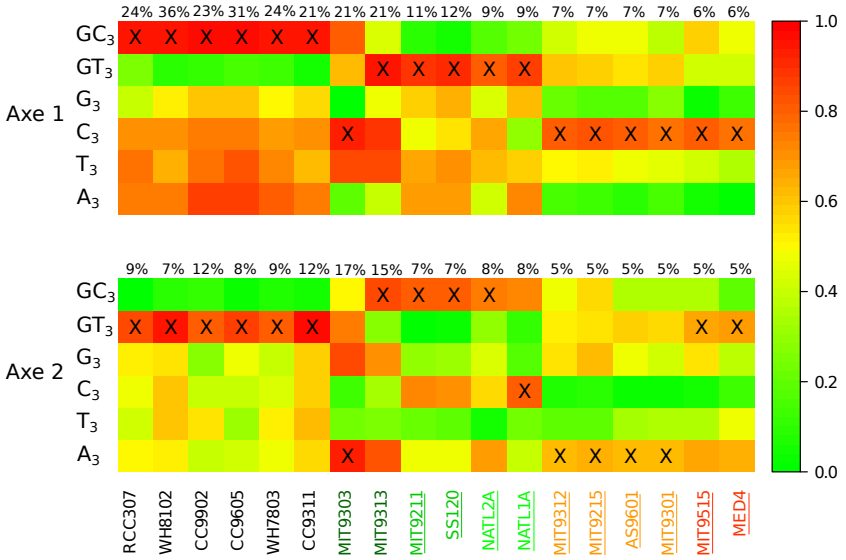


Figure X.9 – Corrélation entre GC_3, GT_3, G_3, C_3, T_3, A_3 et les deux premiers axes des analyses intra-acides aminés des souches de *Prochlorococcus* et de *Synechococcus*
Les couleurs du graphique représentent la valeur absolue du coefficient de corrélation entre GC_3, GT_3, G_3, C_3, T_3, A_3 et les coordonnées des gènes le long des deux premiers axes des analyses intra-acides aminés. Les croix symbolisent les corrélations les plus fortes pour chaque souche.
Pour chaque souche, la variabilité des données expliquée par les deux axes est indiquée en pourcentage.
Les couleurs de noms des souches symbolisent les différents écotypes avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit.

Ainsi, les causes de la variabilité d'usage des codons entre les gènes ont changé depuis la divergence entre *Synechococcus* et *Prochlorococcus*. Les gènes sur les brins précoces et tardifs se distinguent le long du deuxième axe pour les souches de *Synechococcus*, le long du premier axe pour les souches de *Prochlorococcus* LL et sur la combinaison des deux pour les souches de *Prochlorococcus* HL. Ces distinctions s'expliquent par les corrélations fortes des deux axes avec le taux GT_3 car le brin précoce est enrichi en bases G et T surtout visible pour les troisièmes positions des codons. Ainsi, le biais mutationnel de composition des brins se reflète seulement sur l'usage des codons synonymes pour les souches de *Prochlorococcus* LL, indice d'une pression de sélection pour la réplication et la transcription dans ces souches. Cependant, dans ce cas, le brin précoce contient plus de gènes du fait de la sélection réplicationnelle et est enrichi en gènes fortement exprimés par un effet de la sélection transcriptionnelle (McInerney, 1998; Lafay *et al.*, 1999; Das *et al.*, 2005, 2006). Chez *Prochlorococcus* LL, les gènes sont distribués presque équitablement entre les deux brins (Figure X.4b), avec légèrement moins de gènes sur le brin précoce sauf pour *Prochlorococcus* MIT9211 et NATL2A, indiquant ainsi l'absence d'une sélection réplicationnelle claire. Le brin précoce des souches de *Prochlorococcus* LLIV n'est pas enrichi en gènes ribosomaux contrairement des autres souches de *Prochlorococcus* LL (Figure X.4b). L'usage des codons des souches de *Prochlorococcus* LL réduites pourraient

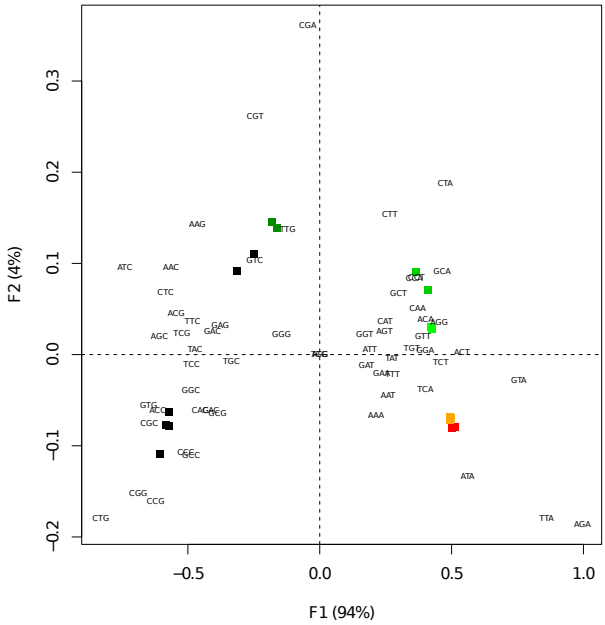


Figure X.10 – Première carte factorielle de l'analyse inter-espèces intra-acides aminés de l'usage des codons pour *Prochlorococcus* et *Synechococcus*
Les couleurs symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI.

ainsi être dirigé par une sélection transcriptionnelle.

L'analyse inter-espèce de l'usage des codons synonymes montre des différences entre les clades dans l'utilisation des codons synonymes (Figure X.10). Le premier axe de l'analyse, expliquant 94% de la variabilité, est fortement anti-corrélé avec le taux de GC_3 ($r = -0.85$, $P < 2{\cdot}10^{-16}$). Il discrimine les souches de *Synechococcus* et non réduites de *Prochlorococcus* des souches réduites de *Prochlorococcus* (Figure X.10). Ainsi, avec l'enrichissement en AT au cours de l'évolution réductive, l'usage des codons a changé pour des codons synonymes terminant principalement par A/T. Le second (4% de la variabilité) distingue les souches de *Prochlorococcus* LL et *Synechococcus* CC9311 et CC9902 des souches de *Prochlorococcus* HL et de *Synechococcus*. Aucune corrélation significative avec des caractères génomiques généraux n'a été trouvée pour cet axe.

X.4 Codons optimaux

Dans les gènes fortement exprimés, certains codons, sont préférés pour une traduction optimale de ces gènes. Ces codons préférés sont appelés codons optimaux. Comme l'enrichissement en bases AT des souches de *Prochlorococcus* a entraîné un changement dans l'utilisation des codons synonymes pour des codons terminant préférentiellement par A/T

dans l'ensemble des gènes, les codons optimaux doivent avoir suivi la même voie si l'usage des codons est principalement dirigé par une sélection traductionnelle.

Hershberg et Petrov (2009) ont estimé les codons optimaux pour 675 génomes bactériens, dont les souches de *Prochlorococcus* et de *Synechococcus* qui nous intéressent. Ils ont utilisé une façon répandue et simple de déterminer les codons favorisés : déterminer les codons codant pour un acide aminé particulier qui augmentent en fréquence dans les gènes fortement exprimés, les plus biaisés dans le choix global des codons synonymes à cause de la sélection traductionnelle (Duret et Mouchiroud, 1999; Vicario *et al.*, 2007; Akashi et Schaeffer, 1997; Akashi, 1995). À partir de la corrélation entre l'ENC' (Novembre, 2002) des gènes et la fréquence de chacun des codons de ces gènes, Hershberg et Petrov (2009) conservent comme codons optimaux ceux dont la corrélation négative significative est la plus forte pour chaque acide aminé. L'ENC' est donc utilisé comme un estimateur du niveau d'expression et les codons optimaux choisis sont ceux utilisés par les gènes les plus fortement exprimés. L'utilisation de l'ENC' pour définir les gènes fortement exprimés pose problème pour *Prochlorococcus* MED4. En effet, l'ENC' n'est pas corrélé au niveau d'expression (Figure X.6). Nous ne pouvons donc pas utiliser les codons optimaux détectés par Hershberg et Petrov (2009) pour analyser les changements de codons optimaux au cours de l'évolution réductive.

Pour récupérer les codons optimaux dans les souches d'intérêt, nous utilisons deux méthodes qui ne reposent pas sur l'ENC' ou un autre estimateur du biais d'usage des codons, pour définir un ensemble de gènes de référence. Notre ensemble de gènes de référence est l'ensemble des gènes codant pour les protéines ribosomales, considérés comme des gènes fortement exprimés même chez *Prochlorococcus* MED4 (Figure X.6). La traduction au sein de ces gènes doit être optimisée pour permettre une production rapide des protéines ribosomales et des ribosomes. Ainsi, ces gènes doivent utiliser préférentiellement les codons optimaux pour la traduction.

Dans une première méthode, les codons optimaux sont déterminés par l'analyse différentielle de l'usage des codons entre les gènes ribosomaux et les autres gènes. Nous utilisons les effectifs des codons au sein de l'ensemble des gènes ribosomaux et des gènes non ribosomaux en supprimant les gènes avec plus d'un codon stop et moins de 100 codons (pour diminuer la proportion d'ELF (Ochman 2002)). Pour chaque codon, nous effectuons un test de χ^2 d'adéquation entre les observations pour les gènes ribosomaux et les gènes non ribosomaux et les effectifs attendus sous l'hypothèse d'une utilisation identique des codons entre les gènes ribosomaux et les gènes non ribosomaux. Les codons dont le test χ^2 est significatif à 5% et dont l'effectif réel est supérieur à l'effectif attendu pour les gènes ribosomaux sont conservés comme codons optimaux. Cette méthode présente quelques limites. En particulier, lorsque le contenu en bases GC est fortement biaisé comme pour les souches réduites de *Prochlorococcus*, l'utilisation des codons est biaisée et les différences d'usage des codons entre les gènes ribosomaux et les autres gènes peuvent être trop faibles pour permettre l'identification de codons optimaux.

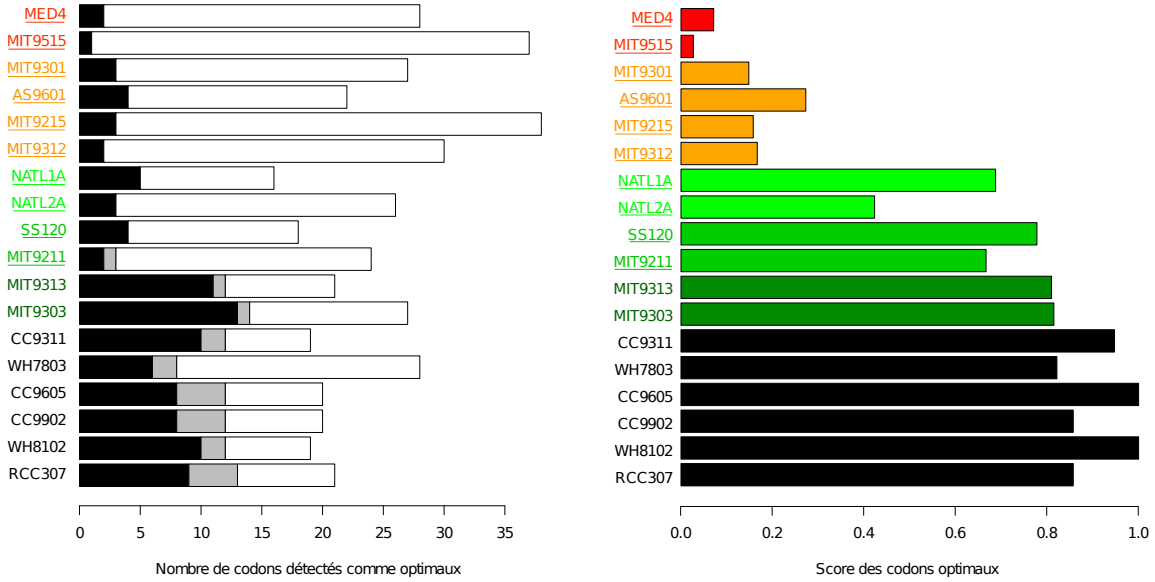
Nous utilisons ainsi une deuxième méthode basée sur les analyses factorielles des correspondances intra-acides aminés des effectifs des codons (Section X.3.2). Les codons

optimaux sont définis ici comme les codons utilisés préférentiellement par les gènes ribosomaux, c'est-à-dire les codons qui se situent dans les même secteurs que les gènes ribosomaux sur les cartes factorielles des analyses intra-acides aminés. Ainsi, nous conservons comme codons optimaux, les codons situés dans le carré de la première carte factorielle dont les limites sont définies pour que 90% des gènes ribosomaux se trouvent dans le carré.

Avec nos deux méthodes, les codons optimaux ne sont pas restreints à un par acide aminé. Dans les analyses où un seul codon optimal par acide aminé est nécessaire, nous choisissons celui dont la différence entre les gènes ribosomaux et les gènes non ribosomaux est la plus grande avec la méthode basée sur l'AFC intra-acides aminés.

Globalement, plus de codons optimaux sont trouvés pour les souches réduites de *Prochlorococcus* (Figure X.11a) mais ils sont principalement trouvés grâce à la seconde méthode et peu par la première méthode. Ainsi, dans les souches réduites de *Prochlorococcus* riches en bases AT, les gènes ribosomaux utilisent peu de codons à des fréquences significativement supérieures à celles des autres gènes. Ils sont donc aussi plus proches des autres gènes dans la première carte factorielle de l'analyse intra-acides aminés, entraînant de fait plus de codons optimaux détectés avec la seconde méthode.

L'identité des codons optimaux semble liée aux taux de GC (Figure X.12) : les souches réduites de *Prochlorococcus* enrichies en bases AT semblent utiliser des codons plus riches en bases AT. Pour étudier plus formellement cette relation, nous utilisons la méthode présentée par Hershberg et Petrov (2009). Les codons de chaque acide aminé sont classés et reçoivent un score : 1 pour les plus riches en GC, -1 pour les plus riches en AT et 0 sinon. Un score est calculé pour chaque génome en sommant les scores des codons optimaux puis en normalisant par le nombre de codons optimaux. Le score d'un génome est alors compris entre -1 et 1, avec -1 si seuls des codons riches en AT sont utilisés et 1 si seuls des codons riches en GC sont utilisés. Le score est moins élevé pour les souches réduites de *Prochlorococcus* que pour les souches de *Synechococcus* et non réduites de *Prochlorococcus* (Figure X.11b), signifiant que les premières utilisent des codons optimaux moins riches en GC que les secondes. Cependant, malgré un taux de GC intergénique inférieur à 40%, le score est systématiquement supérieur à 0 : les souches utilisent plus de codons optimaux riches en GC que de codons riches en AT. Les scores étant très proches de 1 pour les souches de *Synechococcus* et non réduites de *Prochlorococcus*, celles-ci utilisent quasiment exclusivement comme codons optimaux les codons les plus riches en GC. Ainsi, au cours de l'évolution de *Prochlorococcus* et *Synechococcus*, le répertoire des codons optimaux a évolué avec un enrichissement en AT suivant l'enrichissement génomique global pour les souches réduites de *Prochlorococcus*. Cependant, le changement des codons optimaux est plus lent que l'enrichissement des zones intergéniques : les codons optimaux sont donc plus pauvres en GC qu'enrichis en AT.



(a) Nombre de codons optimaux trouvés **(b)** Score des codons optimaux

Figure X.11 – Caractéristiques globales des codons optimaux trouvés chez *Prochlorococcus* et *Synechococcus*
Pour la figure X.11a, les couleurs au sein des barres symbolisent la répartition des codons optimaux trouvés. En noir, ce sont les codons optimaux trouvés par les deux méthodes ; en gris, ceux trouvés seulement par la première méthode (analyse différentielle de l'usage des codons entre les gènes ribosomaux et non ribosomaux) et en blanc ceux trouvés seulement avec la seconde méthode (AFC intra-acides aminés).
Pour la figure X.11b, les scores des codons optimaux sont calculés en classant les codons de chaque acide aminé selon le taux de GC, riche en AT et intermédiaire. Le score des codons est ensuite attribué selon ces classes : 1 pour les codons riches en GC, -1 pour les codons riches en AT et 0 pour les intermédiaires. Un score par génome est calculé en sommant les scores des codons optimaux puis en normalisant par le nombre de codons optimaux. Le score d'un génome est alors entre -1 et 1, avec -1 si seuls des codons riches en AT sont utilisés et 1 si seuls des codons riches en GC sont utilisés.
Les couleurs de noms des souches symbolisent les différents écotypes avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit.

X.5 Gènes ARNt

Chez *Buchnera aphidicola*, la réduction du génome a impacté le contenu en ARN_t et une perte de corrélation entre l'expression des ARN_t et l'usage des codons correspondants (Hansen et Moran, 2012)[1]. Reis *et al.* (2004) ont proposé un modèle de relation entre la taille des génomes et le nombre de gènes ARN_t, lié à la sélection traductionnelle, qui pourrait expliquer les motifs de gènes ARN_t chez *Buchnera*. Les souches de *Prochlorococcus* suivent-elles ce modèle ? Les répertoires de gènes ARN_t ont-ils changé avec l'évolution

[1]Cette perte n'est cependant pas aussi forte qu'attendu dans le cas de la dégénérescence (Charles *et al.*, 2006)

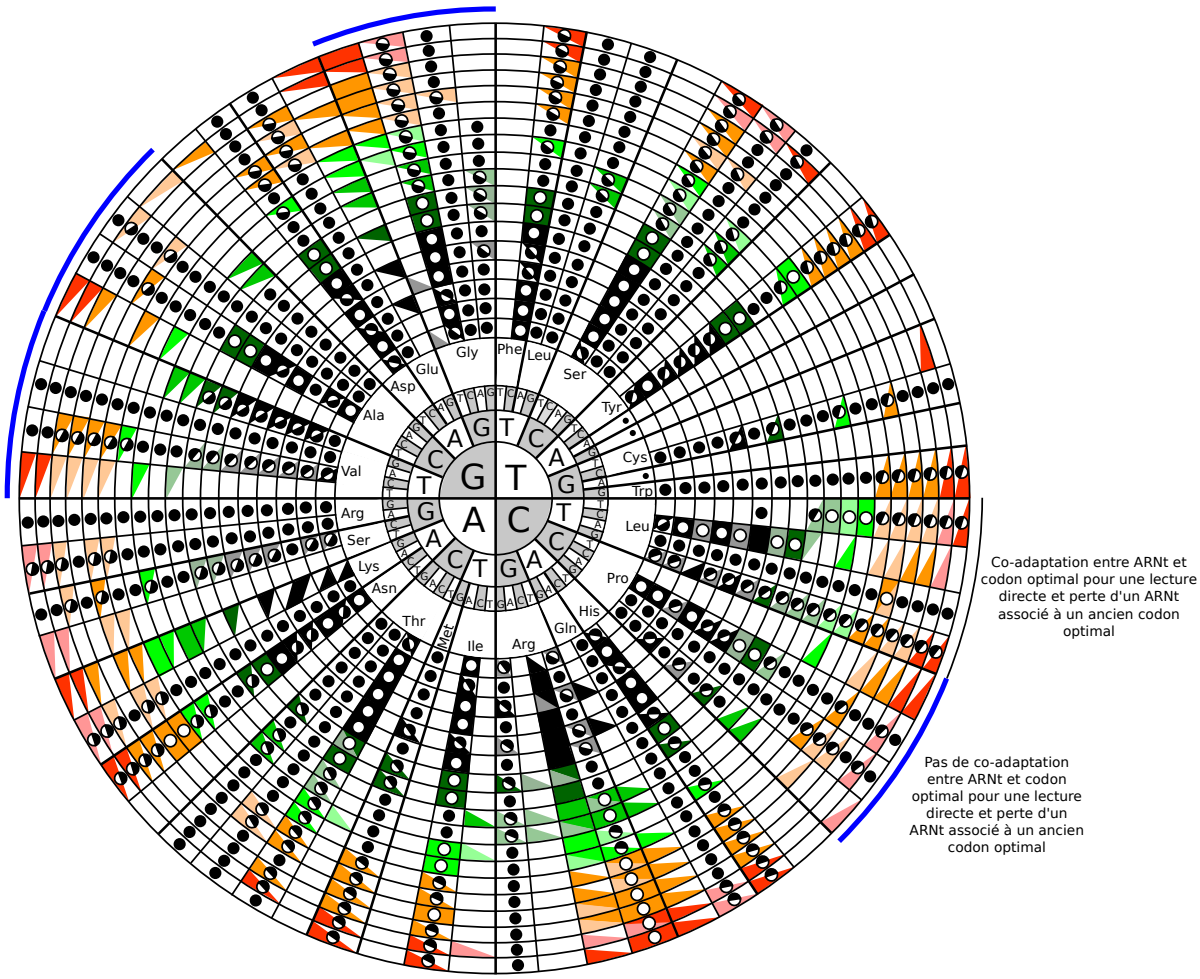


Figure X.12 – Codons optimaux et gènes ARN$_t$ des souches de *Prochlorococcus* et de *Synechococcus*

Les ronds représentent la présence de gènes ARN$_t$ pour décoder le codon correspondant, avec en blanc quand l'ARN$_t$ correspond directement (sans prendre en compte les règles de wobble) à un codon optimal.

Les cases colorées correspondent aux codons optimaux, avec une case colorée totalement lorsque le codon est détecté comme optimal par les deux méthodes d'identification des codons optimaux et une case colorée à moitié lorsque le codon est détecté comme optimal par une des deux méthodes. Les couleurs vives correspondent aux codons préférés au sein de chacun des acides aminés, c'est-à-dire les codons dont la distance est la plus grande avec les gènes non ribosomaux dans la méthode basée sur l'AFC intra-acides aminés.

Les souches réprésentées sont, du centre vers l'extérieur, *Synechococcus* RCC307, *Synechococcus* CC9902, *Synechococcus* CC9605, *Synechococcus* WH8102, *Synechococcus* WH7803, *Synechococcus* CC9311, *Prochlorococcus* MIT9303, *Prochlorococcus* MIT9313, *Prochlorococcus* MIT9211, *Prochlorococcus* SS120, *Prochlorococcus* NATL2A, *Prochlorococcus* NATL1A, *Prochlorococcus* MIT9312, *Prochlorococcus* MIT9215, *Prochlorococcus* AS9601, *Prochlorococcus* MIT9301, *Prochlorococcus* MIT9515 et *Prochlorococcus* MED4. Les couleurs symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI.

Les nucléotides avec un fond gris correspondent aux nucléotides G et C.

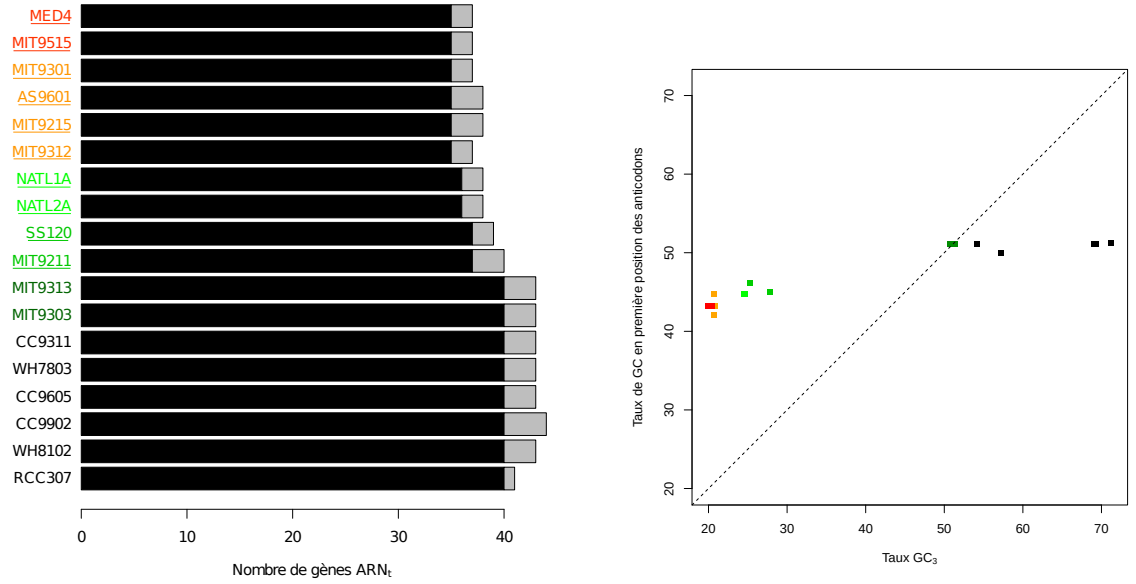
réductive ? Si oui, ont-ils suivi les changements d'usage des codons ou les ont-ils induits ?

Limor-Waisberg *et al.* (2011) avaient exploré le répertoire de gènes ARN_t chez *Prochlorococcus* et *Synechococcus*, mais seulement comme comparaison au répertoire des ARN_t des cyanophages infectant *Prochlorococcus* et *Synechococcus* et avec seulement 11 des 18 souches qui nous intéressent. Ils ont montré que les souches étudiées évitent presque totalement les anticodons démarrant par l'adénosine, c'est-à-dire les codons terminant par la thymine, observation explicable seulement en partie par les interactions wobble[1] (Reis *et al.*, 2004). Inversement, les anticodons démarrant par la thymine sont quasiment tous utilisés. D'autres restrictions du répertoire de gènes ARN_t ont lieu : pour les 61 anticodons possibles, *Synechococcus* WH8102 possède un peu plus de 40 gènes ARN_t et *Prochlorococcus* HL MED4 32. La différence entre les répertoires est principalement attribuée au faible contenu en GC des souches de *Prochlorococcus* HL, évitant certains anticodons démarrant par guanine et cytosine. Le répertoire de gènes des souches de *Prochlorococcus* HL représente alors un répertoire pauvre en anticodon GC plutôt qu'un répertoire riche en anticodon AT. Cependant, le lien entre le répertoire d'ARN_t et l'usage des codons n'est pas clairement étudié dans l'analyse de Limor-Waisberg *et al.* (2011) et une analyse plus complète, avec toutes les souches d'intérêt, est nécessaire.

Comme observé par Limor-Waisberg *et al.* (2011), le nombre de gènes ARN_t est plus faible pour les souches réduites de *Prochlorococcus* que pour les souches non réduites de *Prochlorococcus* et de *Synechococcus*, avec moins d'anticodons différents[2] (Figure X.13a). Les pertes ont eu principalement lieu dans la branche ancestrale aux souches réduites de *Prochlorococcus* (Figure X.14), par la perte d'anticodons associés à des codons terminant par G/C, en cohérence avec l'appauvrissement en base GC observé le long de cette branche. Ainsi, les taux de GC des premières bases des anticodons pour les souches réduites sont inférieurs à ceux des souches non réduites de *Prochlorococcus* et de *Synechococcus* (Figure X.13b). Or, si le répertoire des gènes ARN_t suivait seulement les changements de bases GC, les premières positions des anticodons devraient avoir des taux similaires aux taux de GC_3 génomique. Or, le taux de GC des premières bases des anticodons ne suit pas cette tendance pour nos souches (à l'exception des souches de *Prochlorococcus* LLIV), avec un taux supérieur à celui attendu pour les souches réduites et un taux inférieur pour *Synechococccocus* (Figure X.13b). De plus, en dehors du cas de la leucine pour laquelle le répertoire d'ARN_t s'est adapté au changement de codons optimaux, dans les acides aminés pour lesquels un ARN_t a été perdu, cet ARN_t correspond à un codon optimal ancien et il n'a pas été remplacé par un ARN_t correspondant au nouveau codon optimal (Figure X.12). Comme observé par Rocha (2004) sur 102 bactéries, la composition des anticodons ARN_t s'adapte seulement légèrement à la composition GC génomique et à l'usage des codons en découlant.

[1] L'appariement de wobble est un mode d'appariement non canonique entre les bases nucléotidiques, différent de l'appariement de Watson-Crick par la nature des bases et des liaisons impliquées. Cet appariement utilisé en première position de l'anticodon permet à un ARN_t de reconnaître plusieurs codons. En effet, l'inosine, base modifiée, trouvée fréquemment en première position de l'anticodon, permet des appariements avec les bases U, A et C de l'ARN.

[2] Les gènes ARN_t des souches sont trouvés à l'aide de *tRNAscan-SE* (Lowe et Eddy, 1997) avec les paramètres par défaut pour la recherche de gènes ARNt chez les bactéries.



(a) Nombre de gènes ARN_t trouvés

(b) Taux de GC de la première base des anticodons en fonction du GC_3 des génomes

Figure X.13 – Caractéristiques globales des gènes ARN_t chez *Prochlorococcus* et *Synechococcus*
Pour la Figure X.13a, les couleurs des barres symbolisent la répartition des gènes ARN_t, avec en noir le nombre de gènes codant pour les anticodons différents et en noir et gris le nombre de gènes ARN_t total.
Les couleurs des noms pour la figure X.13a et des points pour la figure X.13b symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelle le génome est réduit.
La droite en pointillé correspond à $y = x$, c'est-à-dire aux valeurs attendues de taux de GC de la première base des anticodons s'il y a une relation entre GC génomique et composition des anticodons ARN_t

Pour comprendre l'évolution des répertoires de gènes ARN_t et leurs différences entre les différentes clades de *Prochlorococcus*, nous intéressons à l'association entre les fréquences des anticodons et le biais d'usage des codons.

Comme la traduction est le processus le plus coûteux énergétiquement dans les cellules croissant exponentiellement, son efficacité est soumise à des pressions sélectives importantes. L'étape limitante dans l'élongation de la chaine de polypeptides est la diffusion du complexe ARN_t-acide aminé (ARN_t-aa) vers le site A du ribosome (Varenne *et al.*, 1984), entrainant un recrutement prédominant du complexe ARN_t-aa le plus abondant par les codons des gènes fortement exprimés (Dong *et al.*, 1996). Les codons les plus favorables sont donc ceux correspondant aux complexes ARN_t-aa les plus abondants et efficaces (Andersson et Kurland, 1990). Comme les concentrations d'ARN_t-aa ne sont pas disponibles, nous considérons que ces concentrations sont proportionnelles au nombre de gènes ARN_t dans le génome (Dong *et al.*, 1996; Percudani *et al.*, 1997; Kanaya *et al.*, 1999; Duret, 2000). En prenant en compte les règles du wobble résumées dans le tableau 2 de Rocha

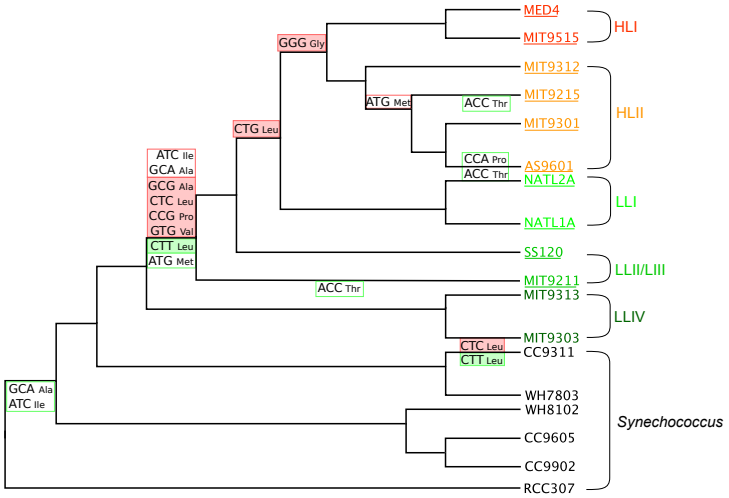
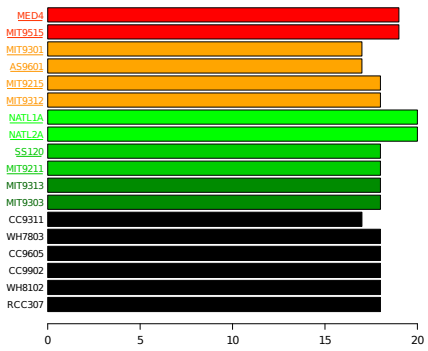
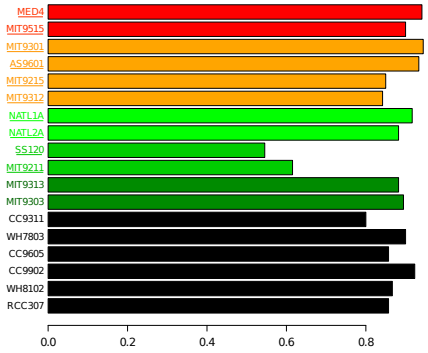


Figure X.14 – Gains et pertes des gènes ARN$_t$ le long de la phylogénie des souches de *Prochlorococcus* et de *Synechococcus*
Les gains et pertes ont été reconstruits à l'aide de la parcimonie de Wagner.
Sont indiqués les codons et acides aminés correspondant directement aux gènes ARN$_t$ gagnés ou perdus. Les rectangles pleins correspondent aux gains ou pertes d'un gène présent en une seule copie et les rectangles vides aux gains ou pertes d'une copie du gène concerné. Les rectangles rouges correspondent aux pertes et les rectangles verts aux gains.
Les couleurs de noms des souches symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit. L'arbre phylogénétique est construit comme décrit dans la section C.4 (Annexe) mais les longueurs des branches sont arbitraires et ne reflètent pas les taux d'évolution.

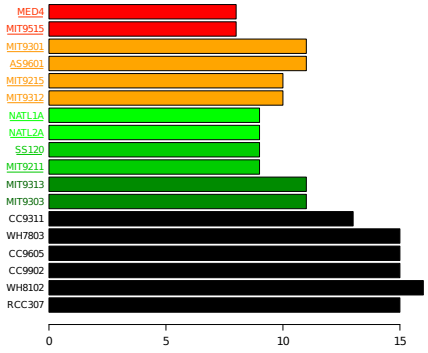
(2004), un codon peut être lu par plusieurs anticodons. Le nombre d'acides aminés où le codon le plus fréquent correspond au codon le plus lu par des ARN$_t$ est relativement bien conservé entre les souches de *Prochlorococcus* et de *Synechococcus* avec des valeurs comprises entre 17 et 20 (Figure X.15a). Ainsi, pour la grande majorité des acides aminés, les codons les plus favorables correspondent aux complexes ARN$_t$-aa les plus abondants. Il en est de même pour les codons optimaux dont la majorité correspondent aux codons les plus lus malgré une forte réduction pour les souches de *Prochlorococcus* LLII/LLIII (Figure X.15b). Les acides aminés pour lesquels le codon le plus fréquent ne correspond pas au codon le plus lu par des ARN$_t$ changent entre les différentes clades : sérine et arginine pour *Synechococcus* et *Prochlorococcus* LLIV, leucine et alanine pour *Prochlorococcus* LLII/LLIII, alanine et glycine pour *Prochlorococcus* HLII et glycine pour *Prochlorococcus* HLI. La glycine, l'alanine et la leucine font parti des acides aminés les moins utilisés par les souches réduites de *Prochlorococcus* par rapport aux souches non réduites (Figure X.8). La perte de correspondance entre le codon le plus fréquent et le nombre d'ARN$_t$ utilisés pour le décoder peut refléter une perte de sélection traductionnelle pour ces acides aminés, en particulier dans le cas d'un changement d'usage des acides aminés où la sélection sur les codons optimaux s'adapte.



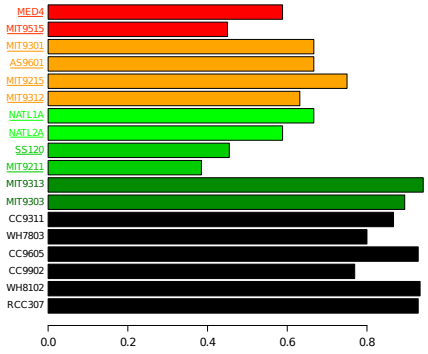
(a) Nombre d'acides aminés où le codon le plus fréquent correspond au codon le plus lu par des ARN_t



(b) Proportion de codons optimaux correspondant au codon le plus lu par des ARN_t



(c) Nombre d'acides aminés où le codon le plus fréquent est lu directement par un ARN_t



(d) Proportion de codons optimaux lus directement par un ARN_t

Figure X.15 – Association entre anticodons et biais d'usage des codons chez *Prochlorococcus* et *Synechococcus*
Le codon le plus fréquent pour chaque acide aminé correspond au codon le plus utilisé dans l'ensemble des gènes de chacune des souches. Le codon le plus lu par des ARN_t correspond au codon lisible par le maximum d'ARN_t associés aux gènes trouvés dans les génomes, étant donné les règles du tableau 2 de Rocha (2004).
Les codons optimaux correspondent aux codons optimaux préférés au sein de chacun des acides aminés, c'est-à-dire les codons dont la distance est la plus grande avec les gènes non ribosomaux dans la méthode basée sur l'AFC intra-acides aminés utilisée pour détecter les codons optimaux.
Les couleurs symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI.

D'après Ikemura (1981), pour augmenter la spécificité et la sensibilité du ribosome, le codon le plus fréquent doit avoir un interaction optimale avec l'anticodon, c'est-à-dire correspondre parfaitement, avec l'appariement canonique de Watson-Crick, à l'anticodon le plus fréquent. Comme la plupart des ARN_t sont en un seul exemplaire chez *Prochlorococcus* et *Synechococcus*, nous regardons le nombre d'acides aminés pour lesquels le codon le plus fréquent est lu directement par un ARN_t sans intervention des règles de wobble (Figure X.15c). Ce nombre est réduit pour les souches de *Prochlorococcus* par rapport aux souches de *Synechococcus*, signifiant qu'il y a une perte de correspondance parfaite

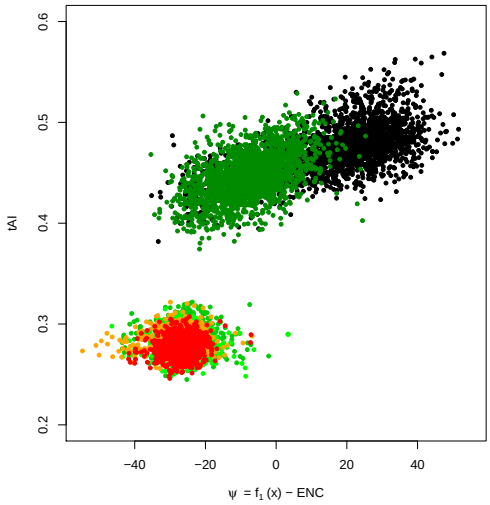


Figure X.16 – *tRNA adaptation index* (tAI) en fonction de $f_1(x) - ENC$ pour *Prochlorococcus* et *Synechococcus*
tAI est calculé sur l'ensemble des gènes des souches selon la méthode expliquée dans Reis *et al.* (2004). Les valeurs d'ENC des gènes ont été calculées à l'aide de l'outil ENCprime sur tous les gènes des souches. $f_1(x) = -6 + GC_3 + \frac{34}{GC_3^2 + (1.025 - GC_3)^2}$ avec GC_3 le taux de bases GC_3 des gènes (Reis *et al.*, 2004). $\psi = f_1(x) - ENC$ représente l'effet de la sélection sur l'usage des codons (Reis *et al.*, 2004).
Les couleurs symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Une souche par clade est représentée, les autres souches ayant des tendances similaires.

entre les gènes ARN_t présents et les codons préférés mais aussi les codons optimaux (Figure X.15d). Cependant, d'après Grosjean et Fiers (1982), les interactions entre codon et anticodon doivent être ni trop fortes, ni trop faibles, car les premières ralentissent le renouvellement des ARN_t dans le ribosome et les dernières peuvent conduire à des erreurs de traduction fréquentes et/ou des taux plus forts de rejet incorrects des ARN_t par le ribosome. Selon ce modèle dit de stabilité, les meilleurs codons démarrant avec deux bases fortes ($S = \{G, C\}$) sont ceux avec une troisième base faible ($W = \{A, U\}$) et inversement, les meilleurs codons démarrant par deux bases faibles doivent avoir une troisième base forte. Pour toutes les souches de *Prochlorococcus* et de *Synechococcus*, le nombre d'acides aminés respectant ces règles est inférieur au nombre d'acides aminés ne les respectant pas, sans réelle différence entre les souches réduites et non réduites. Ainsi, le modèle de stabilité ne semble pas pouvoir expliquer l'association entre les fréquences des anticodons et le biais d'usage des codons dans nos souches d'intérêt.

La perte de correspondance directe entre les ARN_t et les codons les plus favorables est-elle due à un relâchement de la sélection traductionnelle ? Pour étudier cette question, nous avons utilisé l'indice *tRNA adaptation index* (tAI) (Reis *et al.*, 2004) qui permet de mesurer l'usage des ARN_t par les séquences. Comme l'abondance en ARN_t est le facteur

limitant lors de la traduction, la mesure de l'usage des ARN_t dans un gène peut fournir un moyen indirect de détecter la sélection traductionnelle sur un gène en fonction de son adaptation au répertoire de gènes ARN_t. Les gènes des souches réduites de *Prochlorococcus* présentent un *tAI* faible (Figure X.16) et semblent moins bien adaptés à leur répertoire de gènes ARN_t que ne le sont les gènes des souches non réduites de *Prochlorococcus* et de *Synechococccus*, Ceci semble cohérent avec la supériorité du taux de GC des premières bases des anticodons par rapport au taux de GC_3 des souches réduites de *Prochlorococcus* (Figure X.13b). Ainsi, les souches réduites de *Prochlorococcus* utilisent plus de wobble que les souches non réduites et *Synechococcus*, ce qui semble en accord avec la perte de correspondance parfaite entre les codons les plus favorables et les ARN_t (Figures X.15c et X.12).

Le degré de sélection traductionnelle à l'origine de la co-adaptation entre l'usage des codons et le répertoire d'ARN_t peut être inféré par la pente entre tAI et la quantité ψ de sélection agissant sur l'usage des codons. Plus la co-adaptation est due à la sélection traductionnelle, plus forte est la corrélation entre tAI et ψ. Pour les souches réduites de *Prochlorococcus*, la corrélation entre tAI et ψ est la plupart du temps non significative ou alors négative. Au contraire, pour les souches non réduites de *Prochlorococcus* et de *Synechococcus*, la corrélation est systématiquement significative ($P < 0.01$) et comprise entre 0.45 et 0.62. La sélection traductionnelle pour une co-adaptation entre l'usage des codons et les répertoires des gènes ARN_t, présente pour les souches non réduites de *Prochlorococcus* et de *Synechococcus*, semble donc avoir été perdue pour les souches réduites de *Prochlorococcus*.

X.6 Discussion

Le changement de composition des génomes chez *Prochlorococcus* par l'enrichissement en bases AT n'a pas seulement un impact dans les zones intergéniques. Ainsi, les acides aminés, les codons synonymes mais aussi les codons optimaux ou les anticodons utilisés sont moins riches en GC que pour les souches *Synechococcus*. L'usage des codons a ainsi changé au cours de l'évolution et de la diversification de *Prochlorococcus* (Figure X.17). Tout comme les pertes de gènes (Figure VIII.3), les changements de l'usage des codons semblent avoir démarré peu après la divergence entre *Prochlorococcus* et *Synechococcus* avec l'augmentation de l'ENC' et la perte de correspondance parfaite entre les ARN_t et les codons les plus fréquents, synonymes d'un relâchement des pressions de sélection traductionnelle. Mais les principaux changements ont eu lieu le long de la branche ancestrale aux souches réduites avec une homogénéisation de l'usage des codons entre tous les gènes (qu'ils soient fortement exprimés ou pas), la perte de gènes ARN_t, la réduction de l'influence de la sélection traductionnelle sur la co-adaptation entre l'usage des codons et le répertoire ARN_t. Le biais d'usage des codons est alors soumis principalement aux contraintes de composition des génomes et moins à la sélection traductionnelle. Ainsi, l'asymétrie de composition des brins explique une part non négligeable de l'usage des codons mais c'est surtout la composition en bases AT qui joue un rôle important.

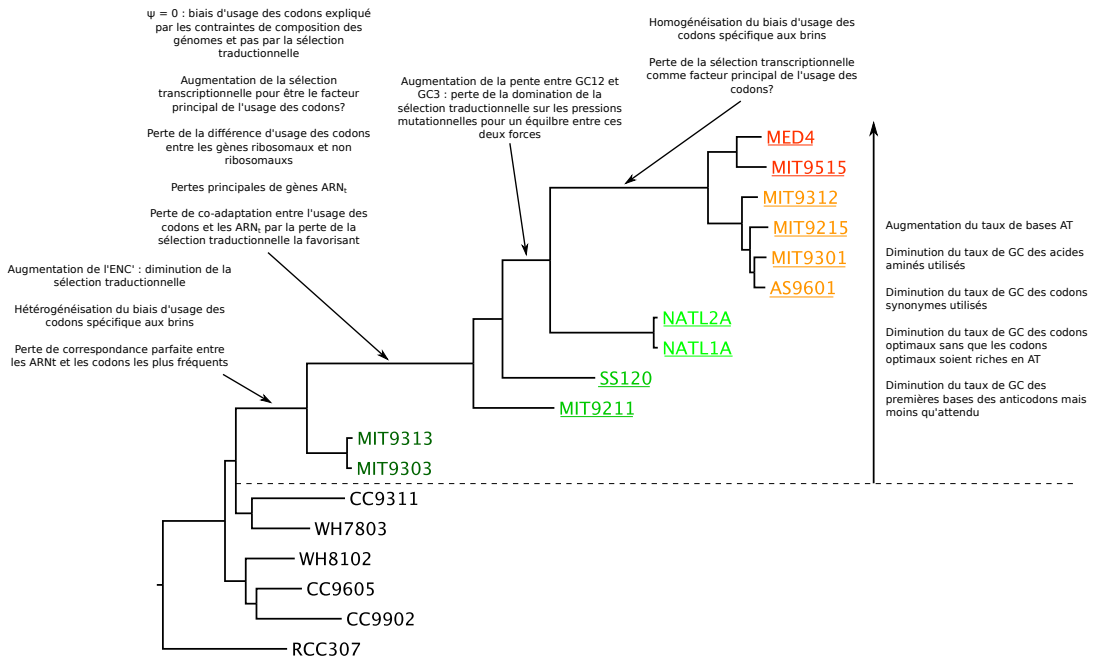


Figure X.17 – Résumé des changements de composition, d'usage des codons, des codons optimaux et des répertoires ARN$_t$ observés chez *Prochlorococcus*
Les couleurs de noms des souches symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit. L'arbre phylogénétique est construit comme décrit dans la section C.4 (Annexe).

D'après Hershberg et Petrov (2009), lorsqu'un génome change de contenu global en bases GC, les codons correspondant au nouveau taux de GC sont progressivement utilisés dans les gènes qui ne sont pas sous forte sélection traductionnelle. Cela affecte globalement leur efficacité de traduction alors qu'individuellement ces gènes ne sont pas assez fortement exprimés pour être sous forte sélection pour l'utilisation des codons optimaux de traduction. Selon cette théorie, les ARN$_t$ correspondant aux codons nouvellement fréquents sont avantagés et leur expression augmente. Les gènes fortement exprimés peuvent ainsi utiliser les codons correspondant au GC global sans perdre trop d'efficacité. L'expression des ARN$_t$ reconnaissant les anciens codons optimaux se réduit par la suppression de sélection pour un fort niveau d'expression. Les génomes sont de nouveau à l'équilibre en terme d'usage des codons et de la sélection traductionnelle. Cette théorie pourrait s'appliquer aux souches de *Prochlorococcus*. Les gènes non ribosomaux ont des taux de GC plus faibles que les gènes ribosomaux et ils utilisent des codons synonymes moins riches en GC. Ainsi, les gènes qui ne sont pas sous forte sélection traductionnelle utilisent les codons correspondant au nouveau taux de GC. Les répertoires d'ARN$_t$ changent mais plus lentement : les taux de bases GC des anticodons sont légèrement plus faibles, les codons les plus fréquents sont les plus lus par des ARN$_t$ mais moins directement. Ainsi, des changements ont aussi lieu dans les gènes ribosomaux sans une grosse perte d'efficacité : les codons optimaux de traduction sont moins riches en GC que dans les souches où

le taux de GC est moyen et les codons optimaux correspondent aux codons les plus lus (faible distinction entre les codons préférentiellement utilisés par les gènes ribosomaux et les autres gènes), à l'exception des souches LLII/LLIII. Ces dernières pourraient être dans une phase moins avancée du processus que les autres souches réduites : les changements de répertoires ARN_t ne sont pas aussi avancés que dans les autres souches. L'expression des ARN_t reconnaissant les anciens codons optimaux s'est réduite : un certain nombre d'entre eux ont été perdus. Les génomes n'ont cependant pas encore atteint l'équilibre pour l'usage des codons. À l'exception de la leucine, les répertoires de gènes ARN_t n'ont pas suivi les changements de codons optimaux. De plus, la plupart des autres indicateurs de sélection traductionnelle montrent que l'usage des codons au sein des souches réduites de *Prochlorococcus* est principalement dominé par des contraintes de composition. Les changements ne seraient donc pas assez avancés pour que la sélection traductionnelle puisse reprendre une place dominante, à moins que la dérive génétique ne soit plus forte et que les changements d'usage des codons n'atteignent pas un équilibre.

Se pose la question de la cause de l'enrichissement en bases AT. Certains gènes impliqués dans la réparation des mutations GT vers AT ont été perdus (Figure VIII.6). Mais pourquoi ont-ils été perdus ? Pourrait-il y avoir une pression de sélection favorisant les génomes riches en AT car potentiellement moins coûteux dans des environnements pauvres en nutriments (Giovannoni *et al.*, 2005, 2014), effaçant ainsi la limite entre des biais mutationnels neutres et une adaptation ? Cependant, les acides aminés utilisés préférentiellement par les souches réduites ne semblent pas avoir des coûts métaboliques plus faibles que ceux utilisés par les souches non réduites ($P = 0.673$, test de Student sur les cinq acides aminés préférés pour les souches réduites et les souches non réduites, avec les coûts métaboliques définis par Akashi et Gojobori (2002)). La dérive génétique pourrait aussi expliquer les observations malgré les grandes tailles efficaces de populations, d'autant que les changements d'usage des codons semblent avoir démarré avant les pertes de gènes de réparation. La question des pressions de sélection est abordée dans le chapitre suivant.

Chapitre XI

Évolution des séquences et pressions de sélection

Certaines des caractéristiques observées dans les chapitres précédents semblaient pouvoir être imputées à des changements de pressions de sélection, et peut-être à de la dégénérescence comme observé chez les endosymbiotes (Moran, 1996). Cependant, les précédentes études sur les pressions de sélection chez *Prochlorococcus* n'ont pas mis en évidence une réduction de la pression de sélection au niveau protéique (Paul *et al.*, 2010; Hu et Blanchard, 2009; Yu *et al.*, 2012). Ces analyses utilisent l'approche classique d'étude des forces de sélection purificatrice et positive par des estimations de $^{d_N}/_{d_S}$, nommé aussi $^{K_a}/_{K_s}$, c'est-à-dire le rapport entre le taux de fixation des substitutions non synonymes (d_N) et le taux de fixation des substitutions synonymes (d_S). Si la sélection n'a pas d'effet sur les individus et leur fitness, les substitutions synonymes et non synonymes sont fixées à des taux similaires, $^{d_N}/_{d_S} = 1$. Si la sélection purificatrice est la force principale dans l'évolution des séquences, les substitutions non synonymes sont délétères et donc moins facilement fixées que les substitutions synonymes, $d_N < d_S$ et $^{d_N}/_{d_S} < 1$. Au contraire, si les organismes subissent une évolution adaptative des protéines par une forte sélection positive, les mutations non synonymes seront favorisées, $d_N > d_S$ et $^{d_N}/_{d_S} > 1$. Ainsi, $^{d_N}/_{d_S}$ donne accès à l'état des pressions de sélection au niveau protéique pour des séquences d'intérêts. Ces pressions peuvent être étudiées soit au niveau des sites des séquences pour trouver, par exemple, des sites sous sélection adaptative, soit au niveau des branches d'un arbre phylogénétique pour détecter des changements de pressions de sélection.

Pour *Prochlorococcus*, Hu et Blanchard (2009) et Dufresne *et al.* (2005) ont montré une accélération des taux d'évolution au fur et à mesure des divergences de *Prochlorococcus*. Cette évolution accélérée des séquences ne semble pas due à un relâchement des pressions de sélection purificatrice, au contraire, car $^{d_N}/_{d_S}$ est plus faible pour les souches réduites de *Prochlorococcus* que pour les souches de *Synechococcus* (Hu et Blanchard, 2009; Yu *et al.*, 2012). Comme le $^{d_N}/_{d_S}$ est inversement proportionnel à $N_e s$ (avec N_e la taille efficace de population et s le coefficient de sélection), la réduction du $^{d_N}/_{d_S}$ pour *Prochlorococcus* semble indiquer soit une augmentation de N_e, soit une augmentation de

s pour *Prochlorococcus* par rapport à *Synechococcus*. La dérive génétique, comme cause de l'évolution réductive chez *Prochlorococcus*, semble donc peu probable et l'évolution accélérée serait alors due à un taux de mutation élevé. Certains gènes du "core-genome" semblent positivement sélectionnés avec $^{d_N}/_{d_S} > 1$ (Paul *et al.*, 2010; Yu *et al.*, 2012), principalement entre les souches de *Prochlorococcus* HLI et les souches non réduites de *Prochlorococcus*. Paul *et al.* (2010) suggèrent ainsi que la sélection positive aurait joué un rôle dans l'adaptation de *Prochlorococcus* peu après la divergence avec *Synechococcus* puis la dérive aurait pris le pas dans la diversification de *Prochlorococcus* et l'évolution réductive. Cette conclusion semble en contradiction avec les valeurs de $^{d_N}/_{d_S}$ plus faibles pour les souches réduites de *Prochlorococcus* que pour les souches de *Synechococcus* (Hu et Blanchard, 2009; Yu *et al.*, 2012; Sun et Blanchard, 2014). Sun et Blanchard (2014) suggèrent une rapide augmentation de la taille de population chez *Prochlorococcus* peu après la divergence avec *Synechococcus*. Les petites tailles de génomes seraient donc le résultat d'une augmentation de la sélection à l'échelle des génomes et non une conséquence d'une niche écologique réduite ou de la sélection relâchée à cause de la dérive.

Cependant, un $^{d_N}/_{d_S}$ plus faible chez *Prochlorococcus* par rapport à *Synechococcus* semble principalement imputable à une augmentation de d_S chez *Prochlorococcus* (Yu *et al.*, 2012). Ceci pourrait résulter d'une augmentation des taux de mutation lié à un relâchement récent des contraintes sélectives sur les motifs d'usage des codons (Morton, 1997). Ainsi, la sélection sur l'usage des codons aurait changé entre *Prochlorococcus* et *Synechococcus*, comme suggéré dans le Chapitre X. Les changements de $^{d_N}/_{d_S}$ pourraient ainsi refléter seulement ces changements de sélection sur l'usage des codons et non des changements de sélection des protéines, rendant l'estimation et l'interprétation de $^{d_N}/_{d_S}$ difficiles. Les méthodes d'estimation du $^{d_N}/_{d_S}$ utilisées jusqu'ici pour *Prochlorococcus* ne permettent pas de discriminer les causes d'un d_S élevé entre un fort taux de mutation, une contrainte sélective relâchée et une sélection positive de l'usage des codons. Il parait ainsi nécessaire d'analyser les différences de $^{d_N}/_{d_S}$ entre les souches de *Prochlorococcus* et de *Synechococcus* en utilisant des modèles où les paramètres mutationnels, de sélection sur l'usage des codons et de sélection sur le contenu protéique sont distincts (Yang et Nielsen, 2008; Pouyet *et al.*, 2013; Michalik, 2014). Ainsi, Juraj Michalik, dans le cadre de son stage de master (Michalik, 2014), a appliqué le modèle de Pouyet *et al.* (2013) à nos souches d'intérêt afin de reconstruire les états ancestraux et d'estimer des indicateurs des différentes pressions de sélection pour les branches de l'arbre phylogénétique de *Prochlorococcus* et de *Synechococcus*. Les analyses ont été effectuées en dehors de ce stage, dans le cadre de ce travail de thèse.

XI.1 Vitesses d'évolution des séquences

D'après Hu et Blanchard (2009) et Dufresne *et al.* (2005), l'évolution des séquences s'est accélérée pour les souches de *Prochlorococcus*, avec les taux d'évolution les plus élevés pour les souches de *Prochlorococcus* dont la divergence est la plus récente (les souches de *Prochlorococcus* HL). L'accélération semble ainsi de plus en plus importante au fur et à

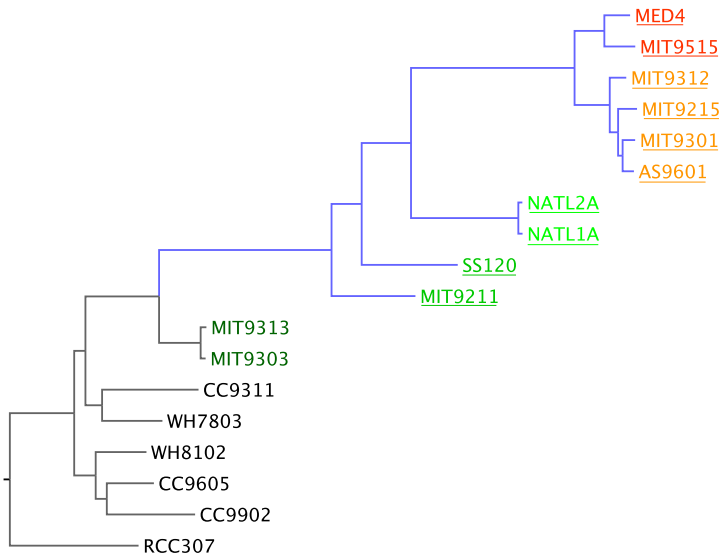
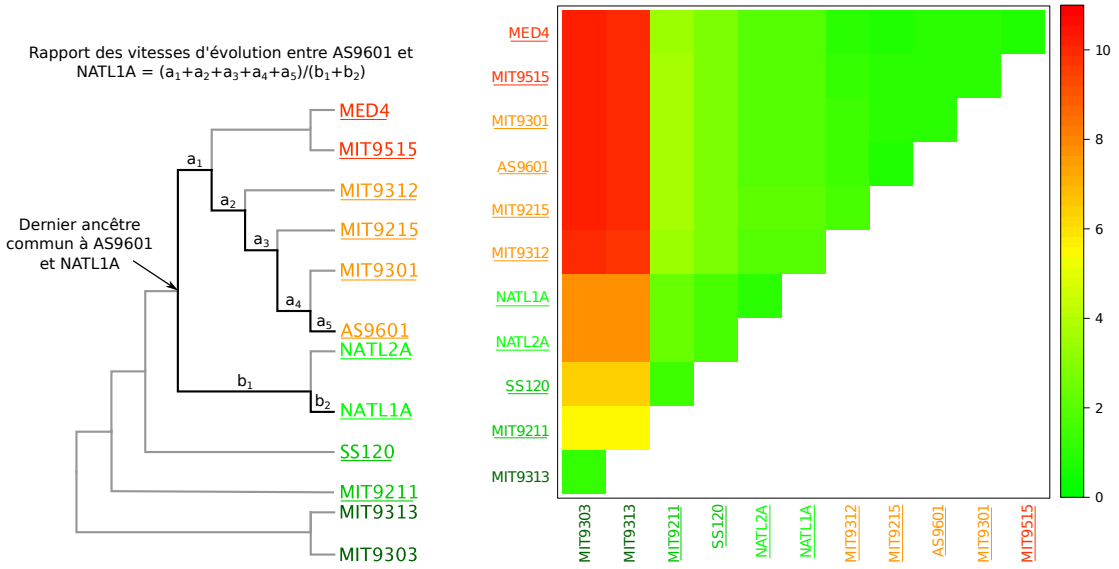


Figure XI.1 – Arbre phylogénétique de souches *Synechococcus* et de *Prochlorococcus*
L'arbre a été construit selon la méthode détaillée dans la section C.4 (Annexe), sur les 693 familles de gènes orthologues.
Les couleurs de noms de souches symbolisent les différentes clades avec en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit.
Les branches en bleu correspondent aux branches conduisant aux souches de *Prochlorococcus*, riches en bases AT et dont le génome est réduit.

mesure des divergences de différentes clades. Cette accélération continue-t-elle au sein des souches de *Prochlorococcus* HL ? Certaines souches évoluent-elles plus vite que les autres au sein des différentes clades ? L'accélération correspondant à la divergence des différentes clades ne serait-elle pas un biais des souches choisies ? Hu et Blanchard (2009) et Dufresne *et al.* (2005) n'ayant pas exploré toutes les souches au sein de chacun des clades, nous comparons les vitesses d'évolution entre les différentes souches de *Prochlorococcus* afin d'identifier les branches où l'évolution des séquences fut la plus rapide et sur lesquelles les analyses des pressions de sélection devront se concentrer.

Les longueurs des branches reflètent le nombre attendu de substitutions par site dans la branche concernée. Certaines branches sont plus longues que d'autres, informant ainsi d'une potentielle accélération de l'évolution des séquences au sein de ces branches ou de temps de divergence plus longs. Au sein de l'arbre de *Prochlorococcus*, des différences de longueurs de branches sont aussi observées (Figure XI.1). Ainsi, la branche ancestrale aux souches réduites de *Prochlorococcus* ou celle ancestrale aux souches de *Prochlorococcus* HL sont relativement longues, par rapport à celle ancestrale aux souches non réduites de *Prochlorococcus*. Ces branches sont-elles plus longues car le temps de divergence est plus long ou car les séquences ont évolué plus rapidement au sein de cette branche ? Comme toutes les branches conduisent à des souches non éteintes, il s'est écoulé la même durée entre un nœud de l'arbre et toutes ses feuilles. Nous pouvons ainsi en déduire



(a) Explication de la mesure utilisée pour comparer les vitesses d'évolution

(b) Comparaison des vitesses d'évolution pour *Prochlorococcus*

Figure XI.2 – Rapport entre les vitesses d'évolution des souches en ligne et des souches en colonne pour *Prochlorococcus*
Chaque case correspond au rapport entre la somme des longueurs de branches depuis l'ancêtre commun de la souche en ligne et de la souche en colonne jusqu'à la souche en ligne et la somme des longueurs de branches depuis l'ancêtre commun à la souche en ligne et la souche en colonne jusqu'à la souche en colonne, comme expliqué dans la Figure XI.2a.
Les couleurs des noms des souches symbolisent les différentes clades avec en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit.

des différences de vitesse d'évolution entre deux souches en remontant jusqu'à l'ancêtre commun le plus récent et en faisant le rapport entre les longueurs des branches conduisant de cet ancêtre commun à chacune des souches (Figure XI.2a).

Ainsi, les souches HL semblent évoluer jusqu'à 10 fois plus vite que les souches LLIV, 3 fois plus vite que les souches LLII/LLIII et 2 fois plus vite que les souches LLI (Figure XI.2b). De même, les souches LLI évoluent plus vite que les souches LLII/LLIII, qui elles-mêmes évoluent plus vite que les souches LLIV. Au sein des souches de *Prochlorococcus* HL, les vitesses d'évolution sont relativement homogènes : les rapports des vitesses d'évolution vont de 1.03 entre MIT9215 et MIT9515 à 1.65 entre MIT9312 et MIT9215 (Figure XI.2b). Ainsi, l'évolution des séquences semblent s'être accélérée au fur et à mesure des divergences, comme suggéré par Hu et Blanchard (2009).

Ces accélérations des taux d'évolution pourraient être dues au cliquet de Muller, bien que peu probable chez *Prochlorococcus*, mais aussi à une augmentation des taux de mutation causée par la perte de gènes de réparation (Marais *et al.*, 2008; Partensky et Garczarek,

2010). Afin de discriminer entre ces hypothèses, des analyses plus poussées des changements dans les séquences doivent être effectuées, en particulier l'estimation des $^{d_N}/_{d_S}$ et des pressions de sélection, en examinant particulièrement les branches ancestrales aux souches réduites de *Prochlorococcus* et aux souches de *Prochlorococcus* HL.

XI.2 Pressions de sélection

Comme mentionné précédemment, les méthodes utilisées jusqu'à présent pour estimer $^{d_N}/_{d_S}$ présentent un certain nombre de défauts qui peuvent conduire à des biais dans les estimations. Ainsi, dans l'analyse de Paul *et al.* (2010), la méthode de Jukes et Cantor (1969) modifiée par Nei et Gojobori (1986) ne prend en compte ni les différences entre les taux de transition et de transversion[1], ni le biais d'usage des codons. Or ce dernier n'est pas le même selon les souches étudiées (Chapitre X). Les analyses de Hu et Blanchard (2009) et Yu *et al.* (2012) sont basées sur la méthode de Yang et Nielsen (1998, 2000), qui prend en compte à la fois la structure du code génétique et le biais entre les taux de transition et de transversion. Cependant, cette méthode ne distingue pas les biais mutationnels des pressions de sélection sur l'usage des codons. Or, dans le cas de *Prochlorococcus*, les changements de biais mutationnels observés peuvent entrainer une surestimation de d_S.

La méthode de Yang et Nielsen (2008) peut pallier ce problème en modélisant explicitement le processus de substitution d'un codon par un autre, c'est-à-dire les mutations, la sélection sur l'usage des codons et la sélection sur la protéine par l'introduction de paramètres de préférence de chacun des codons. Dans une amélioration de ce modèle proposée par Pouyet *et al.* (2013), les paramètres d'usage des codons et les paramètres d'usage des acides aminés sont complètement séparés afin de pouvoir mesurer les différentes pressions indépendamment les unes des autres. Le modèle repose alors sur trois couches distinctes : une couche nucléotidique pour l'influence du déséquilibre des fréquences nucléotidiques, une couche codon où chaque codon possède une préférence se basant sur les forces sélectives de biais d'usage des codons et une couche acide aminé pour l'influence des pression de sélection et des mutations sur les changements de fréquences des acides aminés. Dans ce modèle basé sur les codons, les taux de substitutions entre le codon i et le codon j sont calculés selon la formule suivante (Pouyet *et al.*, 2013) :

[1]Les transitions sont des substitutions entre deux purines (A $\leftrightarrow$ G) ou entre deux pyrimidines (C $\leftrightarrow$ T), alors qu'une substitution entre une purine et une pyrimidine est appelée transversion (C, T $\leftrightarrow$ A, G).

$$Q_{i\to j} = \begin{cases} 0 & \text{si les codons diffèrent de plus d'un nucléotide} \\ \pi_{j_p} \cdot \frac{-log\left(\frac{\phi_{i|aa_i}}{\phi_{j|aa_j}}\right)}{1-\frac{\phi_{i|aa_i}}{\phi_{j|aa_j}}} & \text{si les codons diffèrent par une transversion synonyme au site } p \\ \kappa \cdot \pi_{j_p} \cdot \frac{-log\left(\frac{\phi_{i|aa_i}}{\phi_{j|aa_j}}\right)}{1-\frac{\phi_{i|aa_i}}{\phi_{j|aa_j}}} & \text{si les codons diffèrent par une transition synonyme au site } p \\ \omega \cdot \pi_{j_p} \cdot \frac{-log\left(\frac{\psi_{aa_i}\cdot\phi_{i|aa_i}}{\psi_{aa_j}\cdot\phi_{j|aa_j}}\right)}{1-\frac{\psi_{aa_i}\cdot\phi_{i|aa_i}}{\psi_{aa_j}\cdot\phi_{j|aa_j}}} & \text{si les codons diffèrent par une transversion non synonyme au site } p \\ \kappa \cdot \omega \cdot \pi_{j_p} \cdot \frac{-log\left(\frac{\psi_{aa_i}\cdot\phi_{i|aa_i}}{\psi_{aa_j}\cdot\phi_{j|aa_j}}\right)}{1-\frac{\psi_{aa_i}\cdot\phi_{i|aa_i}}{\psi_{aa_j}\cdot\phi_{j|aa_j}}} & \text{si les codons diffèrent par une transition non synonyme au site } p \end{cases} \quad \text{(XI.1)}$$

avec π_{j_p} la fréquence d'équilibre du nucléotide j_p, κ le rapport entre les taux de transition et de transversion, $\omega = {}^{d_N}/_{d_S}$, ψ_{aa_i} la préférence de l'acide aminé correspondant au codon i par rapport aux autres acides aminés et $\phi_{i|aa_i}$ la préférence du codon i par rapport aux codons synonymes de l'acide aminé. Ainsi, la couche nucléotidique est représentée par κ et π_{j_p}, la couche codon par $\phi_{i|aa_i}$ et la couche acide aminé par ψ_{aa_i}.

Ce modèle, implémenté dans la librairie *bpp-phyl* de la suite de programme *Bio++* (Guéguen *et al.*, 2013), a été appliqué aux 693 familles de gènes orthologues (Section C.1.2, en annexe) des souches de *Prochlorococcus* et de *Synechococcus*. Les familles ont été regroupées en sept groupes de 99 familles chacun selon la valeur du F_{op}[1] dans la souche *Synechococcus* CC93311. Les alignements des familles de gènes au sein des groupes sont concaténés pour obtenir sept concaténats, où le nombre de sites assure une quantité suffisante d'informations pour ajuster correctement les modèles, mais pas trop pour permettre au modèle de converger. Étant donné la différence de contenu en bases GC et d'usage des codons (Chapitre X), utiliser un seul modèle pour l'ensemble des souches pourrait entrainer des estimations trop grossières des différents paramètres. Deux modèles sont ainsi ajustés aux données : le modèle 1 englobe les souches riches en GC et les branches correspondantes (branches en gris dans la figure XI.1) alors que le modèle 2 est ajusté sur les souches riches en AT (branches en bleu dans la figure XI.1). Comme les souches riches en AT correspondent aussi aux souches réduites, le modèle 1 concerne les souches non réduites et le modèle 2 les souches réduites (Michalik, 2014). À partir des modèles inférés, des arbres et des alignements, les séquences ancestrales les plus probables *a posteriori* sont reconstruites grâce à *bpp-ancestor* de *Bio++* (Guéguen *et al.*, 2013). Afin d'estimer le ${}^{d_N}/_{d_S}$, le rapport entre le taux de transition et transversion et d'autres indicateurs pour chaque branche de l'arbre phylogénétique, une méthode de comptage probabiliste des substitutions est appliquée aux données à l'aide de l'outil *MapNH* de *Bio++* (Guéguen *et al.*,

[1] La valeur de F_{op} d'un gène correspond à la fréquence d'utilisation des codons optimaux dans le gène. Dans cette analyse, les codons optimaux utilisés ne correspondent pas à ceux présentés dans la section X.4, car ces derniers n'avaient pas été déterminés de façon fiable. Les codons optimaux, utilisés dans cette analyse, sont les codons utilisés préférentiellement dans les gènes ribosomaux, avec un codon par acide aminé.

2013). Les résultats présentés dans la suite sont les moyennes pour les sept concaténats de gènes.

XI.2.1 Équilibre des modèles

Le modèle de Pouyet *et al.* (2013) est irréductible mais les processus évolutifs atteignent une distribution stationnaire ou d'équilibre. Dans cet état, la fréquence d'équilibre d'un codon sans sélection au niveau protéique est :

$$\pi_c = \begin{cases} \prod_{p=1}^{3} \pi_{c_p} \cdot \phi_{c|aa_c} \cdot \psi_{aa_c} & \text{pour le modèle complet (3 couches)} \\ \prod_{p=1}^{3} \pi_{c_p} & \text{pour la couche nucléotidique (biais nucléotidique)} \\ \phi_{c|aa_c} & \text{pour la couche codon (sélection sur l'usage des codons)} \\ \psi_{aa_c} & \text{pour la couche acide aminé (sélection sur l'usage des acides aminés)} \end{cases} \tag{XI.2}$$

À partir de ces fréquences d'équilibre, nous pouvons calculer la composition nucléotidique des deux modèles aux différentes positions des codons, en différenciant les trois couches pour en déduire le contenu en bases GC et GT (Figure XI.3).

Les taux de GC aux différentes positions des codons sont similaires pour la couche nucléotidique car cette couche met en avant les biais nucléotidiques avant filtrage de la sélection sur l'usage des codons et la sélection au niveau protéique. Dans le modèle 1, le biais est orienté vers un enrichissement en bases GC (GC $\simeq$ 65%) alors que dans le modèle 2, il est orienté vers un enrichissement en bases AT (GC $\simeq$ 23%) (Figure XI.3), correspondant ainsi à la définition des modèles (modèle 1 pour les branches conduisant aux souches riches en GC et modèle 2 pour les branches conduisant aux souches riches en AT). La composition en bases GT, potentiellement liée aux brins, ne semble pas biaisée : les valeurs de GT sont d'environ 47% pour le modèle 1 et d'environ 42% pour le modèle 2.

Pour le modèle 2, le biais vers AT semble être la principale force influençant la composition en codons des séquences. En effet, les valeurs d'équilibre de GC pour le modèle complet sont plus proches de celles de la couche nucléotidique que des autres couches, ces dernières ayant des valeurs nettement supérieures (Figure XI.3). Au contraire dans le modèle 1, les valeurs d'équilibre de GC pour le modèle complet sont plus proches de celles de la couche codon, sauf pour GC_3 qui est principalement dirigé par la couche nucléotidique et le GC_{12} qui semble être composite de la couche codon et de la couche acides aminés. Ainsi, le biais nucléotidique vers un enrichissement en bases GC semble contré par la sélection sur l'usage des codons mais aussi par la sélection sur les protéines, même si celle-ci est plus faible.

Pour la couche codon, les valeurs de GC sont relativement similaires pour les deux modèles ($\simeq$ 45%, Figure XI.3), alors que les souches riches en AT et riches en GC n'utilisent

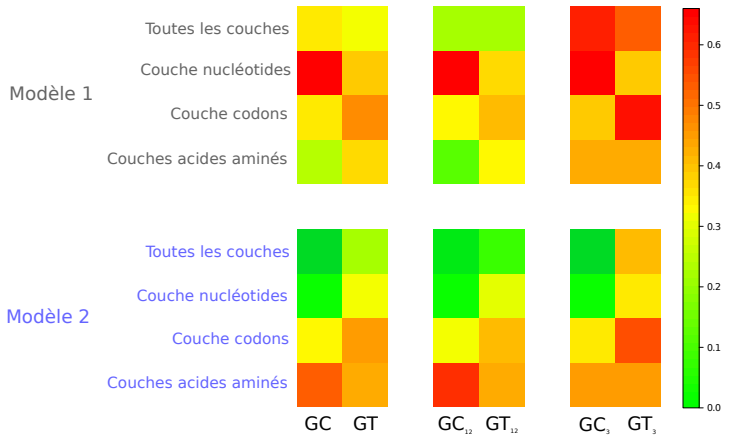


Figure XI.3 – Contenu en bases GC et GT à l'équilibre pour les deux modèles appliqués aux souches de *Prochlorococcus* et de *Synechococcus*
Le contenu en GC et GT à l'équilibre est calculé à partir des fréquences d'équilibre des codons pour les différentes couches des modèles calculées selon l'équation XI.2.
Le modèle 1 correspond au modèle ajusté sur les branches conduisant aux souches riches en GC (non réduites) et le modèle 2 au modèle ajusté sur les branches conduisant aux souches riches en AT (réduites).

pas les mêmes codons synonymes (Figure X.10). Les deux modèles semblent ainsi avoir convergé vers l'utilisation de codons synonymes à mi-chemin entre les valeurs de biais nucléotidiques. Est-ce l'équilibre souhaité pour avoir une utilisation optimale de la machinerie de traduction bactérienne, quelque soit le contenu en GC ? Cet équilibre semble atteint par le modèle 1.

Dans la couche acide aminé du modèle 2, le taux de GC est relativement élevé ($\simeq 58\%$, Figure XI.3) par rapport aux valeurs de GC des souches. Ainsi, les acides aminés préférés pour le modèle 2 par rapport au modèle 1 sont plus riches en GC que ceux préférés par le modèle 1. Le fort enrichissement observé au niveau nucléotidique et codon est compensé, au moins en partie, au niveau de la couche acide aminé, permettant ainsi de conserver un certain contenu en acides aminés malgré les modifications observées (Figure X.8).

La sélection sur l'usage des codons semble jouer un rôle dans le modèle 1 mais peu dans le modèle 2. Il serait ainsi intéressant d'avoir des mesures de la force de la sélection sur l'usage des codons. Comme le modèle de Pouyet *et al.* (2013) est basé sur celui de Yang et Nielsen (2008), nous nous inspirons de mesures utilisées par Yang et Nielsen (2008) en les adaptant.

La mutation du codon I au codon J, qui change le nucléotide i_k en j_k à la position k du codon, a lieu à un taux $\mu_{i_k j_k}$, avec une "fitness" de :

$$S_{IJ} = \begin{cases} -log\left(\frac{\phi_{I|aa_I}}{\phi_{J|aa_J}}\right) & \text{pour une mutation synonyme} \\ -log\left(\frac{\psi_{aa_I}\cdot\phi_{I|aa_I}}{\psi_{aa_J}\cdot\phi_{J|aa_J}}\right) & \text{pour une mutation non synonyme} \end{cases} \quad \text{(XI.3)}$$

À l'équilibre, la proportion de mutations $I \to J$ parmi toutes les mutations est :

$$m_{IJ} = \frac{\pi_I \cdot \mu_{i_k j_k}}{\sum_{M \neq L} \pi_M \cdot \mu_{m_k l_k}} \text{ avec } \mu_{i_k j_k} = \begin{cases} \pi_{j_k} & \text{pour une transversion} \\ \kappa \cdot \pi_{j_k} & \text{pour une transition} \end{cases} \tag{XI.4}$$

Nous pouvons ainsi en déduire la proportion de mutations avantageuses parmi toutes les mutations :

$$P_+ = \sum_{I \neq J} m_{IJ} \mathbb{I}_{S_{IJ}>0} \text{ avec } \mathbb{I}_{S_{IJ}>0} = \begin{cases} 1 & \text{si } S_{IJ} > 0 \\ 0 & \text{sinon} \end{cases} \tag{XI.5}$$

La proportion de mutations délétères parmi toutes les mutations est de 77% pour le modèle 1 et de 65% pour le modèle 2. Ainsi, dans les deux cas, la plupart des mutations sont délétères pour l'usage des codons.

La force de la sélection positive sur une mutation avantageuse moyenne peut être ensuite mesurée :

$$\overline{S}_+ = \sum_{I \neq J} \frac{\pi_I \mu_{i_k j_k} \mathbb{I}_{S_{IJ}>0}}{\sum_{I \neq J} \pi_I \mu_{i_k j_k} \mathbb{I}_{S_{IJ}>0}} S_{IJ} \mathbb{I}_{S_{IJ}>0} \tag{XI.6}$$

ainsi que la force de la sélection purificatrice sur une mutation délétère moyenne avec le même principe que pour l'équation XI.6 avec $\mathbb{I}_{S_{IJ}<0} = \begin{cases} 1 & \text{si } S_{IJ} < 0 \\ 0 & \text{sinon} \end{cases}$

La force de la sélection positive sur une mutation avantageuse moyenne modifiant un codon est plus élevée pour le modèle 1 ($\overline{S}_+ = 0.77$) que pour le modèle 2 ($\overline{S}_+ = 0.51$), de même pour la sélection négative (modèle 1 :$\overline{S}_- = -1.83$, modèle 2 :$\overline{S}_- = -0.79$). En valeur absolue, la sélection purificatrice sur l'usage des codons est plus forte que la sélection positive pour les deux modèles et quelque soit le type de sélection, elle est plus forte pour le modèle 1. Ces résultats sont cohérents avec la prédominance de la couche codon dans le modèle 1 et pas dans le modèle 2 : les mutations avantageuses de codons sont ainsi plus favorisées dans le modèle 1 par rapport à celles dans le modèle 2 et les mutations délétères plus sous contraintes dans le modèle 1. La sélection sur l'usage des codons dans le modèle 2 est donc plus faible que dans le modèle 1. De plus, en accord avec les valeurs de $^{d_N}/_{d_S}$ globalement inférieures à 1 dans les analyses précédentes (Hu et Blanchard, 2009; Yu *et al.*, 2012), la sélection positive est moins forte que la sélection purificatrice.

XI.2.2 Transitions, transversions et évolution du contenu en GC le long de la phylogénie

Les transitions, substitutions entre deux purines (A $\leftrightarrow$ G), sont souvent plus fréquentes que les transversions, substitutions entre une purine et une pyrimidine(C, T $\leftrightarrow$ A, G). De fait, le rapport (κ) entre le taux de transition (T_s) et le taux de transversion (T_v) est supérieur à 1 pour les deux modèles à l'équilibre avec une valeur plus élevée pour le modèle 2 ($\kappa = 1.64$) que pour le modèle 1 ($\kappa = 1.29$). Cette différence est significative lorsque nous comparons les valeurs de κ pour les branches du modèle 2 et les branches du modèle 1 (Figure XI.4, $P = 2.624 \cdot 10^{-7}$, test de Mann-Whitney), principalement car les valeurs de κ pour les branches conduisant aux souches de *Prochlorococcus* sont supérieures aux κ pour les branches conduisant aux souches de *Synechococcus* ($P = 8.389 \cdot 10^{-8}$, test de Mann-Witney). Le long de la phylogénie de *Prochlorococcus*, κ augmente avec la divergence des différents clades jusqu'à atteindre une valeur maximale pour la branche ancestrale aux souches MIT9301 et AS9601 (Figure XI.4). Cependant, pour de nombreuses branches, les valeurs brutes de taux de transition et de taux de transversion sont relativement faibles (Figure XI.4) et le ratio κ entre les deux taux exacerbe des différences relativement faibles, comme pour la branche ancestrale aux souches MIT9301 et AS9601.

Le long de la branche ancestrale aux souches de *Prochlorococcus* HL, les taux de transitions et de transversions sont plus élevés, supérieurs à ceux de la branche ancestrale à l'ensemble des souches réduites de *Prochlorococcus*. Or, c'est le long de cette dernière qu'ont lieu les principaux changements associés à l'enrichissement en bases AT avec une augmentation du contenu en bases AT par rapport à la branche parente et plus de substitutions GC $\rightarrow$ AT que de substitutions AT $\rightarrow$ GC (Figure XI.5). Dans les branches filles, c'est-à-dire les branches conduisant à la diversification des souches de *Prochlorococcus*, l'enrichissement est moindre, avec même parfois moins de substitutions GC $\rightarrow$ AT que de substitutions AT $\rightarrow$ GC (Figure XI.5), alors que la composition en bases des souches est plus riche en GC que l'équilibre du modèle et devrait tendre vers celui-ci. L'effet est contraire dans les branches du modèle 1, à l'exception de la branche ancestrale aux souches non réduites de *Prochlorococcus*, expliquant ainsi les rapports entre le nombre de substitutions GC $\rightarrow$ AT et le nombre de substitutions AT $\rightarrow$ GC supérieurs à 1. Les souches riches en GC évoluent vers un contenu riche en base AT, mais cet enrichissement en AT semble se stabiliser autour d'un contenu plus riche en GC que l'équilibre attendu.

XI.2.3 Usage des codons et acides aminés

L'usage des codons synonymes a changé au cours de l'évolution de *Prochlorococcus* (Chapitre X), en particulier avec l'évolution réductive. Ces changements ont principalement eu lieu le long de la branche ancestrale aux souches réduites de *Prochlorococcus* (Figure XI.6), mais aussi, par ordre décroissant, dans la branche ancestrale aux souches non réduites de *Prochlorococcus*, la branche ancestrale à l'ensemble des souches de *Prochlorococcus* et la branche ancestrale aux souches de *Prochlorococcus* HL. Ainsi, les changements de fré-

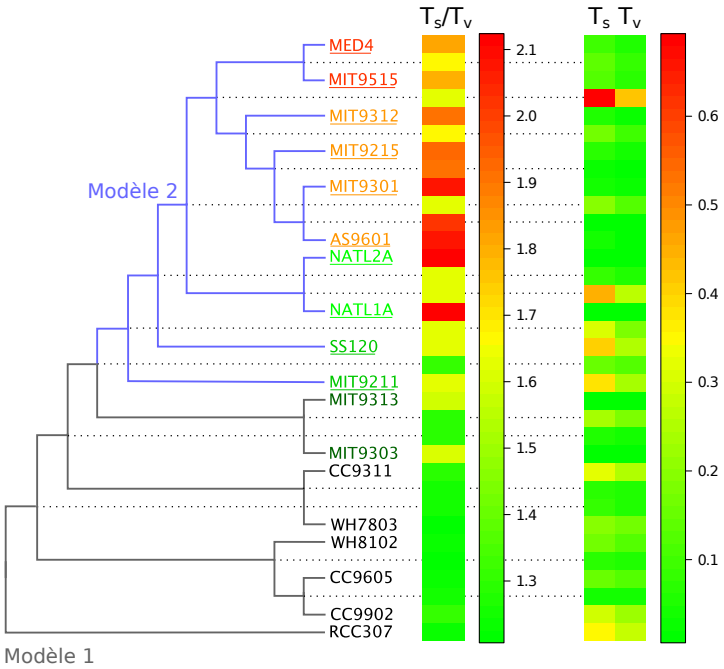


Figure XI.4 – Taux de transition (T_s) et taux de transversion (T_v) le long de la phylogénie de *Prochlorococcus* et de *Synechococcus*
Les valeurs des taux ont été inférées pour chacune des branches par une méthode de comptage.
Deux modèles ont été ajustés aux branches de la phylogénie de *Prochlorococcus* et de *Synechococcus* : un pour les branches conduisant aux souches riches en GC (branches en gris) et un pour les branches conduisant aux souches riches en AT et réduites (branches en bleu).
Les couleurs de noms des souches symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit. L'arbre phylogénétique est construit comme décrit dans la section C.4 (Annexe) mais les longueurs des branches sont arbitraires et ne reflètent pas les taux d'évolution.

quence d'usage des codons ne touchent pas seulement les souches réduites mais toutes les souches de *Prochlorococcus*, avec des changements similaires. Les codons terminant par les bases A/T augmentent en fréquence au détriment des codons terminant par G/C, ce qui est cohérent avec l'enrichissement en bases AT (Figure XI.5).

Le long de la branche ancestrale aux souches HL, la proportion de bases AT augmente par rapport à la branche ancestrale immédiate (Figure XI.5) mais principalement par des changements en première base des codons, contrairement à la branche ancestrale aux souches réduites où les changements ont principalement lieu en troisième base des codons. Ainsi, les principaux changements de codons synonymes le long de la branche ancestrale aux souches HL sont des changements au sein d'acides aminés dégénérés six fois comme l'arginine avec une augmentation de la fréquence du codon AGA au détriment de CGA (Figure XI.6). Les changements de codons synonymes au sein des autres acides aminés sont réduits. A ce stade, la plupart des changements possibles de codons synonymes au

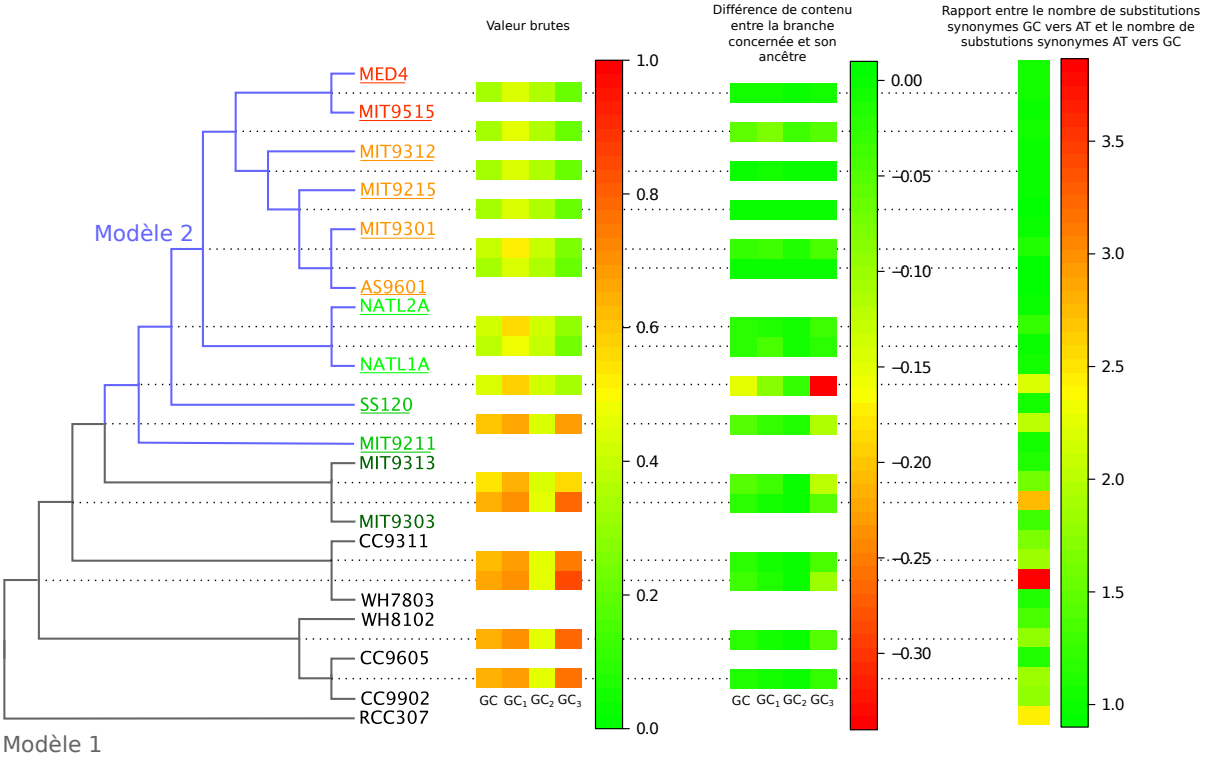


Figure XI.5 – Contenu en GC, différences par rapport aux branches ancestrales et rapport entre le nombre de substitutions de GC vers AT et le nombre de substitutions de AT vers GC le long de la phylogénie de *Prochlorococcus* et de *Synechococcus*

Le contenu en bases GC le long des branches de la phylogénie est calculé à partir des fréquences des codons inférées par la reconstruction des états ancestraux avec *bpp-ancestor*. Le contenu en GC aux trois positions des codons d'une branche est comparé au contenu de la branche ancestrale immédiate. Les rapports entre le nombre de substitutions de GC vers AT et le nombre de substitutions de AT vers GC ont été inférés pour chacune des branches par une méthode de comptage.

Deux modèles ont été ajustés aux branches de la phylogénie de *Prochlorococcus* et de *Synechococcus* : un pour les branches conduisant aux souches riches en GC (branches en gris) et un pour les branches conduisant aux souches riches en AT et réduites (branches en bleu).

Les couleurs de noms des souches symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit. L'arbre phylogénétique est construit comme décrit dans la section C.4 (Annexe) mais les longueurs des branches sont arbitraires et ne reflètent pas les taux d'évolution.

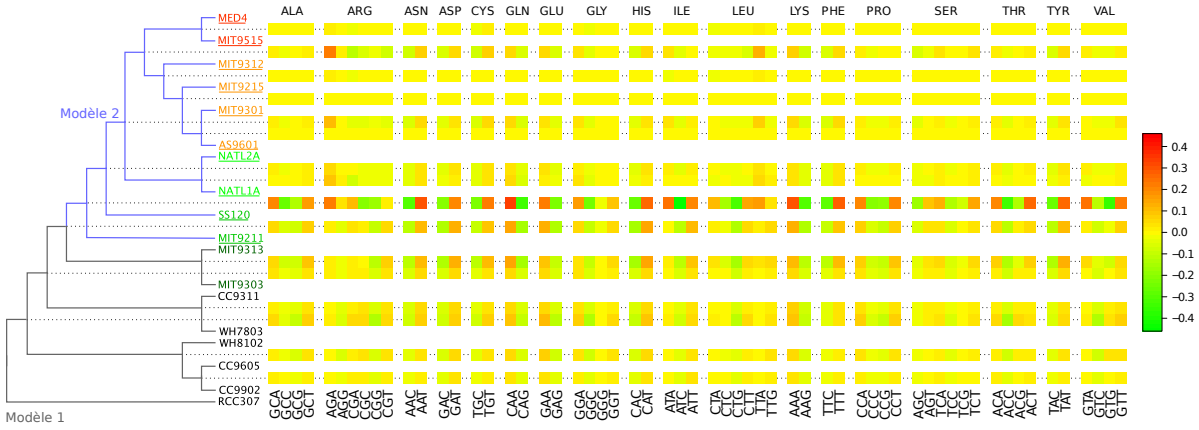


Figure XI.6 – Différence d'usage des codons synonymes au sein de chaque acide aminé par rapport à la branche ancestrale immédiate pour les branches de la phylogénie de *Prochlorococcus* et de *Synechococcus*
Les fréquences des codons dans les différentes branches sont inférées par la reconstruction des états ancestraux avec *bpp-ancestor* puis comparées aux fréquences de la branche ancestrale immédiate.
Deux modèles ont été ajustés aux branches de la phylogénie de *Prochlorococcus* et de *Synechococcus* : un pour les branches conduisant aux souches riches en GC (branches en gris) et un pour les branches conduisant aux souches riches en AT et réduites (branches en bleu).
Les couleurs de noms des souches symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit. L'arbre phylogénétique est construit comme décrit dans la section C.4 (Annexe) mais les longueurs des branches sont arbitraires et ne reflètent pas les taux d'évolution.

sein des acides aminés dégénérés deux et quatre fois pourraient avoir été effectués et tout changement supplémentaire serait trop délétère (traduction moins efficace, ...). Or, l'enrichissement en bases AT continue. Les changements de codons sont alors plus "risqués" avec des changements en première base des codons mais aussi avec des changements d'acides aminés (Figure XI.7).

Les changements de fréquences d'acides aminés ont ainsi principalement lieu le long de la branche ancestrale aux souches HL puis le long de celle ancestrale aux souches réduites, celle ancestrale aux souches LLI, celle ancestrale aux souches non réduites de *Prochlorococcus* puis celle ancestrale à l'ensemble des souches de *Prochlorococcus* (Figure XI.7). Le long de la branche conduisant aux souches HL, seuls six acides aminés augmentent en fréquence, mais de façon importante, et cinq d'entre eux sont des acides aminés dont les premières bases sont A et/ou T (Figures XI.7, X.8). Les acides aminés commençant par des bases G et/ou C font parti des acides aminés dont la fréquence diminue le plus. La tendance est similaire pour la branche ancestrale aux souches réduites, mais change légèrement pour les autres branches avec des changements plus ou moins importants. Les acides aminés commençant par A et/ou T augmentent en fréquence (à l'exception de la tyrosine le long de la branche ancestrale aux souches de *Prochlorococcus*), mais la sérine

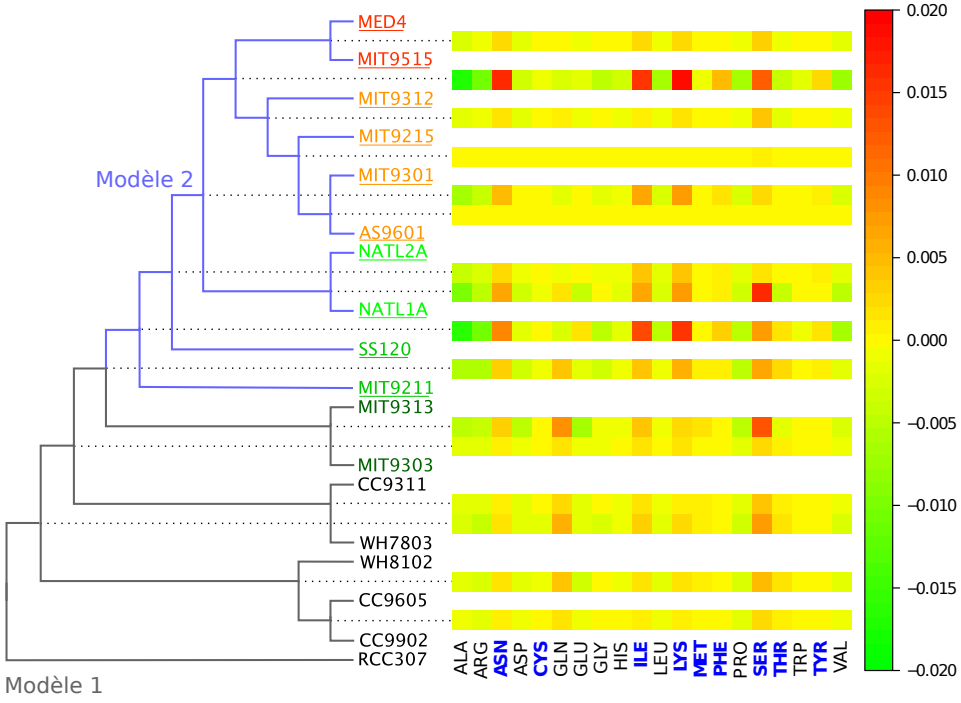


Figure XI.7 – Différence d'usage des acides aminés par rapport à la branche ancestrale immédiate pour les branches de la phylogénie de *Prochlorococcus* et de *Synechococcus*
Les fréquences des codons dans les différentes branches sont inférées par la reconstruction des états ancestraux avec *bpp-ancestor*. À partir de ces fréquences, les fréquences des acides aminés sont calculées pour chacune des branches puis comparées aux fréquences de la branche ancestrale directe. Les acides aminés en bleu et en gras sont ceux dont les codons démarrent par A et/ou T.
Deux modèles ont été ajustés aux branches de la phylogénie de *Prochlorococcus* et de *Synechococcus* : un pour les branches conduisant aux souches riches en GC (branches en gris) et un pour les branches conduisant aux souches riches en AT et réduites (branches en bleu).
Les couleurs de noms des souches symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit. L'arbre phylogénétique est construit comme décrit dans la section C.4 (Annexe) mais les longueurs des branches sont arbitraires et ne reflètent pas les taux d'évolution.

est l'acide aminé dont la fréquence augmente le plus. Ainsi, l'enrichissement en bases AT touche aussi les fréquences des acides aminés en favorisant ceux enrichis en bases AT, mais ces changements semblent plus tardifs que les changements de codons synonymes, comme s'il fallait atteindre un point où les principaux changements de codons synonymes ont été effectués et où l'enrichissement en bases AT ne peut continuer qu'en changeant les acides aminés.

XI.2.4 Ratio $^{d_N}/_{d_S}$

Le modèle de Yang et Nielsen (1998, 2000), utilisé sur les précédentes analyses du $^{d_N}/_{d_S}$ (Hu et Blanchard, 2009; Yu *et al.*, 2012), est appliqué au jeu de données, pour comparer les observations à celles faites avec le modèle de Yang et Nielsen (2008), amélioré par Pouyet *et al.* (2013). Les résultats sont cohérents avec les observations précédentes (Hu et Blanchard, 2009; Yu *et al.*, 2012). En effet, $^{d_N}/_{d_S}$ est plus faible pour les souches réduites de *Prochlorococcus* que pour les souches non réduites ($P = 7.543 \cdot 10^{-9}$, test de Mann-Whitney) (Figure XI.8). Dans ce modèle, les taux de substitutions entre le codon i et le codon j sont calculés selon la formule suivante Yang et Nielsen (2000, 1998) :

$$Q_{i \to j} = \begin{cases} 0 & \text{si les codons diffèrent de plus d'un nucléotide} \\ \pi_{j_p} & \text{si les codons diffèrent par une transversion synonyme au site } p \\ \kappa \cdot \pi_{j_p} & \text{si les codons diffèrent par une transition synonyme au site } p \\ \omega \cdot \pi_{j_p} & \text{si les codons diffèrent par une transversion non synonyme au site } p \\ \kappa \cdot \omega \cdot \pi_{j_p} & \text{si les codons diffèrent par une transition non synonyme au site } p \end{cases} \quad \text{(XI.7)}$$

avec π_{j_p} la fréquence d'équilibre du nucléotide j_p, κ le rapport entre le taux de transition et transversion, $\omega = {}^{d_N}/_{d_S}$.

Ce modèle (YN98) est ainsi plus simple que celui de Yang et Nielsen (2008), amélioré par Pouyet *et al.* (2013) (YN08*) (Equation XI.1) et ne distingue pas l'effet des mutations de celui de la sélection mais aussi l'effet de la sélection sur l'usage des codons de la sélection sur les protéines. $^{d_N}/_{d_S}$ est ainsi composite : $\omega(YN98) = \omega(YN08^*) \cdot {}^{-log\left(\psi_{aa_i} \cdot \phi_{i|aa_i} / \psi_{aa_j} \cdot \phi_{j|aa_j}\right)}/{}_{\left(1 - \psi_{aa_i} \cdot \phi_{i|aa_i} / \psi_{aa_j} \cdot \phi_{j|aa_j}\right)}$ avec ψ_{aa_i} la préférence de l'acide aminé de i par rapport aux autres acides aminés et $\phi_{i|aa_i}$ la préférence du codon i par rapport aux codons synonymes de l'acide aminé. $\omega(YN08^*)$ représente, plus clairement que $\omega(YN98)$, l'effet de la sélection naturelle sur les protéines, indépendamment de la sélection sur l'usage des codons.

$^{d_N}/_{d_S}$ de YN08* est plus homogène que $^{d_N}/_{d_S}$ de YN98 : les différences entre les branches conduisant aux souches réduites et les autres branches se sont effacées (Figure XI.8, $P = 0.7843$, test de Mann-Whitney). En effet, $^{d_N}/_{d_S}$ pour les branches du modèle 2 est sous-estimé avec YN98 par rapport à YN08* : d_N sous-estimé et d_S surestimé (Figure XI.8). De fait, les changements d'usage des codons liés à l'enrichissement en bases AT sont inclus

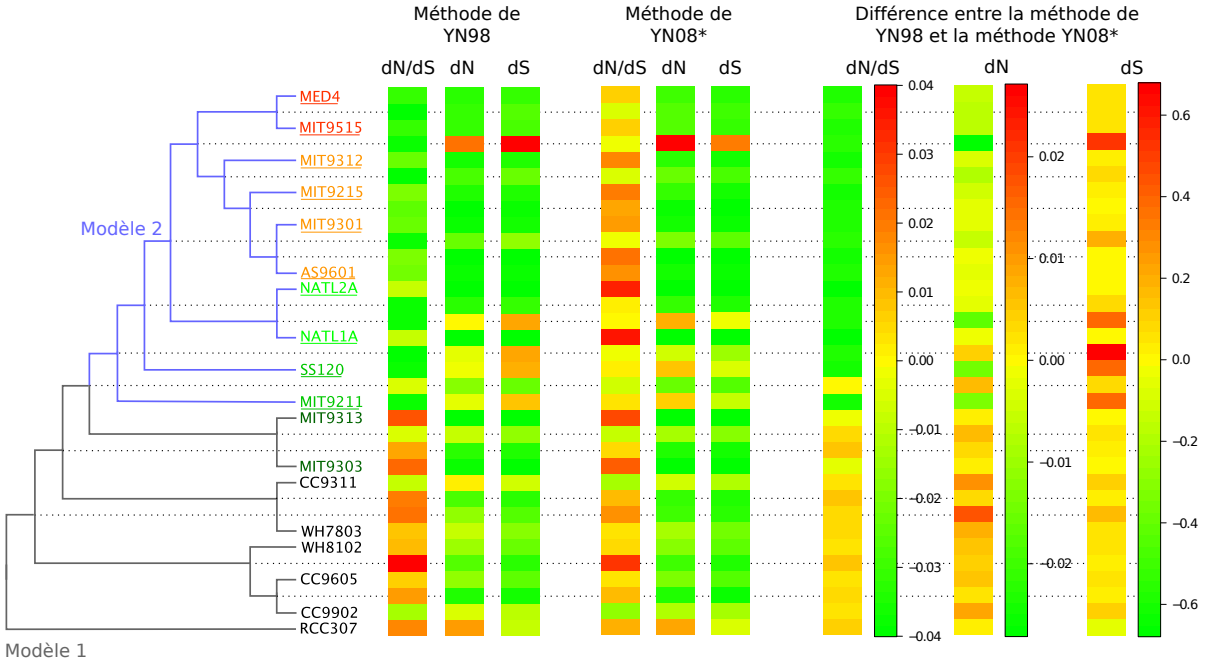


Figure XI.8 – d_N/d_S pour les différentes branches de la phylogénie de *Prochlorococcus* et de *Synechococcus*

Deux méthodes ont été testées pour comparaison : la méthode de Yang et Nielsen (2000, 1998) (YN98) et la méthode de Yang et Nielsen (2008) améliorée par Pouyet *et al.* (2013) (YN08*). Les valeurs de d_N/d_S, d_N et d_S ont été inférées pour chacune des branches par une méthode de comptage. Deux modèles ont été ajustés aux branches de la phylogénie de *Prochlorococcus* et de *Synechococcus* : un pour les branches conduisant aux souches riches en GC (en gris) et un pour les branches conduisant aux souches riches en AT et réduites (en bleu).

Par souci de lisibilité, seules les échelles de valeurs de comparaison entre les deux méthodes sont représentées. Pour les autres, elles sont identiques pour chacun des indicateurs. Ainsi, l'échelle utilisée pour d_N/d_S de la méthode YN98 est la même que d_N/d_S de la méthode YN08 et couvre les valeurs de 0.04 à 0.11 linéairement. Pour d_N, les valeurs vont de 0 à 0.12 et pour d_S de 0.03 à 2.03.

Les couleurs de noms des souches symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit. L'arbre phylogénétique est construit comme décrit dans la section C.4 (Annexe) mais les longueurs des branches sont arbitraires et ne reflètent pas les taux d'évolution.

dans d_S pour YN98 mais surtout dans le changement de préférence des codons pour YN08*. De plus, d_S de YN08* ne comptabilise que les changements de codons synonymes qui sont en contradiction avec les préférences des codons au sein des acides aminés.

Ainsi, contrairement à ce qui a pu être suggéré par les analyses précédentes (Hu et Blanchard, 2009; Yu *et al.*, 2012), les pressions de sélection sur les protéines ne semblent pas avoir changé avec l'évolution réductive. Les valeurs extrêmes de $^{d_N}/_{d_S}$ se retrouvent dans les branches externes avec le minimum pour *Synechococcus* CC9902 et le maximum pour *Prochlorococcus* NATL1A. Dans les branches internes, $^{d_N}/_{d_S}$ est minimal pour la branche ancestrale aux souches non réduites de *Prochlorococcus* et maximal pour la branche ancestrale à trois souches de *Synechococcus*. Enfin, les $^{d_N}/_{d_S}$ des branches internes du modèle 1 et du modèle 2 sont similaires (Figure XI.8, $P = 0.536$, test de Mann-Whitney).

Comme pour $^{T_s}/_{T_v}$ (Figure XI.4), $^{d_N}/_{d_S}$ de la branche ancestrale aux souches de *Prochlorococcus* HL est similaire à celui des autres branches alors que les valeurs d_N et d_S sont nettement supérieures, particulièrement par rapport aux valeurs de la branche ancestrale aux souches réduites (Figure XI.8). Ainsi, le long de cette branche, les évènements de substitutions non synonymes et synonymes sont nombreux, potentiellement liés à une augmentation des taux de mutation ou à l'adaptation à l'environnement différent qu'est le haut de la colonne d'eau.

XI.3 Discussion

Comme chez les endosymbiotes, les séquences des souches réduites de *Prochlorococcus* évoluent plus rapidement que les séquences des souches non réduites. Cependant, cette évolution accélérée ne semble pas être la conséquence d'un relâchement des pressions de sélection ou de sélections plus fortes pour l'adaptation des protéines. En effet, les valeurs de $^{d_N}/_{d_S}$ sont relativement constantes le long des branches internes de l'arbre phylogénétique de *Prochlorococcus*, malgré quelques modifications. Ainsi, contrairement à ce qui est observé dans les analyses précédentes (Hu et Blanchard, 2009; Yu *et al.*, 2012), les pressions de sélection sur les protéines ne semblent pas avoir changé au cours de l'évolution réductive tout comme les tailles efficaces de populations, ou alors d'une façon telle que les changements de pression de sélection et de taille efficace de population se compensent pour conserver une valeur de $^{d_N}/_{d_S}$ constant. La réduction des génomes et les autres caractéristiques de l'évolution réductive ne peuvent donc pas être imputables à des changements de pressions de sélection ou à des modifications de tailles efficaces de population dues à des modes de vie différents, comme pour les endosymbiotes.

Cependant, en élargissant l'analyse, le détail des différents indicateurs montre des changements le long des branches conduisant aux souches de *Prochlorococcus* et principalement aux souches réduites. Ainsi, les souches de *Prochlorococcus* semblent subir un enrichissement en bases AT, de même que les souches non réduites (bien qu'il soit plus faible), entraînant un changement de l'usage des codons et de la sélection sur celui-ci. Ainsi, le

biais de composition nucléotidique est la force principale dirigeant la composition en codon des séquences pour les souches riches en bases AT, alors que c'est la sélection traductionnelle qui domine pour les souches riches en bases GC. Dans ces souches, la sélection contre les mutations délétères vis-à-vis de l'usage des codons est plus forte que pour les souches riches en AT. Ainsi, alors que la sélection sur les protéines ne semble pas avoir changé avec l'évolution réductive, la sélection sur l'usage des codons a quant à elle perdu de l'importance au détriment des biais mutationnels. Le biais vers l'enrichissement en bases AT est relativement fort. En effet, ce biais semble d'abord toucher les préférences des codons au sein des acides aminés, principalement le long de la branche ancestrale aux souches réduites de *Prochlorococcus*, riches en AT. Quand le maximum de modifications ont été effectuées sans altérer la qualité des protéines ($^{d_N}/_{d_S}$ constant), ce sont les préférences en acides aminés qui sont modifiées, le long de la branche ancestrale aux souches de *Prochlorococcus* HL. D'où vient ce biais mutationnel fort s'il n'est pas dû à un relâchement des pressions de sélection ? Serait-il dû à une pression de l'environnement pour une réduction de la dépense énergétique des cellules, qui serait aussi à l'origine de la réduction de la taille des génomes ? Dans ce cas, pourquoi touche-t-il toutes les souches de *Prochlorococcus* et dans une moindre mesure les souches de *Prochlorococcus* LLIV ?

Pour aller plus loin dans l'analyse des pressions de sélection, en utilisant le modèle de Pouyet *et al.* (2013), les valeurs de $^{d_N}/_{d_S}$ gène à gène devraient être estimées et comparées entre les différentes souches, étudier l'homogénéité de la conservation de $^{d_N}/_{d_S}$ mais aussi analyser les sites sous pression de sélection positive afin de déterminer la part d'adaptation dans les changements observés et les gènes touchés.

Chapitre XII

Synthèse : L'évolution réductive chez *Prochlorococcus*

Dans les chapitres précédents, nous avons analysé un grand nombre de caractéristiques génomiques dans un contexte phylogénétique afin de déterminer où ont principalement eu lieu les changements au sein du genre *Prochlorococcus* mais aussi pour donner des clés de compréhension de l'évolution réductive chez *Prochlorococcus*.

D'après la reconstruction de l'évolution des contenus en gènes (Chapitre VIII, Figure XII.1), l'évolution réductive semble s'être initiée au moment de la divergence de *Prochlorococcus* et *Synechococcus*, comme suggéré par Sun et Blanchard (2014), et non après la divergence entre les souches réduites et non réduites. Elle serait ainsi induite par la spéciation de *Prochlorococcus*. Cependant, contrairement à ce qui a été observé par Sun et Blanchard (2014), l'évolution réductive toucherait toutes les branches conduisant aux souches réduites et ne se serait pas arrêtée avec la divergence des différentes souches réduites (Figure XII.1), comme observé par Luo *et al.* (2011).

Seule une branche interne conduisant à des souches réduites montre plus de gains de gènes que de pertes : la branche ancestrale aux souches HL (Figure XII.1). Les souches ayant évolué le long de cette branche ont changé d'environnement, remontant la colonne d'eau, et les gains de gènes pourraient ainsi avoir favorisé l'adaptation au nouvel environnement.

Des gènes liés à la réparation, la réplication et la recombinaison ont été perdus au cours de l'évolution réductive (Dufresne *et al.*, 2005, 2003; Partensky et Garczarek, 2010; Kettler *et al.*, 2007), principalement le long de 4 branches (Figure XII.1) : la branche ancestrale à l'ensemble des souches de *Prochlorococcus*, la branche ancestrale aux souches réduites, la branche ancestrale aux souches LLI et HL et la branche ancestrale aux souches HL. Ainsi, les taux de mutation dans l'ensemble des *Prochlorococcus* seraient élevés et la perte de gènes de réparation des mutations GC→AT dans les branches ancestrales aux souches réduites et aux souches HL pourrait expliquer l'enrichissement en bases AT observé pour les souches réduites. L'augmentation des taux de mutation pourrait être transitoire, au

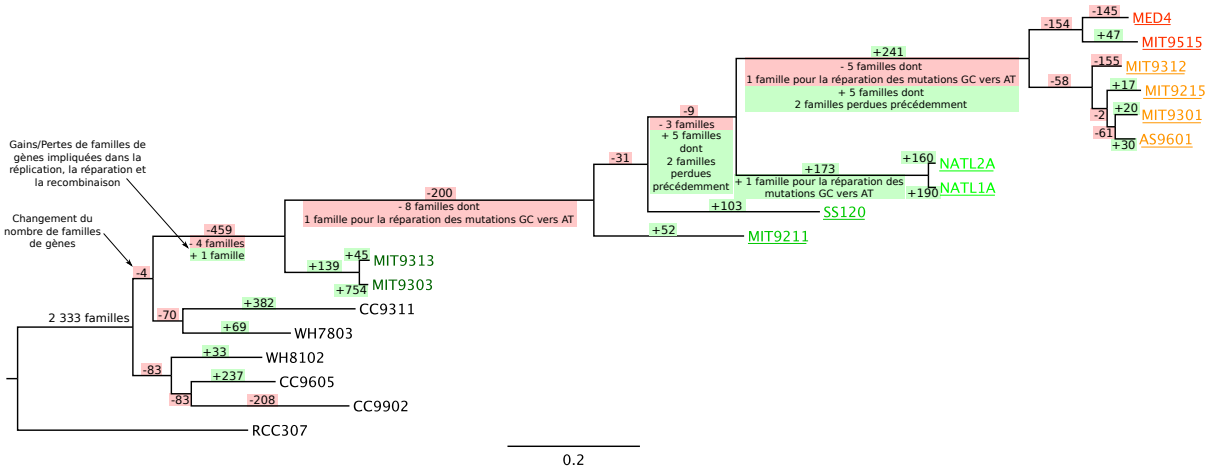


Figure XII.1 – Gains et pertes de gènes le long de la phylogénie de *Prochlorococcus* et de *Synechococcus*
Les valeurs au-dessus de chaque branche correspondent à la différence entre le nombre de familles de gènes gagnées et le nombre de familles perdues le long de la branche en question (Figure VIII.3). Les rectangles rouges symbolisent les cas où le nombre de familles perdues excède le nombre de familles gagnées et les rectangles verts l'inverse.
En dessous des branches, sont inscrits les gains et pertes de familles de gènes potentiellement impliquées dans la réplication, la réparation et la recombinaison (Figure VIII.6), avec en rouge les pertes et en vert les gains. Plus de détail est fourni sur les familles impliquées dans la réparation des mutations GC vers AT, ou des familles perdues ou gagnées précédemment dans la phylogénie.
Les couleurs de noms des souches symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit. L'arbre phylogénétique est construit comme décrit dans la section C.4 (Annexe).

moins pour certaines souches. En effet, les gènes de réparation perdus dans la branche ancestrale à *Prochlorococcus* sont regagnés dans les branches ancestrales aux souches de *Prochlorococcus* LLI et HL, en particulier les gènes impliqués dans la réparation des dommages causés par les UV, utiles en haut de la colonne d'eau où les UV sont plus forts. Pour vérifier cela, il faudrait avoir des estimations fiables des taux spontanés de mutation dans les différents écotypes.

L'évolution réductive par la perte de gènes démarre ainsi avec la divergence entre *Prochlorococcus* et *Synechococcus*. Cependant, l'évolution réductive ne touche pas seulement la quantité de gènes mais aussi la longueur des gènes, même si la phase de réduction de la longueur des gènes est plus tardive que celle des pertes de gènes (Figure XII.2) : elle démarre le long de la branche ancestrale aux souches réduites et s'intensifie au fur et à mesure des divergences jusqu'aux souches HL, principalement par des pertes de bases à l'intérieur des gènes et non aux extrémités. *In fine*, les gènes au sein des souches HL sont au minimum 15 bases plus petits que ceux à la racine de l'arbre (Figure XII.2).

Avec la perte d'un gène de réparation des mutations GC→AT le long de la branche ances-

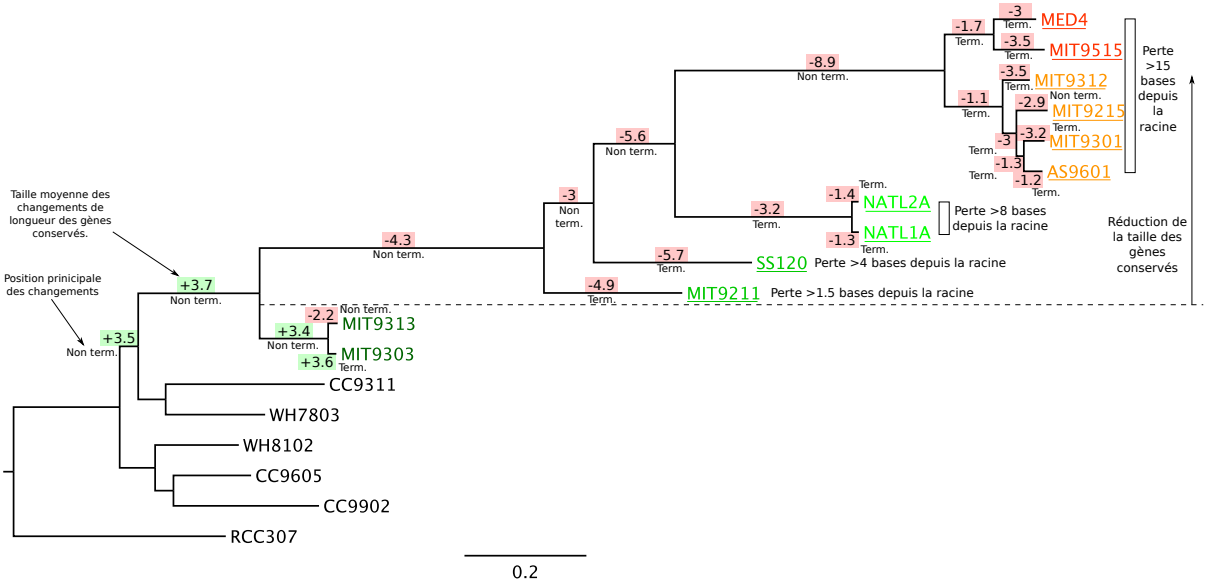


Figure XII.2 – Changement de la longueur des gènes au sein de 693 familles de gènes orthologues le long de la phylogénie de *Prochlorococcus* et de *Synechococcus*
Les valeurs au-dessus de chaque branche correspondent à la différence entre le nombre moyen de bases gagnées par gène et le nombre moyen de bases perdues par gène via des insertions et des délétions (Figure IX.10). Les rectangles rouges symbolisent les cas où la taille moyenne des gènes diminue et les rectangles verts l'augmentation de la taille moyenne des gènes.
En-dessous des branches, est indiquée l'origine principale des gains et pertes de bases, c'est-à-dire des bases principalement aux extrémités (Term.) ou à l'intérieur (Non term.) des gènes.
Les couleurs de noms des souches symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit. L'arbre phylogénétique est construit comme décrit dans la section C.4 (Annexe).

trale aux souches réduites, les séquences s'enrichissent en bases AT (Figure XII.3) induisant un changement des fréquences des codons synonymes mais aussi des changements des fréquences des acides aminés, principalement le long de la branche où le gène a été perdu. Un autre gène impliqué dans la réparation des mutations GC→AT est perdu dans la branche ancestrale aux souches HL. Cette perte semble avoir principalement entraîné un changement des fréquences des acides aminés (Figure XII.3), les codons synonymes étant déjà, à ce stade, assez enrichis en bases AT. Ainsi, les taux de substitution non synonyme et synonyme sont relativement élevés le long des branches où les changements de codons et d'acides aminés au sein des séquences ont lieu, tout comme les taux de transition et transversion (Figure XII.3) : les taux de mutation sont donc élevés.

Avec l'enrichissement en bases AT et les changements d'usage des codons et des acides aminés, les causes de l'usage des codons ont été modifiées. Ainsi, la sélection traductionnelle ne semble pas être le principal acteur du biais d'usage des codons chez *Prochlorococcus*, contrairement à *Synechococcus* (Figure X.17). Le biais d'usage des codons est alors

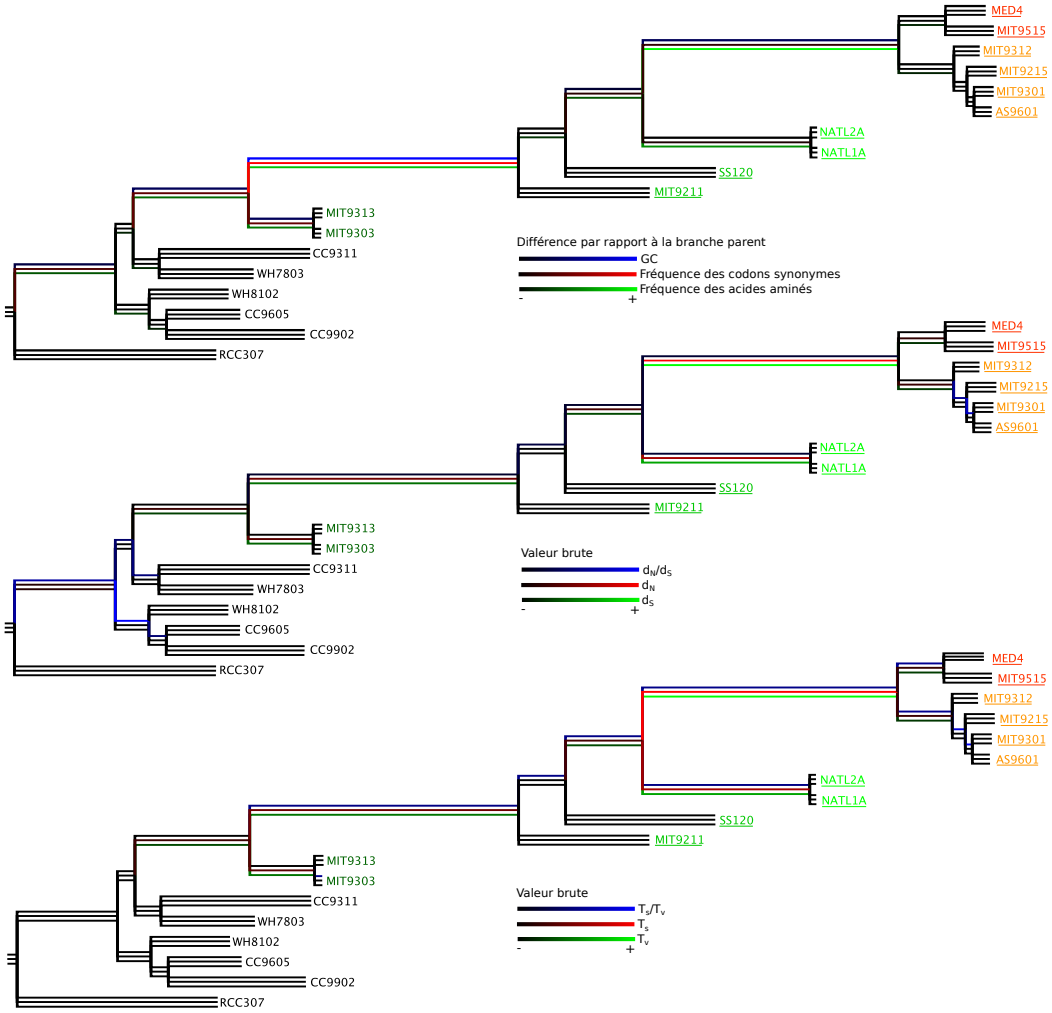


Figure XII.3 – Évolution des séquences au sein de 693 familles de gènes orthologues le long de la phylogénie de *Prochlorococcus* et de *Synechococcus*
Pour l'arbre du haut, chaque branche est colorée en fonction de la différence de taux GC (Figure XI.5), de la somme des valeurs absolues des différences de fréquences des codons synonymes (Figure XI.6) et de la somme des valeurs absolues des différences de fréquences des acides aminés (Figure XI.7) par rapport à la branche ancestrale directe de la branche étudiée.
Pour l'arbre au centre, chaque branche est colorée selon les valeurs de $^{d_N}/_{d_S}$, d_N et d_S estimées par comptage avec la méthode de Pouyet *et al.* (2013) (Figure XI.8).
Pour l'arbre du bas, chaque branche est colorée selon les valeurs de $^{T_s}/_{T_v}$, T_s et T_v estimées par comptage avec la méthode de Pouyet *et al.* (2013) (Figure XI.4).
Les couleurs de noms des souches symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit. Les arbres phylogénétiques sont construits comme décrit dans la section C.4 (Annexe).

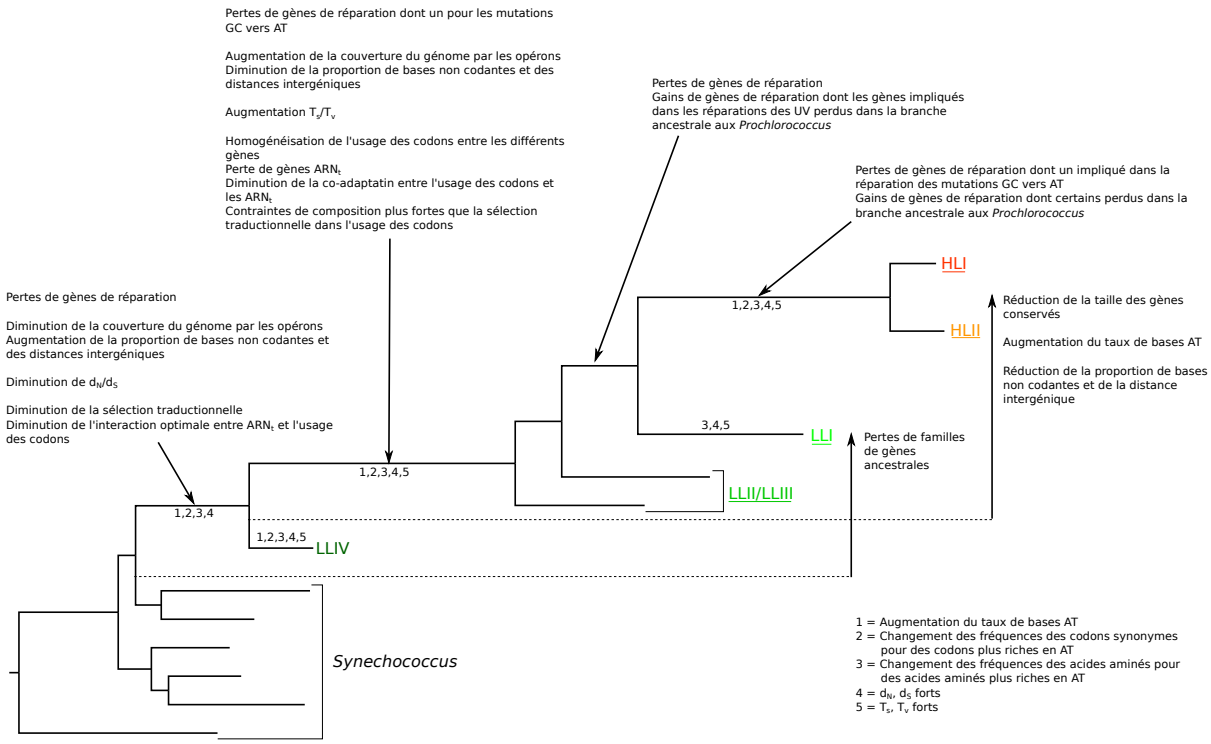


Figure XII.4 – Changements le long de la phylogénie de *Prochlorococcus* et de *Synechococcus* Seuls les différents écotypes sont représentés. Les écotypes dont le nom est souligné contiennent seulement des souches pour lesquelles le génome est réduit. L'arbre phylogénétique est construit comme décrit dans la section C.4 (Annexe).

principalement dirigé par les biais mutationnels, avec une homogénéisation des différences d'usage des codons entre les gènes, la perte de la co-adaptation entre le répertoire d'ARN$_t$ et l'usage des codons, etc.

Enfin, les pressions de sélection sur les protéines, observées par le $^{d_N}/_{d_S}$, sont conservées au sein de *Prochlorococcus*, mais le $^{d_N}/_{d_S}$ semble avoir légèrement diminué au moment de la divergence entre *Prochlorococcus* et *Synechococcus* (Figure XII.3), principalement par l'augmentation de d_S et les changements d'usage des codons liés à l'enrichissement en bases AT.

Les différentes observations effectuées dans les chapitres précédents peuvent être résumées par la figure XII.4. Il en ressort que les évènements liés aux changements génomiques sont concentrés au sein de plusieurs branches, et pas une seule, contrairement à la vision générale de l'évolution réductive chez *Prochlorococcus*. Ainsi, la comparaison entre les souches réduites et non réduites seulement,comme faite dans le tableau I.1, semble insuffisante. Les changements chez *Prochlorococcus* et l'évolution réductive pourraient être résumés à trois phases : la différenciation entre *Prochlorococcus* et *Synechococcus*, la première phase de l'évolution réductive correspondant à la branche ancestrale aux souches réduites et la seconde phase de l'évolution réductive correspondant à la différenciation entre les souches LL et les souches HL (Tableau XII.1).

Caractéristiques	*Prochlorococcus* vs *Synechococcus*					*Prochlorococcus* réduite vs *Prochlorococcus* non réduite					*Prochlorococcus* HL vs *Prochlorococcus* LL réduite				
		CM	AE	FTM	HRN		CM	AE	FTM	HRN		CM	AE	FTM	HRN
Taille du génome	Stable	-	-	+	-	Réduction	+	?	+	+	Réduction	+	+	+	+
Proportion d'ADN codant	Réduction	+	-	+	=	Augmentation	-	?	+	=	Augmentation	-	+	+	=
Distances intergéniques	Augmentation	+	-	+	=	Réduction	-	?	+	=	Réduction	-	+	+	=
%GC	Réduction	+	-	+	=	Réduction	+	?	+	=	Réduction	+	+	+	=
Couverture par les opérons	Réduction	+	-	+	=	Augmentation	-	?	+	=	Stable	-	-	+	=
Longueur des gènes	Stable	-	+	+	+	Réduction	+	?	+	-	Réduction	+	+	+	-
Familles de gènes	Réduction	+	+	+	+	Réduction	+	?	+	+	Augmentation puis réduction	-	+	+	-
Recombinaison intragénique	Stable	-	+	+	+	Stable	-	?	+	+	Stable	-	+	+	+
Gènes de réparation	Pertes	+	-	+	-	Pertes	+		+	-	Pertes	+	-	+	-
Vitesse d'évolution	Stable	-	+	-	+	Augmentation	+	?	+	-	Augmentation	+	-	+	-
$^{d_N}/_{d_S}$	Stable	-	?	+	+	Stable	-	?	+	+	Stable	-	+	+	+
d_N	Augmentation	+	-	+	-	Augmentation	+	?	+	-	Augmentation	+	-	+	-
d_S	Augmentation	+	-	+	-	Augmentation	+	?	+	-	Augmentation	+	-	+	-
$^{T_s}/_{T_v}$	Stable	?	+	?	+	Augmentation	?	?	+	-	Augmentation	?	-	+	-
T_s	Augmentation	+	-	+	-	Augmentation	+	?	+	-	Augmentation	+	-	+	-
T_v	Augmentation	+	-	+	-	Augmentation	+	?	+	-	Augmentation	+	-	+	-
Gènes ARN$_t$	Stable	-	+	+	-	Réduction	+	?	+	-	Stable	-	-	+	+
Interaction optimale entre gènes ARN$_t$ et usage des codons	Réduction	+	-	?	=	Stable mais faible	-	?	?	=	Stable mais faible	-	=	+	=
Co-adaptation entre gènes ARN$_t$ et usage des codons	Stable	-	+	?	=	Réduction	+	?	+	=	Stable mais faible	-	=	+	=
Différence d'usage des codons entre les gènes ribosomaux et non ribosomaux	Stable	-	+	?	=	Homogénéisation	+	?	+	=	Stable mais faible	-	=	+	=
Force de la sélection traductionnelle	Réduction	+	-	+	=	Réduction	+	?	+	=	Stable mais faible	-	=	+	=
Codons optimaux	Stable	-	+	?	=	Réduction	+	?	+	=	Stable	-	=	+	=
Changement dans la constitution en acides aminés	Riche en AT	+	-	+	-	Riche en AT	+	?	+	-	Riche en AT		-	+	-

Table XII.1 – Motifs et hypothèses pour les changements des caractéristiques de *Prochlorococcus*
CM : Cliquet de Muller, AE : adaptation à l'environnement pauvre en nutriments, FTM : fort taux de mutation, HRN : hypothèse de la reine noire. "+" et "-" indiquent les observations qui confirment et contredisent une hypothèse donnée pour l'évolution réductive chez *Prochlorococcus*, respectivement. "=" symbolise une hypothèse ne faisant aucune prédiction pour un motif donné. "?" indique que des travaux théoriques ou des données écologiques supplémentaires sont nécessaires pour étudier la prédiction d'une hypothèse donnée.

Comment expliquer ces différentes phases ? Quelles en sont les causes ? Pour tenter d'élucider cette question, nous comparons, dans le tableau XII.1 comme pour le tableau I.1, les différents motifs observés et les différentes hypothèses proposées (cliquet de Muller, adaptation à l'environnement, fort taux de mutation et hypothèse de la Reine Noire). Nous discutons ces hypothèses en relation avec les résultats des scénarios présentés dans la première partie du manuscrit.

Avec le cliquet de Muller, touchant principalement les petites populations non recombinantes, la sélection naturelle est dépassée par la dérive. Ainsi, les mutations délétères s'accumulent entraînant une dégénérescence, une évolution accélérée des séquences et la perte de gènes non essentiels. Dans le cas de *Prochlorococcus*, le cliquet de Muller peut expliquer les taux d'évolution élevés, l'augmentation des taux de substitutions synonymes, non synonymes et des taux de transition et de transversion, le raccourcissement des gènes, les changements non adaptatifs d'acides aminés, l'enrichissement en bases AT et les pertes de gènes (Tableau XII.1). Dans une évolution réductive typique du cliquet de Muller, les pertes de gènes auraient principalement lieu par pseudogénéisation, augmentant ainsi la proportion de bases non codantes et diminuant la couverture par les opérons. Ces bases seraient ensuite progressivement éliminées. Cependant, lorsque nous simulons les scénarios de diminution de la taille de population et d'arrêt de la recombinaison, ces changements ne sont pas observés. De plus, certaines caractéristiques observées chez *Prochlorococcus* sont difficilement explicables par le cliquet de Muller comme la présence de recombinaison intragénomique (Tableau XII.1). Or, dans le seul scénario lié à l'hypothèse du cliquet de Muller où les changements sont compatibles avec une évolution réductive (diminution de la pression de sélection), la taille efficace de population semble augmenter. Ainsi, des changements de structure de population entraînant une augmentation de la taille efficace de population, comme l'agrandissement de la niche écologique, pourrait être à l'origine de l'évolution réductive chez *Prochlorococcus*. Pour confirmer ces observations, il nous faudrait des estimations plus fiables de la taille efficace de population le long de la phylogénie. L'utilisation du ${}^{d_N}/_{d_S}$ présente quelques lacunes car ce ratio reste un estimateur composite ($N_e s$, avec s le coefficient de sélection). Il en est de même pour le niveau de polymorphisme neutre ($N_e u$, avec u le taux de mutation). Cependant, ce dernier peut permettre d'estimer la taille efficace de population actuelle des différents écotypes, comme pour MIT9312 (Kashtan *et al.*, 2014). En combinant ensuite les deux approches, l'évolution de N_e le long de la phylogénie de *Prochlorococcus* pourrait être reconstruite, permettant ainsi de déterminer si les structures de populations de *Prochlorococcus* ont changé au cours de l'évolution réductive.

L'hypothèse de la Reine Noire repose sur la mutualisation de tâches au sein de communautés bactériennes (Morris *et al.*, 2012). Elle peut expliquer la perte de certains gènes mais pas la perte de gènes de réparation ou de gènes d'ARN$_t$. De plus, cette hypothèse ne fait pas de prédiction pour la plupart des autres caractéristiques génomiques ayant évolué chez *Prochlorococcus* (Tableau XII.1). Les scénarios de simplification de l'environnement (suppression/neutralisation d'un lobe de l'environnement) pourraient simuler cette hypothèse : les gènes devenus inutiles sont perdus sans que les autres caractéristiques génomiques changent, comme prédit. Ainsi, cette hypothèse seule ne suffit pas à expliquer les caractéristiques de l'évolution réductive chez *Prochlorococcus*.

Les souches de *Prochlorococcus* se trouvent dans les eaux tropicales et subtropicales où les changements tout au long de l'année sont peu fréquents, contrairement aux eaux tempérées où vivent les *Synechococococcus*. Dans un environnement stable, une machinerie de régulation sophistiquée et coûteuse utile pour répondre aux variations de concentration des nutriments peut être perdue à faible coût tout comme les gènes impliqués dans la régulation ou utiles pour la vie en zone tempérée tant que le bénéfice de la perte est supérieur à son coût. L'adaptation au nouvel environnement de *Prochlorococcus* peut ainsi expliquer certaines pertes de gènes mais pas celles des gènes de réparation ou des gènes d'ARN$_t$ (Tableau XII.1). Les observations faites dans les trois scénarios liés à la simplification et à l'arrêt de la variation de l'environnement sont cependant en opposition avec ce qui est attendu : pas de pertes de gènes ni de simplification des réseaux de régulation. Cette non-simplification des réseaux de régulation simulés peut être due aux limites du modèle de régulation utilisé (Chapitre V). Ainsi, nous ne pouvons pas nous fonder sur ce résultat de simulation seul pour exclure définitivement le scénario d'adaptation à un environnement plus stable. Une analyse plus détaillée des gains et des pertes des facteurs de transcription le long de la phylogénie de *Prochlorococcus* et de *Synechococcus* serait nécessaire pour conclure.

L'adaptation à une nouvelle hauteur d'eau est la principale hypothèse proposée pour expliquer l'évolution réductive chez *Prochlorococcus* (Rocap *et al.*, 2003; Dufresne *et al.*, 2005; Giovannoni *et al.*, 2005; Partensky et Garczarek, 2010). Cette hypothèse prédit une corrélation entre la réduction des génomes et la hauteur de la colonne d'eau. Cependant, les souches LL réduites se trouvent à des profondeurs équivalentes à celles des souches LL non réduites. Un changement d'environnement pour expliquer les différences entre les souches non réduites et réduites dans leur ensemble est donc peu plausible (Tableau XII.1) et des données écologiques supplémentaires seraient ainsi nécessaires pour déterminer ce qui distingue les souches LL réduites et les souches LL non réduites et donner des pistes sur les causes de l'évolution réductive au sein des souches LL.

Les souches LL et HL se distinguent, quant à elles, par un environnement différent. Les souches LL sont situées en bas de la colonne d'eau et les souches HL en haut de la colonne d'eau où les nutriments sont rares. Dans cet environnement, un petit génome est un avantage en diminuant les besoins en azote et en phosphore, deux éléments très rares dans les eaux de surface, mais aussi en augmentant le ratio cellulaire surface-volume et améliorant ainsi l'assimilation des nutriments (Dufresne *et al.*, 2005; Giovannoni *et al.*, 2005). Ainsi, cette hypothèse peut expliquer les pertes de gènes, le raccourcissement des gènes et du non codant (Tableau XII.1), mais aussi les nombreux gains de gènes le long de la branche ancestrale aux souches HL. En effet, comme observé dans le scénario de déplacement d'un lobe de l'environnement, le changement de niche entraîne de nouveaux besoins et donc la nécessité de recruter de nouveaux gènes pour combler ces besoins. Il semble nécessaire d'avoir une annotation plus précise et fiable des gènes gagnés et perdus le long de cette branche pour déterminer la cause des changements de répertoires géniques, en utilisant les données de Yooseph *et al.* (2010) par exemple. L'ATP étant le moins coûteux des nucléotides à produire (Rocha et Danchin, 2002), l'enrichissement en bases AT, par la perte d'un gène de réparation des mutations GC→AT, pourrait être favorisé par la sélection pour un génome moins coûteux dans un environnement pauvre

en nutriment (Giovannoni *et al.*, 2005, 2014). Cependant, les pertes des 4 autres gènes de réparations dans la branche ancestrale aux souches HL sont difficilement explicables dans ce scénario car le bénéfice de telles pertes n'est pas évident (Marais *et al.*, 2008).

Ces pertes des gènes de réparation peuvent augmenter les taux de mutation. Les séquences évoluent ainsi rapidement et les biais mutationnels sont exacerbés, comme le biais vers la délétion (Mira *et al.*, 2001; Kuo et Ochman, 2009) ou l'enrichissement en bases AT par la perte de gènes impliqués dans la réparation des mutations GC→AT. La sélection traductionnelle est alors dépassée par les biais de composition nucléotidique dans l'usage des codons entrainant une homogénéisation de l'usage des codons entre les différents gènes, une perte de co-adaptation entre les ARN_t et l'usage des codons (Tableau XII.1). Dans le cas d'un changement des taux de mutation (comme cela pourrait être le cas pour la branche ancestrale à *Prochlorococcus*), conformément à ce qui est observé dans le scénario d'augmentation des taux de mutation, les gènes non essentiels sont perdus par l'accumulation de mutations et la pseudogénéisation. La quantité de bases non codantes augmente jusqu'à ce qu'elles soient éliminées par délétion. Dans une seconde phase, la taille du génome se réduit par de nouvelles pertes de gènes, l'élimination des bases non codantes et potentiellement de bases au sein des gènes. À un certain point, tous les gènes non essentiels ont été éliminés et toute perte de gènes est plus désavantageuse que bénéfique. Si les taux de mutation restent élevés, les changements et les réductions toucheront principalement le non codant et les séquences des gènes. Ces changements pourraient correspondre à ceux observés pour *Prochlorococcus* : augmentation de la proportion de non codant, accompagnée de la perte de gènes, peu après la divergence avec *Synechococcus* ; continuation des pertes de gènes dans les ancêtres des souches réduites et réduction de la proportion de non codant mais aussi des séquences géniques jusqu'à atteindre les branches des souches HL. L'augmentation des taux de mutation se reflète aussi dans les taux d'évolution (d_N, d_S, T_s,...) élevés et les nombreuses insertions et délétions au sein des gènes le long de certaines branches. Ces estimations indirectes des pressions mutationnelles se concentrent sur les séquences codantes. Il pourrait être intéressant de faire les analyses d'évolution des séquences sur des portions non codantes, où les pressions de sélection sont moins fortes. Cependant, les séquences non codantes sont difficiles à comparer pour des souches ayant divergé depuis si longtemps. Pour contrer cela, nous pourrions utiliser la méthodologie développée pour l'évolution de la longueur des gènes en élargissant les alignements de part et d'autre des gènes pour étudier les changements en amont et en aval des gènes. Dans le scénario d'augmentation des taux de mutation, le nombre de réarrangements fixés diminue. Cela est compatible avec le fait que, pour les souches réduites de *Prochlorococcus*, l'architecture génomique est plus stable (Dufresne *et al.*, 2005). Il pourrait cependant être intéressant de quantifier les taux de réarrangement le long des branches de la phylogénie. L'augmentation des taux de mutation est néanmoins l'hypothèse expliquant le plus grand nombre de caractéristiques génomiques (Tableau XII.1).

Une augmentation transitoire du taux de mutation peut améliorer l'adaptation des bactéries aux changements environnementaux (Taddei *et al.*, 1997; Tenaillon *et al.*, 1999). Comme *Prochlorococcus*, les organismes mutateurs manquent souvent de gènes de réparation de l'ADN et évoluent rapidement. Marais *et al.* (2008) ont proposé une histoire évolutive où la perte des gènes de réparation dans certains écotypes pourrait pousser ces

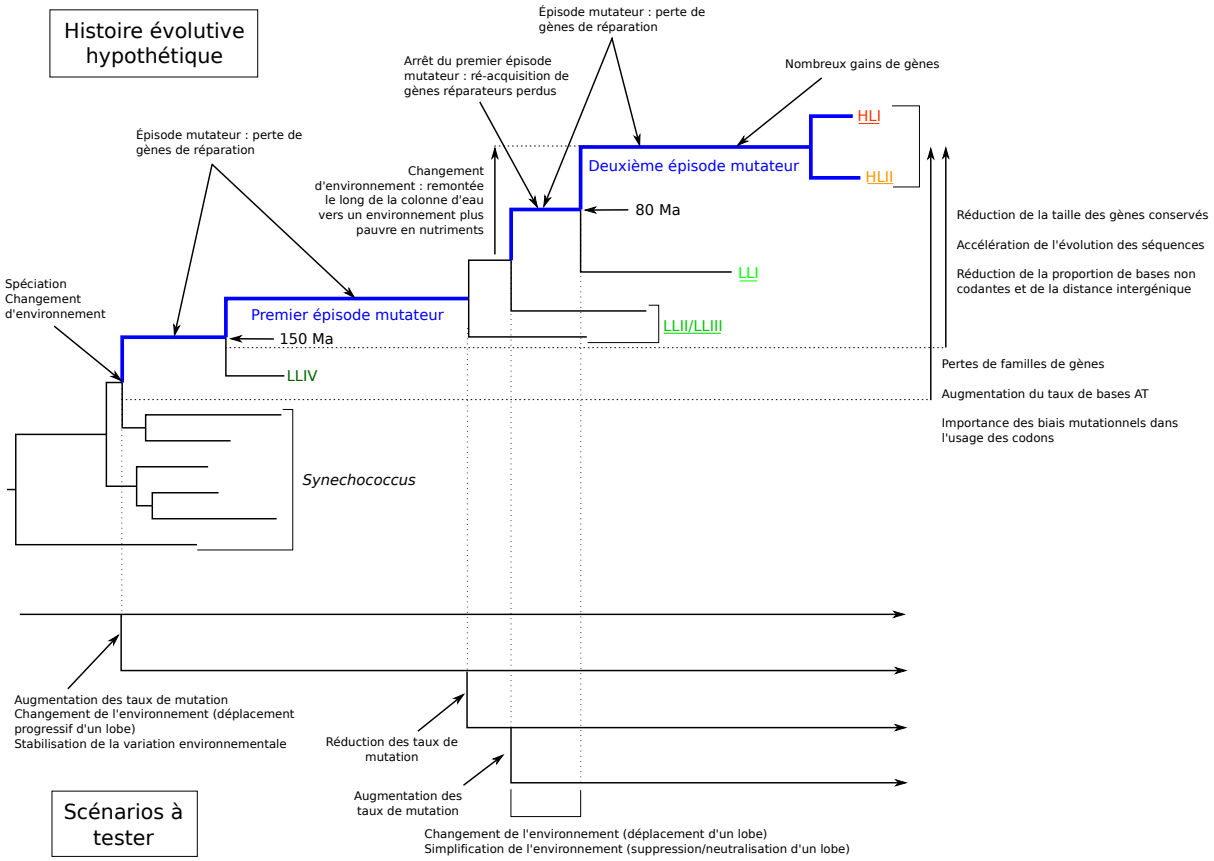


Figure XII.5 – Histoire évolutive hypothétique et scénario à tester pour expliquer les changements génomiques le long de la phylogénie de *Prochlorococcus* et de *Synechococcus*
Seuls les différents écotypes sont représentés. Les écotypes dont le nom est souligné contiennent seulement des souches pour lesquelles le génome est réduit. L'arbre phylogénétique est construit comme décrit dans la section C.4 (Annexe).

derniers à devenir mutateurs, expliquant les pertes de gènes, l'enrichissement en bases AT et l'évolution rapide des séquences (Tableau XII.1), mais n'expliquant pas la perte définitive des gènes de réparation[1]. Partensky et Garczarek (2010) ont ainsi proposé une histoire évolutive combinant la simplification adaptative et l'hypothèse mutatrice pour expliquer la réduction des génomes, avec plusieurs évènements indépendants de souches mutatrices suivies par la restauration d'un taux de mutation plus faible quand la population est adaptée à son nouvel environnement. Ils supposent la présence de trois épisodes mutateurs : la branche ancestrale aux souches réduites, la branche ancestrale aux souches LLI et HL et la branche ancestrale aux souches HL. Quel évènement déclencheur pourrait être à l'origine de l'augmentation des taux de mutation le long de la branche ancestrale aux souches réduites alors que les souches LL réduites et non réduites semblent avoir des

[1]Conserver un taux de mutation élevé est délétère à long terme pour les organismes, comme nous l'avons observé dans la première partie du manuscrit. Après un épisode mutateur, les gènes de réparation sont ainsi réacquis, principalement par transfert, afin de diminuer les taux de mutation. *Prochlorococcus* est capable d'acquérir des gènes par transfert (Dufresne *et al.*, 2008; Luo *et al.*, 2011; Coleman *et al.*, 2006) et devrait donc réacquérir des gènes de réparation quand l'adaptation rapide à l'environnement n'est plus nécessaire, comme c'est le cas pour certains gènes de réparation (Figure VIII.6)

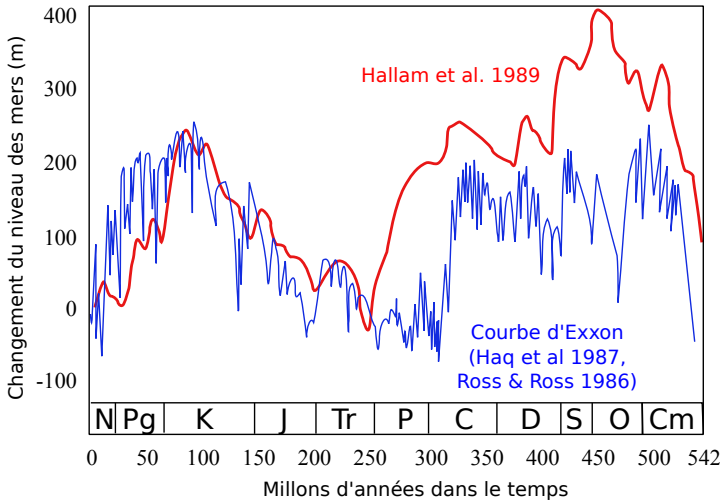


Figure XII.6 – Estimations des fluctuations des niveaux globaux des mers sur les 500 derniers millions d'années
La figure compare les reconstruction des niveaux des mers de Hallam et Cohen (1989) (en rouge) et d'Exxon (Haq *et al.*, 1987; Ross et Ross, 1986) (en bleu).
La figure est issue d'une image de Wikipédia dont les droits d'utilisation et de modification sont libres.

préférences écologiques similaires ?

Nous proposons une histoire évolutive un peu différente de celle de Partensky et Garczarek (2010) dans le nombre d'épisodes mutateurs et la position de ceux-ci le long de la phylogénie (Figure XII.5). Une souche mutatrice de *Synechococcus* aurait acquis un avantage sélectif lui permettant de coloniser une nouvelle niche écologique, donnant ainsi naissance à *Prochlorococcus*. Cette adaptation aurait été facilitée par une augmentation des taux de mutation. Ce premier épisode mutateur se serait arrêté dans les souches ancestrales aux souches non réduites mais aurait continué dans les souches ancestrales aux souches réduites, potentiellement en lien avec la colonisation de nouvelles niches (eaux tropicales et sub-tropicales, ...), avec de nouvelles pertes de gènes de réparation et un impact plus important des mutations sur les séquences (insertions, délétions, substitutions, ...) et la réduction des génomes. Les gènes de réparation perdus initialement sont regagnés pour les souches ancestrales aux souches HL et LL et le premier épisode mutateur s'arrête. Un nouvel épisode s'initie avec de nouvelles pertes de gènes de réparation et la colonisation de nouvelles niches en remontant la colonne d'eau. Le nouvel environnement étant pauvre en nutriments, des changements adaptatifs, principalement par la simplification, sont nécessaires pour limiter les dépenses énergétiques. Ces changements sont favorisés par les taux élevés de mutation.

Pourquoi ces nouvelles niches écologiques n'auraient-elles pas été conquises plus tôt ? Les évènements importants dans la phylogénie de *Prochlorococcus* et en particulier les deux épisodes mutateurs pourraient être liés à des évènements géologiques globaux, comme des crises d'extinctions massives d'espèces laissant des niches écologiques vides dans lesquelles *Prochlorococcus* a pu se développer. Par exemple, un des épisodes de changement d'environnement et d'augmentation des taux de mutation pourrait correspondre à la crise

biologique du Crétacé-Tertiaire (Cr/T, 65 millions d'années) qui a vu l'extinction de plus de 70% des espèces marines et de nombreuses espèces terrestres dont les dinosaures. À ce moment là, de nombreux changements climatiques ont eu lieu avec, par exemple, une réduction du niveau de la mer (Figure XII.6). L'évolution réductive chez *Prochlorococcus* pourrait ainsi être liée à des évènements géologiques majeurs.

Cette histoire évolutive scénario pourra difficilement être prouvée en l'absence des génomes ancestraux. Cependant, elle peut être contredite ou affinée avec les différentes analyses complémentaires proposées précédemment. De plus, les simulations de scénarios combinés, en particulier celui proposé dans la figure XII.5, pourraient aider dans cette démarche en ouvrant de nouvelles pistes d'exploration des génomes.

Conclusions et perspectives

L'évolution réductive chez *Prochlorococcus* est un phénomène complexe dont il est difficile de comprendre les causes étant donné le peu de données disponibles et les nombreux mécanismes qui semblent être impliqués.

Dans la première partie du manuscrit, nous avons testé les hypothèses proposées dans la littérature pour expliquer l'évolution réductive grâce à des expériences d'évolution *in silico*. Celles-ci ont permis d'exclure certaines hypothèses, de poser des questions sur les causes des changements observés et de mettre en avant la présence d'une pression indirecte dans l'évolution des organismes simulés. De nouvelles expériences *in silico*, plus ou moins complexes, sont nécessaires pour confirmer les observations. Les simulations ouvrent aussi des questions sur les caractéristiques génomiques de l'évolution réductive chez *Prochlorococcus* qui doivent être explorées.

Plusieurs d'entre elles ont été effectuées. Ainsi, dans la seconde partie du manuscrit, nous avons présenté une série d'analyses des génomes disponibles des 12 souches de *Prochlorococcus* afin de répondre à certaines questions émises dans la première partie et surtout d'en déduire le déroulement hypothétique des évènements ayant conduit aux réductions observées dans les génomes. Nous avons ainsi proposé une histoire évolutive hypothétique basée sur deux épisodes mutateurs liés à des modifications d'environnement : un premier épisode initié à la divergence entre *Prochlorococcus* et *Synechococcus* et un second lié à la colonisation des écotypes plus haut dans la colonne d'eau. Les temps de divergence entre les différentes souches de *Prochlorococcus* étant proches des temps géologiques, les changements observés le long de la phylogénie devraient être remis dans un cadre géologique avec la reconstruction de l'évolution de l'environnement dans lequel vivent les différents souches de *Prochlorococcus* (variations climatiques, tectonique des plaques, ...). Nous avons ainsi initié une collaboration avec un paléontologue, spécialiste des paléoenvironnements.

Chacune des deux méthodes utilisées dans cette thèse s'est nourri ainsi des résultats obtenus avec l'autre méthode, permettant d'initier des allers-retours entre ces deux méthodes. D'un côté, les expériences d'évolution *in silico* permettent d'aller au-delà du raisonnement verbal pour comprendre l'effet isolé de chaque facteur sur chaque caractéristique génomique, et de révéler des effets indirects, difficiles à prévoir par une simple expériences de pensée. Elles apportent le support pour éliminer certaines hypothèses et en conserver d'autres, au moins temporairement, mais aussi émettre de nouvelles questions (comme celle d'une sélection indirecte d'un compromis entre la robustesse et l'évolvabilité, par

exemple). De l'autre côté, les analyses par génomique comparative permettent d'étudier l'objet biologique dans toute sa complexité, c'est-à-dire comme le résultat de l'interaction de facteurs multiples, d'affiner les hypothèses à tester en simulation et de proposer des histoires évolutives hypothétiques.

Les deux méthodes utilisées dans ce travail de thèse ont ainsi une forte complémentarité dans l'analyse d'une question biologique complexe touchant des temps évolutifs longs. Il faudrait favoriser les échanges entre les deux communautés qui utilisent chacune des deux méthodes. Cette thèse est à notre connaissance une des premières combinant ces deux approches pour répondre à une question évolutive précise. Les principales difficultés de ces échanges résident dans les langages, qui bien que proches peuvent souvent différer, et dans les mesures accessibles. Ainsi, à cause du peu de données disponibles, les caractéristiques mesurées dans les génomes réels sont souvent la conséquence indirecte de certains mécanismes identifiables plus directement dans les expériences d'évolution *in silico*. Cependant, ces mécanismes ont souvent besoin d'être approchés par des indicateurs similaires à ceux utilisés en génomique comparative, afin de permettre le dialogue et la comparaison des résultats entre les deux approches. Par exemple, la notion de coefficient de sélection est très utilisée en génétique des populations et en génomique, alors que la communauté de l'évolution artificielle ne l'utilise quasiment pas. En effet, même si la définition du coefficient de sélection est générale, cette notion est en pratique très souvent pensée dans le contexte implicite d'une population où tous les individus possèdent un même génotype, sauf un mutant dont on cherche à caractériser la valeur sélective. Un tel contexte n'a pas de sens dans les simulations menées en évolution artificielle, où les taux de mutations et les tailles de populations utilisés sont tels qu'il y a une diversité génétique beaucoup plus grande et que la notion de génotype de référence n'a que peu de sens. Pourtant, nous pensons que la communauté de l'évolution artificielle doit faire l'effort de trouver des ponts entre ses mesures habituelles et celles de la génétique des populations, car elle risque sinon de demeurer une approche marginale, sans écho auprès des biologistes. Cette thèse est un premier pas dans cette direction, mais il reste du travail. Par exemple, nous travaillons à l'heure actuelle sur la mesure de la taille efficace des populations simulées, et sur la possibilité d'avoir une mesure de $^{d_N}/_{d_S}$, en utilisant un code génétique dégénéré qui permette des mutations synonymes.

Malgré ces difficultés, nous pensons que la combinaison d'une approche d'évolution *in silico* et d'une approche de génomique comparative permet d'apporter des regards différents pour l'analyse d'une question biologique donnée mais aussi pour le développement de nouvelles méthodes dans chacune des communautés.

Bibliographie

Abbot, P. et Moran, N. A. (2002). Extremely low levels of genetic polymorphism in endosymbionts (*Buchnera*) of aphids (*Pemphigus*). *Mol Ecol*, 11(12):2649–2660.

Abby, S. S., Tannier, E., Gouy, M. et Daubin, V. (2010). Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests. *BMC Bioinformatics*, 11(1):324.

Adachi, J. et Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J Mol Evol*, 42(4):459–468.

Adami, C., Brown, T. et Kellogg, W. (1994). Evolutionary learning in the 2d artificial life system "avida". *In Artificial life IV*, volume 1194, pages 377–381. Cambridge, MA : MIT Press.

Akashi, H. (1995). Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. *Genetics*, 139(2):1067–1076.

Akashi, H. et Gojobori, T. (2002). Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc Natl Acad Sci USA*, 99(6):3695–3700.

Akashi, H. et Schaeffer, S. W. (1997). Natural selection and the frequency distributions of “silent” DNA polymorphism in *Drosophila*. *Genetics*, 146(1):295–307.

Andersson, G. E., Karlberg, O., Canbäck, B. et Kurland, C. G. (2003). On the origin of mitochondria : a genomics perspective. *Phil Trans R Soc Lond B Bio Sci*, 358(1429):165–179.

Andersson, J. O. et Andersson, S. G. (1999). Genome degradation is an ongoing process in *Rickettsia*. *Mol Biol Evol*, 16(9):1178–1191.

Andersson, S. G. et Kurland, C. G. (1990). Codon preferences in free-living microorganisms. *Microbiol Rev*, 54(2):198–210.

Andersson, S. G. et Kurland, C. G. (1998). Reductive evolution of resident genomes. *Trends Microbiol*, 6(7):263–268.

Anisimova, M. et Gascuel, O. (2006). Approximate likelihood-ratio test for branches : A fast, accurate, and powerful alternative. *Syst Biol*, 55(4):539–552.

Anisimova, M., Nielsen, R. et Yang, Z. (2003). Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics*, 164(3):1229–1236.

Arenas, M. (2012). Simulation of molecular data under diverse evolutionary scenarios. *PloS Comput Biol*, 8(5):e1002495.

Arenas, M. et Posada, D. (2007). Recodon : Coalescent simulation of coding DNA sequences with recombination, migration and demography. *BMC Bioinformatics*, 8(1): 458.

Arenas, M. et Posada, D. (2010). Coalescent simulation of intracodon recombination. *Genetics*, 184(2):429–437.

Avrani, S., Wurtzel, O., Sharon, I., Sorek, R. et Lindell, D. (2011). Genomic island variability facilitates *Prochlorococcus*-virus coexistence. *Nature*, 474(7353):604–608.

Barrick, J. E., Yu, D. S., Yoon, S. H., Jeong, H., Oh, T. K., Schneider, D., Lenski, R. E. et Kim, J. F. (2009). Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature*, 461(7268):1243–1247.

Batut, B., Parsons, D. P., Fischer, S., Beslon, G. et Knibbe, C. (2013). *In silico* experimental evolution : a tool to test evolutionary scenarios. *BMC Bioinformatics*, 14(15):S11.

Baumdicker, F., Hess, W. R. et Pfaffelhuber, P. (2012). The infinitely many genes model for the distributed genome of bacteria. *Genome Biol Evol*, 4(4):443–456.

Beiko, R. G. et Charlebois, R. L. (2007). A simulation test bed for hypotheses of genome evolution. *Bioinformatics*, 23(7):825–831.

Beslon, G., Parsons, D., Sanchez-Dehesa, Y., Peña, J.-M. et Knibbe, C. (2010). Scaling laws in bacterial genomes : A side-effect of selection of mutational robustness ? *Biosystems*, 102(1):32–40.

Boussau, B., Brown, J. M. et Fujita, M. K. (2011). Nonadaptive evolution of mitochondrial genome size. *Evolution*, 65(9):2706–2711.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F. W., Zeller, U., Khaitovich, P., Grützner, F., Bergmann, S., Nielsen, R., Pääbo, S. et Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–348.

Bruen, T. C., Philippe, H. et Bryant, D. (2006). A simple and robust statistical test for detecting the presence of recombination. *Genetics*, 172(4):2665–2681.

Çakar, Z. P., Seker, U. O., Tamerler, C., Sonderegger, M. et Sauer, U. (2005). Evolutionary engineering of multiple-stress resistant *Saccharomyces cerevisiae*. *FEMS Yeast Res*, 5(6-7):569–578.

Cartwright, R. A. (2005). DNA assembly with gaps (dawg) : simulating sequence evolution. *Bioinformatics*, 21(Suppl 3):iii31–iii38.

Carvajal-Rodríguez, A. (2008). GENOMEPOP : A program to simulate genomes in populations. *BMC Bioinformatics*, 9(1):223.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*, 17(4):540–552.

Charif, D. et Lobry, J. R. (2007). SeqinR 1.0-2 : A contributed package to the r project for statistical computing devoted to biological sequences retrieval and analysis. *In* Bastolla, D. U., Porto, P. D. M., Roman, D. H. E. et Vendruscolo, D. M., éditeurs : *Structural Approaches to Sequence Evolution*, Biological and Medical Physics, Biomedical Engineering, pages 207–232. Springer Berlin Heidelberg.

Charif, D., Thioulouse, J., Lobry, J. R. et Perrière, G. (2005). Online synonymous codon usage analyses with the ade4 and seqinR packages. *Bioinformatics*, 21(4):545–547.

Charles, H., Calevro, F., Vinuelas, J., Fayard, J.-M. et Rahbe, Y. (2006). Codon usage bias and tRNA over-expression in *Buchnera aphidicola* after aromatic amino acid nutritional stress on its host *Acyrthosiphon pisum*. *Nucleic Acids Res*, 34(16):4583–4592.

Charles, H., Mouchiroud, D., Lobry, J., Gonçalves, I. et Rahbe, Y. (1999). Gene size reduction in the bacterial aphid endosymbiont, *Buchnera*. *Mol Biol Evol*, 16(12):1820–1822.

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet*, 10(3):195–205.

Charlesworth, J. et Eyre-Walker, A. (2006). The rate of adaptive evolution in enteric bacteria. *Mol Biol Evol*, 23(7):1348–1356.

Chase, J. W. et Richardson, C. C. (1977). *Escherichia coli* mutants deficient in exonuclease VII. *J Bacteriol*, 129(2):934–947.

Chow, S. S., Wilke, C. O., Ofria, C., Lenski, R. E. et Adami, C. (2004). Adaptive radiation from resource competition in digital organisms. *Science*, 305(5680):84–86.

Clark, M. A., Moran, N. A. et Baumann, P. (1999). Sequence evolution in bacterial endosymbionts having extreme base compositions. *Mol Biol Evol*, 16(11):1586–1598.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. et Hoon, M. J. L. d. (2009). Biopython : freely available python tools for computational molecular biology and. *Bioinformatics*, 25(11):1422–1423.

Cohen, O. et Pupko, T. (2010). Inference and characterization of horizontally transferred gene families using stochastic mapping. *Mol Biol Evol*, 27(3):703–713.

Coleman, M. L. et Chisholm, S. W. (2010). Ecosystem-specific selection pressures revealed through comparative population genomics. *P Natl Acad Sci USA*, 107(43): 18634–18639.

Coleman, M. L., Sullivan, M. B., Martiny, A. C., Steglich, C., Barry, K., DeLong, E. F. et Chisholm, S. W. (2006). Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science*, 311(5768):1768–1770.

Comings, D. E. (1972). The structure and function of chromatin. *In* Harris, H. et Hirschhorn, K., éditeurs : *Advances in Human Genetics*, numéro 3 de Advances in Human Genetics, pages 237–431. Springer US.

Conrad, T. M., Lewis, N. E. et Palsson, B. Ø. (2011). Microbial laboratory evolution in the era of genome-scale science. *Mol Syst Biol*, 7(1).

Cooper, T. F., Remold, S. K., Lenski, R. E. et Schneider, D. (2008). Expression profiles reveal parallel evolution of epistatic interactions involving the CRP regulon in *Escherichia coli*. *Plos Genet*, 4(2):e35.

Crombach, A. et Hogeweg, P. (2007). Chromosome rearrangements and the evolution of genome structuring and adaptability. *Mol Biol Evol*, 24(5):1130–1139.

Crombach, A. et Hogeweg, P. (2008). Evolution of evolvability in gene regulatory networks. *Plos Comput Biol*, 4(7):e1000112.

Crombach, A. et Hogeweg, P. (2009). Evolution of resource cycling in ecosystems and individuals. *BMC Evol Biol*, 9(1):122.

Csűrös, M. (2010). Count : evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics*, 26(15):1910–1912.

Csűrös, M. et Miklós, I. (2006). A probabilistic model for gene content evolution with duplication, loss, and horizontal transfer. *In* Apostolico, A., Guerra, C., Istrail, S., Pevzner, P. A. et Waterman, M., éditeurs : *Research in Computational Molecular Biology*, numéro 3909 de Lecture Notes in Computer Science, pages 206–220. Springer Berlin Heidelberg.

Dalquen, D. A., Anisimova, M., Gonnet, G. H. et Dessimoz, C. (2012). ALF—a simulation framework for genome evolution. *Mol Biol Evol*, 29(4):1115–1123.

Das, S., Paul, S., Chatterjee, S. et Dutta, C. (2005). Codon and amino acid usage in two major human pathogens of genus *Bartonella* — optimization between replicational-transcriptional selection, translational control and cost minimization. *DNA Res*, 12(2): 91–102.

Das, S., Paul, S. et Dutta, C. (2006). Evolutionary constraints on codon and amino acid usage in two strains of human pathogenic actinobacteria *Tropheryma whipplei*. *J Mol Evol*, 62(5):645–658.

Daubin, V. et Moran, N. A. (2004). Comment on "the origins of genome complexity". *Science*, 306(5698):978–978.

Degnan, P. H., Ochman, H. et Moran, N. A. (2011). Sequence conservation and functional constraint on intergenic spacers in reduced genomes of the obligate symbiont *Buchnera*. *Plos Genet*, 7(9):e1002252.

Denamur, E., Lecointre, G., Darlu, P., Tenaillon, O., Acquaviva, C., Sayada, C., Sunjevaric, I., Rothstein, R., Elion, J., Taddei, F., Radman, M. et Matic, I. (2000). Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. *Cell*, 103(5):711–721.

Domingo-Calap, P., Cuevas, J. M. et Sanjuán, R. (2009). The fitness effects of random mutations in single-stranded DNA and RNA bacteriophages. *Plos Genet*, 5(11): e1000742.

Dong, H., Nilsson, L. et Kurland, C. G. (1996). Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *J Mol Biol*, 260(5):649–663.

Doolittle, W. F. (2013). Is junk DNA bunk ? a critique of ENCODE. *P Natl Acad Sci USA*, 110(14):5294–5300.

Draghi, J. et Wagner, G. P. (2008). Evolution of evolvability in a developmental model. *Evolution*, 62(2):301–315.

Draghi, J. et Wagner, G. P. (2009). The evolutionary dynamics of evolvability in a gene network model. *J Evol Biol*, 22(3):599–611.

Drummond, D. A. et Wilke, C. O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, 134(2):341–352.

Dufresne, A., Garczarek, L. et Partensky, F. (2005). Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biol*, 6(2):1–10.

Dufresne, A., Ostrowski, M., Scanlan, D. J., Garczarek, L., Mazard, S., Palenik, B. P., Paulsen, I. T., Marsac, T. N. d., Wincker, P., Dossat, C., Ferriera, S., Johnson, J., Post, A. F., Hess, W. R. et Partensky, F. (2008). Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome Biol*, 9(5):1–16.

Dufresne, A., Salanoubat, M., Partensky, F., Artiguenave, F., Axmann, I. M., Barbe, V., Duprat, S., Galperin, M. Y., Koonin, E. V., Gall, F. L., Makarova, K. S., Ostrowski, M., Oztas, S., Robert, C., Rogozin, I. B., Scanlan, D. J., Marsac, N. T. d., Weissenbach, J., Wincker, P., Wolf, Y. I. et Hess, W. R. (2003). Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. *P Natl Acad Sci USA*, 100(17):10020–10025.

Duret, L. (2000). tRNA gene number and codon usage in the <i>c. elegans<\i> genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet.*, 16(7): 287–289.

Duret, L. et Mouchiroud, D. (1999). Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *P Natl Acad Sci USA*, 96(8):4482–4487.

Dutheil, J. et Boussau, B. (2008). Non-homogeneous models of sequence evolution in the bio++ suite of libraries and programs. *BMC Evol Biol*, 8(1):255.

Dwight Kuo, P., Banzhaf, W. et Leier, A. (2006). Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplication and divergence. *Biosystems*, 85(3):177–200.

Edgar, R. C. (2004). MUSCLE : multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5):1792–1797.

Eigen, M. (1971). Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58(10):465–523.

Elena, S. F. et Sanjuán, R. (2008). The effect of genetic robustness on evolvability in digital organisms. *BMC Evol Biol*, 8(1):284.

Espinosa-Soto, C. et Wagner, A. (2010). Specialization can drive the evolution of modularity. *Plos Comput Biol*, 6(3):e1000719.

Ewing, G. et Hermisson, J. (2010). MSMS : a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, 26(16):2064–2065.

Excoffier, L., Novembre, J. et Schneider, S. (2000). Computer note. SIMCOAL : a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. *J Hered*, 91(6):506–509.

Eyre-Walker, A. et Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nat Rev Genet*, 8(8):610–618.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences : A maximum likelihood approach. *J Mol Evol*, 17(6):368–376.

Felsenstein, J. (1985). Phylogenies and the comparative method. *Am Nat*, 125(1):1–15.

Felsenstein, J. (1993). *PHYLIP : phylogenetic inference package, version 3.5c.*

Felsenstein, J. (2002). *PHYLIP : Phylogeny Inference Package, version 3.6a3.*

Felsenstein, J. (2005). *Theoretical evolutionary genetics.* University of Washington.

Fischer, S. (2013). *Modélisation de l'évolution de la taille des génomes et de leur densité en gènes par mutations locales et grands réarrangements chromosomiques.* Thèse de doctorat, INSA de Lyon.

Fisher, R. A. (1922). On the dominance ratio. *P Roy Soc Edinb*, (42):321–341.

Fisher, R. A. (1930). The distribution of gene ratios for rare mutations. *P Roy Soc Edinb*, 50:205–220. Reproduced with permission of the Royal Society of Edinburgh.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Syst Biol*, 19(2):99–113.

Fletcher, W. et Yang, Z. (2009). INDELible : A flexible simulator of biological sequence evolution. *Mol Biol Evol*, 26(8):1879–1888.

Fletcher, W. et Yang, Z. (2010). The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol Biol Evol*, 27(10):2257–2267.

Flombaum, P., Gallegos, J. L., Gordillo, R. A., Rincón, J., Zabala, L. L., Jiao, N., Karl, D. M., Li, W. K. W., Lomas, M. W., Veneziano, D., Vera, C. S., Vrugt, J. A. et Martiny, A. C. (2013). Present and future global distributions of the marine cyanobacteria *Prochlorococcus* and *Synechococcus*. *P Natl Acad Sci USA*, 110(24):9824–9829.

Floreano, D., Mitri, S., Magnenat, S. et Keller, L. (2007). Evolutionary conditions for the emergence of communication in robots. *Current Biology*, 17(6):514–519.

Frank, A. C. et Lobry, J. R. (2000). Oriloc : prediction of replication boundaries in unannotated bacterial chromosomes. *Bioinformatics*, 16(6):560–561.

French, S. (1992). Consequences of replication fork movement through transcription units in vivo. *Science*, 258(5086):1362–1365.

Frénoy, A., Taddei, F. et Misevic, D. (2012). Robustness and evolvability of cooperation. *In Proceedings of Artificial Life XIII*, volume 12, pages 53–58. MIT Press.

Frénoy, A., Taddei, F. et Misevic, D. (2013). Genetic architecture promotes the evolution and maintenance of cooperation. *Plos Comput Biol*, 9(11):e1003339.

García-Fernández, J. M., Marsac, N. T. d. et Diez, J. (2004). Streamlined regulation and gene loss as adaptive mechanisms in *Prochlorococcus* for optimized nitrogen utilization in oligotrophic environments. *Microbiol Mol Biol Rev*, 68(4):630–638.

Garland, T. et Carter, P. A. (1994). Evolutionary physiology. *Annu Rev Physiol*, 56(1):579–621.

Gautier, C. (2000). Compositional bias in DNA. *Curr Opin Genet Dev*, 10(6):656–661.

Giovannoni, S. J., Cameron Thrash, J. et Temperton, B. (2014). Implications of streamlining theory for microbial ecology. *ISME J*, 8:1553–1565.

Giovannoni, S. J., Tripp, H. J., Givan, S., Podar, M., Vergin, K. L., Baptista, D., Bibbs, L., Eads, J., Richardson, T. H., Noordewier, M., Rappé, M. S., Short, J. M., Carrington, J. C. et Mathur, E. J. (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *Science*, 309(5738):1242–1245.

Gouy, M. et Delmotte, S. (2008). Remote access to ACNUC nucleotide and protein sequence databases at PBIL. *Biochimie*, 90(4):555–562.

Gouy, M. et Gautier, C. (1982). Codon usage in bacteria : correlation with gene expressivity. *Nucleic Acids Res*, 10(22):7055–7074.

Gouy, M. et Grantham, R. (1980). Polypeptide elongation and tRNA cycling in *Escherichia coli* : a dynamic approach. *FEBS Lett*, 115(2):151–155.

Grosjean, H. et Fiers, W. (1982). Preferential codon usage in prokaryotic genes : the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene*, 18(3):199–209.

Guéguen, L., Gaillard, S., Boussau, B., Gouy, M., Groussin, M., Rochette, N. C., Bigot, T., Fournier, D., Pouyet, F., Cahais, V., Bernard, A., Scornavacca, C., Nabholz, B., Haudry, A., Dachary, L., Galtier, N., Belkhir, K. et Dutheil, J. Y. (2013). Bio ++ : Efficient extensible libraries and tools for computational molecular evolution. *Mol Biol Evol*, page mst097.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W. et Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies : Assessing the performance of PhyML 3.0. *Syst Biol*, 59(3):307–321.

Guindon, S. et Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, 52(5):696–704.

Gupta, S. K. et Ghosh, T. C. (2001). Gene expressivity is the main factor in dictating the codon usage variation among the genes in *Pseudomonas aeruginosa*. *Gene*, 273(1):63–70.

Hahn, M. W., Bie, T. D., Stajich, J. E., Nguyen, C. et Cristianini, N. (2005). Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res*, 15(8):1153–1160.

Hall, B. G. (2008). Simulating DNA coding sequence evolution with EvolveAGene 3. *Mol Biol Evol*, 25(4):688–695.

Hallam, A. et Cohen, J. M. (1989). The case for sea-level change as a dominant causal factor in mass extinction of marine invertebrates. *Phil Trans R Soc Lond B*, 325(1228):437–455.

Hanage, W. P., Spratt, B. G., Turner, K. M. E. et Fraser, C. (2006). Modelling bacterial speciation. *Philos Trans R Soc Lond B Biol Sci*, 361(1475):2039–2044.

Hansen, A. K. et Moran, N. A. (2012). Altered tRNA characteristics and 3' maturation in bacterial symbionts with reduced genomes. *Nucleic Acids Res*, 40(16):7870–7884.

Haq, B. U., Hardenbol, J. et Vail, P. R. (1987). Chronology of fluctuating sea levels since the triassic. *Science*, 235(4793):1156–1167.

Hasegawa, M., Kishino, H. et Yano, T.-a. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, 22(2):160–174.

Hernandez, R. D. (2008). A flexible forward simulator for populations subject to selection and demography. *Bioinformatics*, 24(23):2786–2787.

Hershberg, R. et Petrov, D. A. (2009). General rules for optimal codon choice. *Plos Genet*, 5(7):e1000556.

Hindré, T., Knibbe, C., Beslon, G. et Schneider, D. (2012). New insights into bacterial adaptation through *in vivo*<|*i*> *and* <*i*>*in silico* experimental evolution. *Nat Rev Microbiol*, 10(5):352–365.

Hoban, S., Bertorelle, G. et Gaggiotti, O. E. (2012). Computer simulations : tools for population and evolutionary genetics. *Nat Rev Genet*, 13(2):110–122.

Hu, J. et Blanchard, J. L. (2009). Environmental sequence data from the sargasso sea reveal that the characteristics of genome reduction in *Prochlorococcus* are not a harbinger for an escalation in genetic drift. *Mol Biol Evol*, 26(1):5–13.

Hurst, L. D. (1995). Evolutionary genetics. the silence of the genes. *Curr Biol*, 5(5):459–461.

Ikemura, T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes : a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J Mol Biol*, 151(3):389–409.

Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol*, 2(1):13–34.

Itoh, T., Martin, W. et Nei, M. (2002). Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. *P Natl Acad Sci USA*, 99(20):12944–12948.

Jameson, E., Joint, I., Mann, N. H. et Mühling, M. (2008). Application of a novel rpoC1-RFLP approach reveals that marine *Prochlorococcus* populations in the atlantic gyres are composed of greater microdiversity than previously described. *Microb Ecol*, 55(1):141–151.

Jenkins, D. J. et Stekel, D. J. (2010). De novo evolution of complex, global and hierarchical gene regulatory mechanisms. *J Mol Evol*, 71(2):128–140.

Jermiin, L. S., Ho, S. Y. W., Ababneh, F., Robinson, J. et Larkum, A. W. D. (2003). Hetero : a program to simulate the evolution of DNA on a four-taxon tree. *Appl Bioinformatics*, 2(3):159–163.

Johnson, Z. I., Zinser, E. R., Coe, A., McNulty, N. P., Woodward, E. M. S. et Chisholm, S. W. (2006). Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science*, 311(5768):1737–1740.

Jordan, G. et Goldman, N. (2012). The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol Biol Evol*, 29(4):1125–1139.

Jukes, T. et Cantor, C. (1969). Evolution of protein molecules. *In* Munro, M., éditeur : *Mammalian protein metabolism*, volume III, pages 21–132. Academic Press.

Kanaya, S., Yamada, Y., Kudo, Y. et Ikemura, T. (1999). Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs : gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene*, 238(1):143–155.

Kaneko, K. (2011). Proportionality between variances in gene expression induced by noise and mutation : consequence of evolutionary robustness. *BMC Evol Biol*, 11(1):27.

Kano-Sueoka, T., Lobry, J. R. et Sueoka, N. (1999). Intra-strand biases in bacteriophage t4 genome. *Gene*, 238(1):59–64.

Kashtan, N. et Alon, U. (2005). Spontaneous evolution of modularity and network motifs. *Proc Natl Acad Sci USA*, 102(39):13773–13778.

Kashtan, N., Roggensack, S. E., Rodrigue, S., Thompson, J. W., Biller, S. J., Coe, A., Ding, H., Marttinen, P., Malmstrom, R. R., Stocker, R., Follows, M. J., Stepanauskas, R. et Chisholm, S. W. (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science*, 344(6182):416–420.

Katoh, K., Kuma, K.-i., Toh, H. et Miyata, T. (2005). MAFFT version 5 : improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*, 33(2):511–518.

Kawabe, A. et Miyashita, N. T. (2003). Patterns of codon usage bias in three dicot and four monocot plant species. *Genes Gent Syst*, 78(5):343–352.

Kenyon, L. J. et Sabree, Z. L. (2014). Obligate insect endosymbionts exhibit increased ortholog length variation and loss of large accessory proteins concurrent with genome shrinkage. *Genome Biol Evol*, 6(4):763–775.

Kettler, G. C., Martiny, A. C., Huang, K., Zucker, J., Coleman, M. L., Rodrigue, S., Chen, F., Lapidus, A., Ferriera, S., Johnson, J., Steglich, C., Church, G. M., Richardson, P. et Chisholm, S. W. (2007). Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *Plos Genet*, 3(12):e231.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, 16(2):111–120.

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248.

Kissling, G. E., Grogan, D. W. et Drake, J. W. (2013). Confounders of mutation-rate estimators : Selection and phenotypic lag in *Thermus thermophilus*. *Mut Res Fundam Mol Mech Mutagen*, 749(1–2):16–20.

Knibbe, C. (2006). *Structuration des génomes par sélection indirecte de la variabilité mutationnelle : une approche de modélisation et de simulation*. Thèse de doctorat, INSA de Lyon.

Knibbe, C., Coulon, A., Mazet, O., Fayard, J.-M. et Beslon, G. (2007a). A long-term evolutionary pressure on the amount of noncoding DNA. *Mol Biol Evol*, 24(10): 2344–2353.

Knibbe, C., Mazet, O., Chaudier, F., Fayard, J.-M. et Beslon, G. (2007b). Evolutionary coupling between the deleteriousness of gene mutations and the amount of non-coding sequences. *J Theor Biol*, 244(4):621–630.

Koonin, E. V. (2004). A non-adaptationist perspective on evolution of genomic complexity or the continued dethroning of man. *Cell Cycle*, 3(3):280–285.

Krishnan, A., Tomita, M. et Giuliani, A. (2008). Evolution of gene regulatory networks : Robustness as an emergent property of evolution. *Physica A*, 387(8–9):2170–2186.

Kuo, C.-H., Moran, N. A. et Ochman, H. (2009). The consequences of genetic drift for bacterial genome complexity. *Genome Res*, 19(8):1450–1454.

Kuo, C.-H. et Ochman, H. (2009). Deletional bias across the three domains of life. *Genome Biol Evol*, 1:145–152.

Kyte, J. et Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, 157(1):105–132.

Lafay, B., Atherton, J. C. et Sharp, P. M. (2000). Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*. *Microbiology*, 146(4):851–860.

Lafay, B., Sharp, P. M., Lloyd, A. T., McLean, M. J., Devine, K. M. et Wolfe, K. H. (1999). Proteome composition and codon usage in *Spirochaetes* : Species-specific and DNA strand-specific mutational biases. *Nucleic Acids Res*, 27(7):1642–1649.

Lanave, C., Preparata, G., Sacone, C. et Serio, G. (1984). A new method for calculating evolutionary substitution rates. *J Mol Evol*, 20(1):86–93.

Lenski, R. E., Ofria, C., Pennock, R. T. et Adami, C. (2003). The evolutionary origin of complex features. *Nature*, 423(6936):139–144.

Limor-Waisberg, K., Carmi, A., Scherz, A., Pilpel, Y. et Furman, I. (2011). Specialization versus adaptation : two strategies employed by cyanophages to enhance their translation efficiencies. *Nucleic Acids Res*, 39(14):6016–6028.

Lindell, D., Sullivan, M. B., Johnson, Z. I., Tolonen, A. C., Rohwer, F. et Chisholm, S. W. (2004). Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *P Natl Acad Sci USA*, 101(30):11013–11018.

Lobry, J. R. (1996). A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. *Biochimie*, 78(5):323–326.

Lobry, J. R. et Chessel, D. (2003). Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria. *J Appl Genet*, 44:235–261.

Lowe, T. M. et Eddy, S. R. (1997). tRNAscan-SE : A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*, 25(5):0955–964.

Löytynoja, A. et Goldman, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. *P Natl Acad Sci USA*, 102(30):10557–10562.

Luo, H., Friedman, R., Tang, J. et Hughes, A. L. (2011). Genome reduction by deletion of paralogs in the marine cyanobacterium *Prochlorococcus*. *Mol Biol Evol*, 28(10):2751–2760.

Luo, H., Shi, J., Arndt, W., Tang, J. et Friedman, R. (2008). Gene order phylogeny of the genus *Prochlorococcus*. *Plos One*, 3(12):e3837.

Lynch, M. (2006). The origins of eukaryotic gene structure. *Mol Biol Evol*, 23(2):450–468.

Lynch, M. (2007). *The Origins of Genome Architecture*. Sinauer Associates Inc.

Lynch, M. (2011). Statistical inference on the mechanisms of genome evolution. *Plos Genet*, 7(6):e1001389.

Lynch, M. (2012). The evolution of multimeric protein assemblages. *Mol Biol Evol*, 29(5):1353–1366.

Lynch, M. et Abegg, A. (2010). The rate of establishment of complex adaptations. *Mol Biol Evol*, 27(6):1404–1414.

Lynch, M., Bobay, L.-M., Catania, F., Gout, J.-F. et Rho, M. (2011). The repatterning of eukaryotic genomes by random genetic drift. *Annu Rev Genomics Hum Genet*, 12(1):347–366.

Lynch, M. et Conery, J. S. (2003). The origins of genome complexity. *Science*, 302(5649):1401–1404.

Lynch, M., Koskella, B. et Schaack, S. (2006). Mutation pressure and the evolution of organelle genomic architecture. *Science*, 311(5768):1727–1730.

Malmstrom, R. R., Coe, A., Kettler, G. C., Martiny, A. C., Frias-Lopez, J., Zinser, E. R. et Chisholm, S. W. (2010). Temporal dynamics of *Prochlorococcus* ecotypes in the atlantic and pacific oceans. *ISME J*, 4(10):1252–1264.

Marais, G. A. B., Calteau, A. et Tenaillon, O. (2008). Mutation rate and genome reduction in endosymbiotic and free-living bacteria. *Genetica*, 134(2):205–210.

Mary, I., Tu, C.-J., Grossman, A. et Vaulot, D. (2004). Effects of high light on transcripts of stress-associated genes for the cyanobacteria *Synechocystis* sp. PCC 6803 and *Prochlorococcus* MED4 and MIT9313. *Microbiology*, 150(5):1271–1281.

Mattiussi, C. et Floreano, D. (2007). Analog genetic encoding for the evolution of circuits and networks. *IEEE T Evolut Comput*, 11(5):596–607.

Maynard Smith, J. (1983). Models of evolution. *Proc R Soc Lond B Biol Sci*, 219:315–325.

McCutcheon, J. P. et Moran, N. A. (2012). Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol*, 10(1):13–26.

McGrath, C. L. et Katz, L. A. (2004). Genome diversity in microbial eukaryotes. *Trends Ecol Evol*, 19(1):32–38.

McInerney, J. O. (1997). Prokaryotic genome evolution as assessed by multivariate analysis of codon usage patterns. *Microb Comp Genomics*, 2(1):89–97.

McInerney, J. O. (1998). Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. *P Natl Acad Sci USA*, 95(18):10698–10703.

McLean, M. J., Wolfe, K. H. et Devine, K. M. (1998). Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J Mol Evol*, 47(6):691–696.

Médigue, C., Rouxel, T., Vigier, P., Hénaut, A. et Danchin, A. (1991). Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J Mol Biol*, 222(4):851–856.

Memon, D., Singh, A. K., Pakrasi, H. B. et Wangikar, P. P. (2013). A global analysis of adaptive evolution of operons in cyanobacteria. *Antonie van Leeuwenhoek*, 103(2):331–346.

Michalik, J. (2014). *Application de méthodes de comptage probabiliste pour estimer l'évolution et l'hétérogénéité temporelle du biais d'usage des codons de souches de Prochlorococcus et Synechococcus*. Thèse de doctorat.

Mira, A. et Moran, N. A. (2002). Estimating population size and transmission bottlenecks in maternally transmitted endosymbiotic bacteria. *Microb Ecol*, 44(2):137–143.

Mira, A., Ochman, H. et Moran, N. A. (2001). Deletional bias and the evolution of bacterial genomes. *Trends Genet*, 17(10):589–596.

Misevic, D., Frénoy, A., Parsons, D. P. et Taddei, F. (2012). Effects of public good properties on the evolution of cooperation. *In Proceedings of Artificial Life XIII*, volume 12, pages 218–225. MIT Press.

Misevic, D., Ofria, C. et Lenski, R. E. (2006). Sexual reproduction reshapes the genetic architecture of digital organisms. *Proc R Soc Lond B Biol Sci*, 273(1585):457–464.

Moore, L. R., Coe, A., Zinser, E. R., Saito, M. A., Sullivan, M. B., Lindell, D., Frois-Moniz, K., Waterbury, J. et Chisholm, S. W. (2007). Culturing the marine cyanobacterium *Prochlorococcus*. *Limnol Oceanogr-Meth*, 5:353–362.

Moran, N. A. (1996). Accelerated evolution and muller's rachet in endosymbiotic bacteria. *P Natl Acad Sci USA*, 93(7):2873–2878.

Moran, N. A., McCutcheon, J. P. et Nakabachi, A. (2008). Genomics and evolution of heritable bacterial symbionts. *Annu Rev Genet*, 42(1):165–190.

Moran, N. A., McLaughlin, H. J. et Sorek, R. (2009). The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science*, 323(5912):379–382.

Morris, J. J., Lenski, R. E. et Zinser, E. R. (2012). The black queen hypothesis : Evolution of dependencies through adaptive gene loss. *mBio*, 3(2).

Morrison, D. A. (2009). A framework for phylogenetic sequence alignment. *Plant Syst Evol*, 282(3-4):127–149.

Morton, B. R. (1997). Rates of synonymous substitution do not indicate selective constraints on the codon use of the plant psbA gene. *Mol Biol Evol*, 14(4):412–419.

Mozhayskiy, V. et Tagkopoulos, I. (2012a). Guided evolution of *in silico* microbial populations in complex environments accelerates evolutionary rates through a step-wise adaptation. *BMC Bioinformatics*, 13(Suppl 10):S10.

Mozhayskiy, V. et Tagkopoulos, I. (2012b). Horizontal gene transfer dynamics and distribution of fitness effects during microbial *in silico* evolution. *BMC Bioinformatics*, 13(Suppl 10):S13.

Muller, H. (1964). The relation of recombination to mutational advance. *Mutat Res-Fund Mol M*, 1(1):2–9.

Nei, M. et Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*, 3(5):418–426.

Nelson, C. W. et Sanford, J. C. (2011). The effects of low-impact mutations in digital organisms. *Theor Biol Med Model*, 8(1):1–17.

Nielsen, R., DuMont, V. L. B., Hubisz, M. J. et Aquadro, C. F. (2007). Maximum likelihood estimation of ancestral codon usage bias parameters in *Drosophila*. *Mol Biol Evol*, 24(1):228–235.

Novembre, J. A. (2002). Accounting for background nucleotide composition when measuring codon usage bias. *Mol Biol Evol*, 19(8):1390–1394.

Ochman, H. (2002). Distinguishing the ORFs from the ELFs : short bacterial genes and the annotation of genomes. *Trends Genet*, 18(7):335–337.

O'Fallon, B. (2008). Population structure, levels of selection, and the evolution of intracellular symbionts. *Evolution*, 62(2):361–373.

Ofria, C. et Wilke, C. O. (2004). Avida : A software platform for research in computational evolutionary biology. *Artificial Life*, 10(2):191–229.

Ohta, T. (1972). Fixation probability of a mutant influenced by random fluctuation of selection intensity. *Genet Res*, 19(01):33–38.

Osburne, M. S., Holmbeck, B. M., Coe, A. et Chisholm, S. W. (2011). The spontaneous mutation frequencies of *Prochlorococcus* strains are commensurate with those of other bacteria. *Environ Microbiol Rep*, 3(6):744–749.

Pál, C. et Hurst, L. D. (2004). Evidence against the selfish operon theory. *Trends Genet*, 20(6):232–234.

Pál, C., Papp, B. et Hurst, L. D. (2001). Highly expressed genes in yeast evolve slowly. *Genetics*, 158(2):927–931.

Pang, A., Smith, A. D., Nuin, P. A. et Tillier, E. R. (2005). SIMPROT : Using an empirically determined indel distribution in simulations of protein evolution. *BMC Bioinformatics*, 6(1):236.

Parsons, D. (2011). *Sélection indirecte en évolution darwinienne : mécanismes et implications*. Thèse de doctorat, INSA de Lyon.

Parsons, D. P., Knibbe, C. et Beslon, G. (2010). Importance of the rearrangement rates on the organization of genome transcription. *In Proceedings of Artificial Life XII*, pages 479–486. MIT Press.

Partensky, F. et Garczarek, L. (2010). *Prochlorococcus* : Advantages and limits of minimalism. *Annu Rev Mar Sci*, 2(1):305–331.

Partensky, F., Hess, W. R. et Vaulot, D. (1999). *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol Mol Biol Rev*, 63(1):106–127.

Paul, S., Dutta, A., Bag, S. K., Das, S. et Dutta, C. (2010). Distinct, ecotype-specific genome and proteome signatures in the marine cyanobacteria *Prochlorococcus*. *BMC Genomics*, 11(1):103.

Penel, S., Arigon, A.-M., Dufayard, J.-F., Sertier, A.-S., Daubin, V., Duret, L., Gouy, M. et Perrière, G. (2009). Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*, 10(Suppl 6):S3.

Percudani, R., Pavesi, A. et Ottonello, S. (1997). Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. *J Mol Biol*, 268(2):322–330.

Pérez-Brocal, V., Gil, R., Ramos, S., Lamelas, A., Postigo, M., Michelena, J. M., Silva, F. J., Moya, A. et Latorre, A. (2006). A small microbial genome : The end of a long symbiotic relationship ? *Science*, 314(5797):312–313.

Peris, J. B., Davis, P., Cuevas, J. M., Nebot, M. R. et Sanjuán, R. (2010). Distribution of fitness effects caused by single-nucleotide substitutions in bacteriophage f1. *Genetics*, 185(2):603–609.

Perrière, G. et Thioulouse, J. (2002). Use and misuse of correspondence analysis in codon usage studies. *Nucleic Acids Res*, 30(20):4548–4555.

Pettersson, M. E. et Berg, O. G. (2007). Muller's ratchet in symbiont populations. *Genetica*, 130(2):199–211.

Pouyet, F., Jacquemetton, J., Bailly-Bechet, M. et Guégen, L. (2013). Codon usage in *Escherichia coli* : an evolutionary approach.

Price, M. N., Arkin, A. P. et Alm, E. J. (2006). The life-cycle of operons. *Plos Genet*, 2(6):e96.

Price, M. N., Huang, K. H., Arkin, A. P. et Alm, E. J. (2005). Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res*, 15(6):809–819.

Rambaut, A. et Grass, N. C. (1997). Seq-gen : an application for the monte carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci*, 13(3):235–238.

Reis, M. d., Savva, R. et Wernisch, L. (2004). Solving the riddle of codon usage preferences : a test for translational selection. *Nucleic Acids Res*, 32(17):5036–5044.

Rispe, C. et Moran, N. A. (2000). Accumulation of deleterious mutations in endosymbionts : Muller's ratchet with two levels of selection. *Am Nat*, 156(4):425–441.

Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N. A., Arellano, A., Coleman, M., Hauser, L., Hess, W. R., Johnson, Z. I., Land, M., Lindell, D., Post, A. F., Regala, W., Shah, M., Shaw, S. L., Steglich, C., Sullivan, M. B., Ting, C. S., Tolonen, A., Webb, E. A., Zinser, E. R. et Chisholm, S. W. (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature*, 424(6952):1042–1047.

Rocha, E. P. et Danchin, A. (2002). Base composition bias might result from competition for metabolic resources. *Trends Genet*, 18(6):291–294.

Rocha, E. P. C. (2004). Codon usage bias from tRNA's point of view : Redundancy, specialization, and efficient decoding for translation optimization. *Genome Res.*, 14(11): 2279–2286.

Rocha, E. P. C., Danchin, A. et Viari, A. (1999). Universal replication biases in bacteria. *Mol Microbiol*, 32(1):11–16.

Romero, H., Zavala, A. et Musto, H. (2000). Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. *Nucleic Acids Res*, 28(10):2084–2090.

Rosenberg, M. S. (2007). MySSP : Non-stationary evolutionary sequence simulation, including indels. *Evol Bioinform Online*, 1:81–83.

Ross, C. A. et Ross, J. R. P. (1986). Sea-level changes : An integrated approach. *Geology*, 14(6):535–535.

Sanchez-Dehesa, Y. (2009). *RAevol : un modèle de génétique digital pour étudier l'évolution des réseaux de régulation génétique.* Thèse de doctorat, INSA de Lyon.

Sanjuán, R. (2010). Mutational fitness effects in RNA and single-stranded DNA viruses : common patterns revealed by site-directed mutagenesis studies. *Phil Trans R Soc Lond B*, 365(1548):1975–1982.

Sanjuán, R., Moya, A. et Elena, S. F. (2004). The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proc Natl Acad Sci USA*, 101(22):8396–8401.

Scanlan, D. J., Hess, W. R., Partensky, F., Newman, J. et Vaulot, D. (1996). High degree of genetic variation in *Prochlorococcus* (*Prochlorophyta*) revealed by RFLP analysis. *Eur J Phycol*, 31(1):1–9.

Sharp, P. M. et Li, W.-H. (1987). The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*, 15(3): 1281–1295.

Sharp, P. M., Tuohy, T. M. F. et Mosurski, K. R. (1986). Codon usage in yeast : cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res*, 14(13):5125–5143.

Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y. et Ishikawa, H. (2000). Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature*, 407(6800):81–86.

Sipos, B., Massingham, T., Jordan, G. E. et Goldman, N. (2011). PhyloSim - monte carlo simulation of sequence evolution in the r statistical computing environment. *BMC Bioinformatics*, 12(1):104.

Soyer, O. S. et Bonhoeffer, S. (2006). Evolution of complexity in signaling pathways. *Proc Natl Acad Sci USA*, 103(44):16337–16342.

Spencer, C. C. A. et Coop, G. (2004). SelSim : a program to simulate population genetic data with natural selection and recombination. *Bioinformatics*, 20(18):3673–3675.

Stamatakis, A. (2006). RAxML-VI-HPC : maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690.

Stanley, D., Fraser, S., Chambers, P. J., Rogers, P. et Stanley, G. A. (2010). Generation and characterisation of stable ethanol-tolerant mutants of *Saccharomyces cerevisiae*. *J Ind Microbiol Biotechnol*, 37(2):139–149.

Stoye, J., Evers, D. et Meyer, F. (1998). Rose : generating sequence families. *Bioinformatics*, 14(2):157–163.

Sueoka, N. (1961). Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. *Proc Natl Acad Sci USA*, 47(8):1141–1149.

Sueoka, N. (1962). On the genetic basis of variation and heterogeneity of DNA base composition. *Proc Natl Acad Sci USA*, 48(4):582–592.

Sueoka, N. (1988). Directional mutation pressure and neutral molecular evolution. *P Natl Acad Sci USA*, 85(8):2653–2657.

Sueoka, N. (1995). Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J Mol Evol*, 40(3):318–325.

Sullivan, M. B., Coleman, M. L., Weigele, P., Rohwer, F. et Chisholm, S. W. (2005). Three *Prochlorococcus* cyanophage genomes : Signature features and ecological interpretations. *Plos Biol*, 3(5):e144.

Sullivan, M. B., Waterbury, J. B. et Chisholm, S. W. (2003). Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*. *Nature*, 424(6952):1047–1051.

Sun, Z. et Blanchard, J. L. (2014). Strong genome-wide selection early in the evolution of *Prochlorococcus* resulted in a reduced genome through the loss of a large number of small effect genes. *Plos One*, 9(3):e88837.

Swofford, D. (2003). *PAUP. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4*. Sinauer Associates.

Taddei, F., Radman, M., Maynard-Smith, J., Toupance, B., Gouyon, P. H. et Godelle, B. (1997). Role of mutator alleles in adaptive evolution. *Nature*, 387(6634): 700–702.

Tagkopoulos, I., Liu, Y.-C. et Tavazoie, S. (2008). Predictive behavior within microbial genetic networks. *Science*, 320(5881):1313–1317.

Takuno, S., Kado, T., Sugino, R. P., Nakhleh, L. et Innan, H. (2012). Population genomics in bacteria : A case study of *Staphylococcus aureus*. *Mol Biol Evol*, 29(2):797–809.

Talavera, G. et Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol*, 56(4):564–577.

Talevich, E., Invergo, B. M., Cock, P. J. et Chapman, B. A. (2012). Bio.phylo : A unified toolkit for processing, analyzing and visualizing phylogenetic trees in biopython. *BMC Bioinformatics*, 13(1):209.

Tamas, I., Klasson, L., Canbäck, B., Näslund, A. K., Eriksson, A.-S., Wernegreen, J. J., Sandström, J. P., Moran, N. A. et Andersson, S. G. E. (2002). 50 million years of genomic stasis in endosymbiotic bacteria. *Science*, 296(5577):2376–2379.

Tamura, K. et Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol*, 10(3):512–526.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J. et Natale, D. A. (2003). The COG database : an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41.

Tatusov, R. L., Galperin, M. Y., Natale, D. A. et Koonin, E. V. (2000). The COG database : a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res*, 28(1):33–36.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *American Mathematical Society : Lectures on Mathematics in the Life Sciences*, 17:57–86.

Tenaillon, O., Toupance, B., Nagard, H. L., Taddei, F. et Godelle, B. (1999). Mutators, population size, adaptive landscape and the adaptation of asexual populations of bacteria. *Genetics*, 152(2):485–493.

Thioulouse, J., Chessel, D., Dolédec, S. et Olivier, J.-M. (1997). ADE4 : a multivariate analysis and graphical display software. *Stat Comput*, 7(1):75–83.

Thomas, C. A. (1971). The genetic organization of chromosomes. *Annual Review of Genetics*, 5(1):237–256.

Thompson, C. C., Silva, G. G. Z., Vieira, N. M., Edwards, R., Vicente, A. C. P. et Thompson, F. L. (2013). Genomic taxonomy of the genus *Prochlorococcus*. *Microb Ecol*, pages 1–11.

Thompson, J. D., Higgins, D. G. et Gibson, T. J. (1994). CLUSTAL w : improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22): 4673–4680.

Toft, C. et Andersson, S. G. E. (2010). Evolutionary microbial genomics : insights into bacterial host adaptation. *Nat Rev Genet*, 11(7):465–475.

Tolonen, A. C., Aach, J., Lindell, D., Johnson, Z. I., Rector, T., Steen, R., Church, G. M. et Chisholm, S. W. (2006). Global gene expression of *Prochlorococcus* ecotypes in response to changes in nitrogen availability. *Mol Syst Biol*, 2.

Toprak, E., Veres, A., Michel, J.-B., Chait, R., Hartl, D. L. et Kishony, R. (2012). Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nat Genet*, 44(1):101–105.

Tsuda, M. E. et Kawata, M. (2010). Evolution of gene regulatory networks by fluctuating selection and intrinsic constraints. *Plos Comput Biol*, 6(8):e1000873.

Tusscher, K. H. t. et Hogeweg, P. (2009). The role of genome and gene regulatory network canalization in the evolution of multi-trait polymorphisms and sympatric speciation. *BMC Evol Biol*, 9(1):159.

van Ham, R. C., Kamerbeek, J., Palacios, C., Rausell, C., Abascal, F., Bastolla, U., Fernández, J. M., Jiménez, L., Postigo, M. et Silva, F. J. (2003). Reductive genome evolution in *Buchnera aphidicola*. *P Natl Acad Sci USA*, 100(2):581–586.

Varenne, S., Buc, J., Lloubes, R. et Lazdunski, C. (1984). Translation is a non-uniform process : Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. *J Mol Biol*, 180(3):549–576.

Vicario, S., Moriyama, E. N. et Powell, J. R. (2007). Codon usage in twelve species of *Drosophila*. *BMC Evol Biol*, 7(1):226.

Viklund, J., Ettema, T. J. G. et Andersson, S. G. E. (2012). Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. *Mol Biol Evol*, 29(2):599–615.

Wagner, A. (2008). Neutralism and selectionism : a network-based reconciliation. *Nat Rev Genet*, 9(12):965–974.

Waibel, M., Floreano, D. et Keller, L. (2011). A quantitative test of hamilton's rule for the evolution of altruism. *Plos Biol*, 9(5):e1000615.

Wang, B., Lu, L., Lv, H., Jiang, H., Qu, G., Tian, C. et Ma, Y. (2014). The transcriptome landscape of *Prochlorococcus* MED4 and the factors for stabilizing the core genome. *BMC Microbiol*, 14(1):11.

Wang, M., Kurland, C. G. et Caetano-Anollés, G. (2011). Reductive evolution of proteomes and protein structures. *Proc Natl Acad Sci USA*, 108(29):11954–11958.

Wernegreen, J. J. (2002). Genome evolution in bacterial endosymbionts of insects. *Nat Rev Genet*, 3(11):850–861.

Wernegreen, J. J. et Moran, N. A. (1999). Evidence for genetic drift in endosymbionts (*Buchnera*) : analyses of protein-coding genes. *Mol Biol Evol*, 16(1):83–97.

Whitney, K. D., Boussau, B., Baack, E. J. et Garland, T. (2011). Drift and genome complexity revisited. *Plos Genet*, 7(6):e1002092.

Whitney, K. D. et Garland, T. (2010). Did genetic drift drive increases in genome complexity ? *Plos Genet*, 6(8):e1001080.

Wilke, C. O., Wang, J. L., Ofria, C., Lenski, R. E. et Adami, C. (2001). Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412(6844):331–333.

Wright, F. (1990). The 'effective number of codons' used in a gene. *Gene*, 87(1):23–29.

Wright, S. (1931). Evolution in mendelian populations. *Genetics*, 16(2):97–159.

Yang, Z. (1997). PAML : a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*, 13(5):555–556.

Yang, Z. (2007). PAML 4 : Phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, 24(8):1586–1591.

Yang, Z. et Nielsen, R. (1998). Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J Mol Evol*, 46(4):409–418.

Yang, Z. et Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol*, 17(1):32–43.

Yang, Z. et Nielsen, R. (2008). Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol*, 25(3):568–579.

Yang, Z. et Rannala, B. (2012). Molecular phylogenetics : principles and practice. *Nat Rev Genet*, 13(5):303–314.

Yooseph, S., Nealson, K. H., Rusch, D. B., McCrow, J. P., Dupont, C. L., Kim, M., Johnson, J., Montgomery, R., Ferriera, S., Beeson, K., Williamson, S. J., Tovchigrechko, A., Allen, A. E., Zeigler, L. A., Sutton, G., Eisenstadt, E., Rogers, Y.-H., Friedman, R., Frazier, M. et Venter, J. C. (2010). Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature*, 468(7320):60–66.

Yu, T., Li, J., Yang, Y., Qi, L., Chen, B., Zhao, F., Bao, Q. et Wu, J. (2012). Codon usage patterns and adaptive evolution of marine unicellular cyanobacteria *Synechococcus* and *Prochlorococcus*. *Mol Phylogenet Evol*, 62(1):206–213.

Zeidner, G., Bielawski, J. P., Shmoish, M., Scanlan, D. J., Sabehi, G. et Béjà, O. (2005). Potential photosynthesis gene recombination between *Prochlorococcus* and *Synechococcus* via viral intermediates. *Environ Microbiol*, 7(10):1505–1513.

Zhao, F. et Qin, S. (2007). Comparative molecular population genetics of phycoerythrin locus in *Prochlorococcus*. *Genetica*, 129(3):291–299.

Zhaxybayeva, O., Doolittle, W. F., Papke, R. T. et Gogarten, J. P. (2009). Intertwined evolutionary histories of marine *Synechococcus* and *Prochlorococcus marinus*. *Genome Biol Evol*, 1:325–339.

Zhaxybayeva, O., Gogarten, J. P., Charlebois, R. L., Doolittle, W. F. et Papke, R. T. (2006). Phylogenetic analyses of cyanobacterial genomes : Quantification of horizontal gene transfer events. *Genome Res.*, 16(9):1099–1108.

Zwickl, D. J. (2006). *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion*. Thèse de doctorat, Université du Texas, Austin.

Annexe A

Etude de la variation environnementale

Dans les simulations de scénarios, la cible environnementale varie au cours des générations afin d'éviter l'érosion du non-codant induite par le schéma de sélection. Cependant, l'effet de la variation environnementale sur la construction des génomes n'a pas été caractérisée dans aevol. Une étude préliminaire aux scénarios a donc été réalisée afin de pouvoir comprendre la structure des génomes dans les populations souches. Elle a été réalisée sur la cible environnementale par défaut de aevol, qui est constituée de troix gaussiennes chevauchantes (Figure A.1). De plus, la variation touchait les valeurs moyennes des gaussiennes, c'est-à-dire leur position sur l'axe, plutôt que leur hauteur, comme dans les populations souches construites pour les scénarios de la première partie. Mais, étant donné le chevauchement des gaussiennes, ce sont surtout les hauteurs des pics de la cible environnementale qui sont touchés (Figure A.1). Ainsi, cette étude peut être réutilisée pour le choix des paramètres de variation dans les populations souches.

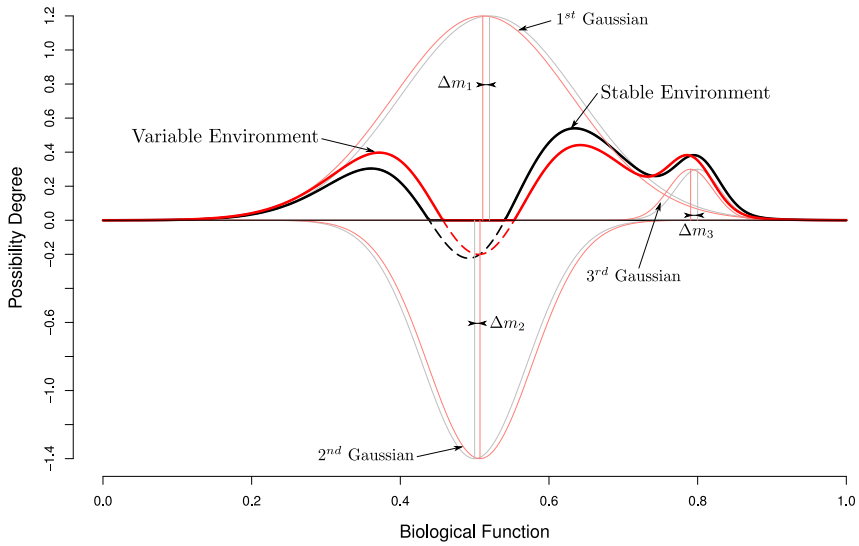


Figure A.1 – Fluctuation des moyennes des trois fonctions gaussiennes formant la distribution cible f_E

Comme mentionné précédemment, la cible environnementale est constituée de 3 gaussiennes (Figure A.1). Ces gaussiennes varient autour de valeurs moyennes, indépendamment les unes des autres, en suivant un processus autorégressif d'ordre 1 de paramètres σ et τ : $m_i(t+1) = \bar{m_i} + \Delta m_i(t+1)$ avec $\Delta m_i(t+1) = \Delta m_i(t)\left(1 - \frac{t}{\tau}\right) + \frac{\sigma}{\tau}\sqrt{2\tau - 1}\varepsilon(t)$. Les $\varepsilon(t) \sim N(0,1)$ pour chaque gaussienne sont indépendants les uns des autres et sont normalement distribués. σ représente l'amplitude de variation et τ le temps moyen de retour à la moyenne. Quels sont les influences de σ et τ sur la construction des génomes ? Afin de quantifier les effets, nous avons conduit une campagne de simulations, avec les mêmes paramètres que ceux utilisés pour les populations souches (Section III.2), à l'exception du transfert et du biais spontané favorisant les petites délétions, pendant 300 000 générations. 5 valeurs de σ et 21 valeurs de τ ont été testées. Chacun des 105 couples σ-τ est testé sur 5 répétitions, différant seulement par la graine du génération de nombres aléatoires, ont été effectuées.

La vitesse de variation de l'environnement (τ) a un impact sur de nombreux indicateurs génomiques (Figure A.2) avec une relation en forme de cloche. Pour de petits et grands τ, le génome est plus court que pour les valeurs moyennes de τ. La différence est principalement due à la quantité de bases non codantes : la proportion de bases codantes a une relation avec τ en forme de cloche inversée.

Ce phénomène est dû à une sélection indirecte du niveau de variabilité mutationnelle. En effet, les taux spontanés de mutations locales et de réarrangements étant fixés par bases, la taille du génome module la quantité totale de mutations locales et de réarrangements par reproduction et peut donc faire l'objet d'une sélection indirecte par "autostop" : lorsqu'une mutation locale ou un réarrangement avantageux est sélectionné, la taille du génome dans lequel il s'est produit est de fait sélectionné également. Plus la cible environnementale varie vite (plus τ est petit), plus des mutations locales ou des réarrangements avantageux sont fixés régulièrement et ces évènements ont statistiquement plus de chances de s'être produits dans de grands génomes. Ainsi la sélection indirecte de grands génomes est d'autant plus forte que la cible varie fréquemment, sauf au-delà d'une certaine vitesse de variation ; lorsque la cible varie tellement vite que les mutations avantageuses à une génération t ne le sont déjà plus à la suivante, la sélection devient de fait stabilisatrice plutôt que directionnelle et le phénotype se stabilise autour de la moyenne temporelle de la cible environnementale. Dans ces conditions ($\tau < 10$), les grands génomes, subissant de nombreuses mutations, sont contre-sélectionnés car les mutations sont majoritairement contre-sélectionnées.

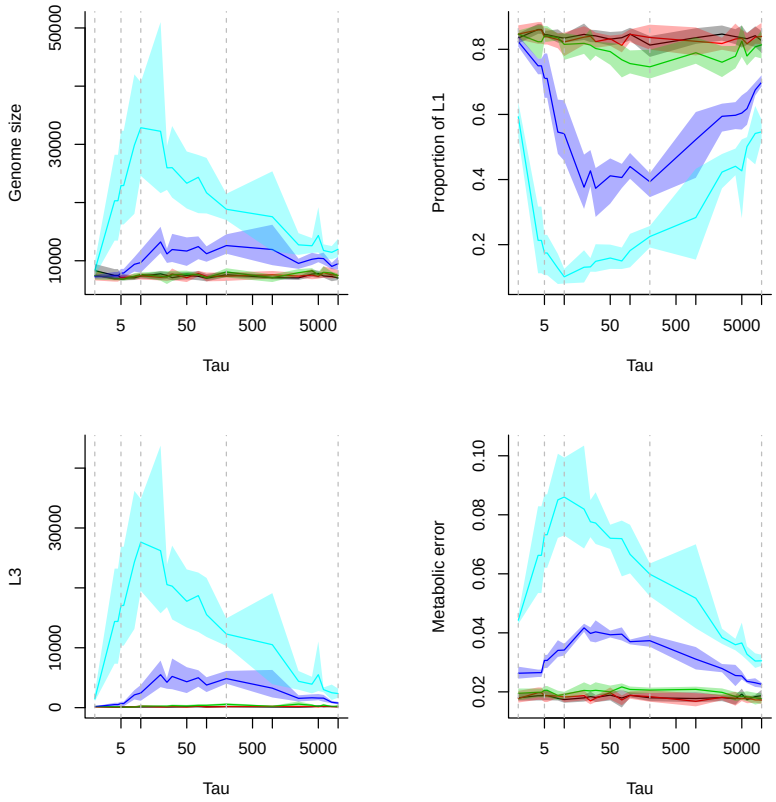
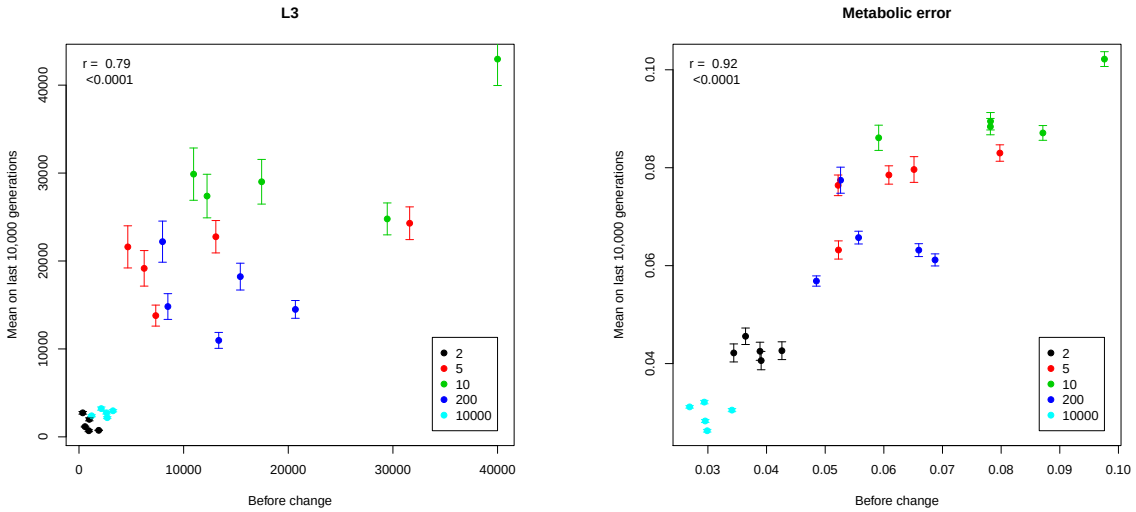


Figure A.2 – Impact de σ et τ sur la taille du génome, la proportion de bases dans des gènes codants (L1), le nombre de bases non codantes (L3) et l'erreur métabolique (différence entre phénotype et cible environnementale)
Les différentes valeurs de σ sont représentées par les différentes couleurs : noir pour $\sigma = 0.0005$, rouge pour $\sigma = 0.001$, vert pour $\sigma = 0.002$, bleu pour $\sigma = 0.005$ et cyan pour $\sigma = 0.009$. Les valeurs correspondent aux moyennes et intervalle de confiance pour les 5 simulations de chaque couples σ-τ, calculés sur une moyenne sur les 10 000 dernières générations pour le meilleur individu. Les lignes verticales en pointillées symbolisent les valeurs de τ utilisées pour les post-traitements.

σ, en revanche, ne change pas le type de la relation entre τ et les différents indicateurs (Figure A.2) mais amplifie cette relation et la décale vers les valeurs de τ plus faibles.

L'impact de τ sur les génomes est relativement robuste. En effet, dans une expérience complémentaire, nous avons supprimé tout l'ADN non codant des génomes évolués, pour les simulations menées avec $\sigma = 0.009$ et τ correspondant aux lignes verticales dans la figure A.2. En quelques générations, les différents indicateurs génomiques ont retrouvé leurs valeurs d'avant suppression (Figure A.3).

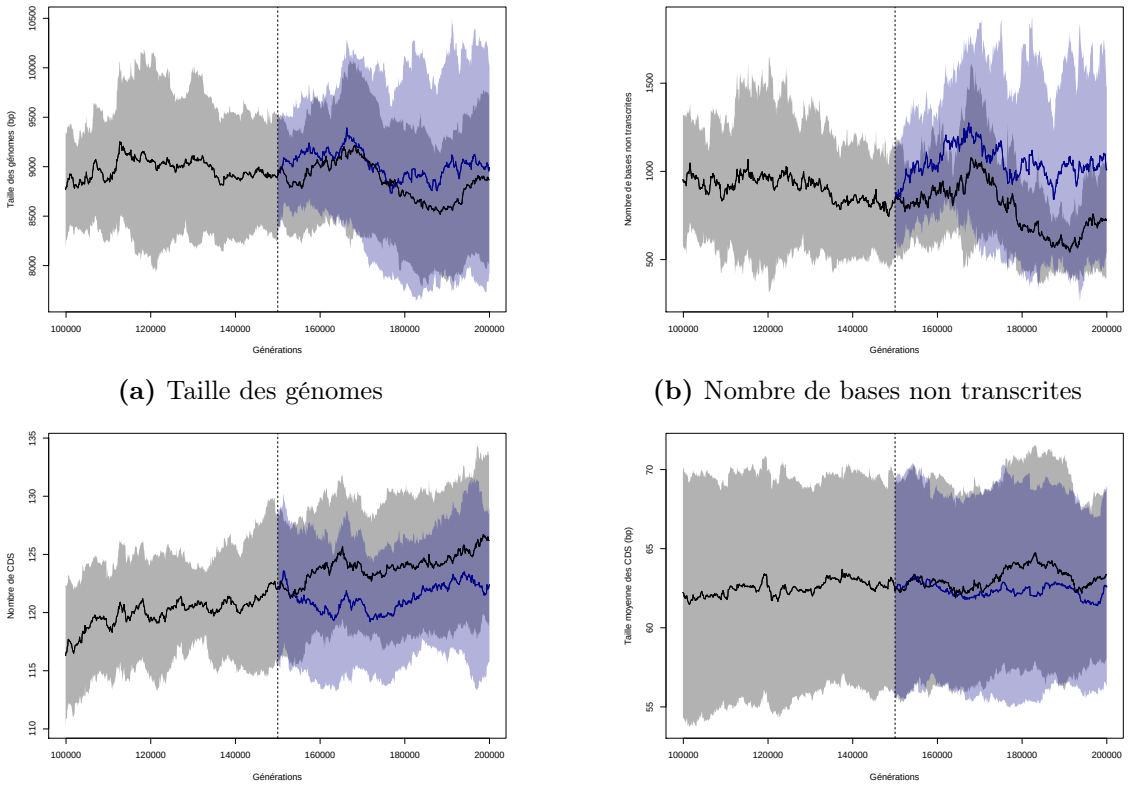


(a) Nombre de bases non transcrites **(b)** Erreur métabolique

Figure A.3 – Nombre de bases non transcrites (L3) et erreur métabolique juste avant et 100 000 générations après la suppression du non codant
Les différentes valeurs de τ sont représentées par les différentes couleurs : noir pour $\tau = 2$, rouge pour $\tau = 5$, vert pour $\tau = 10$, bleu pour $\tau = 200$ et cyan pour $\tau = 10\,000$.

Annexe B

Figures détaillées des séries temporelles des scénarios



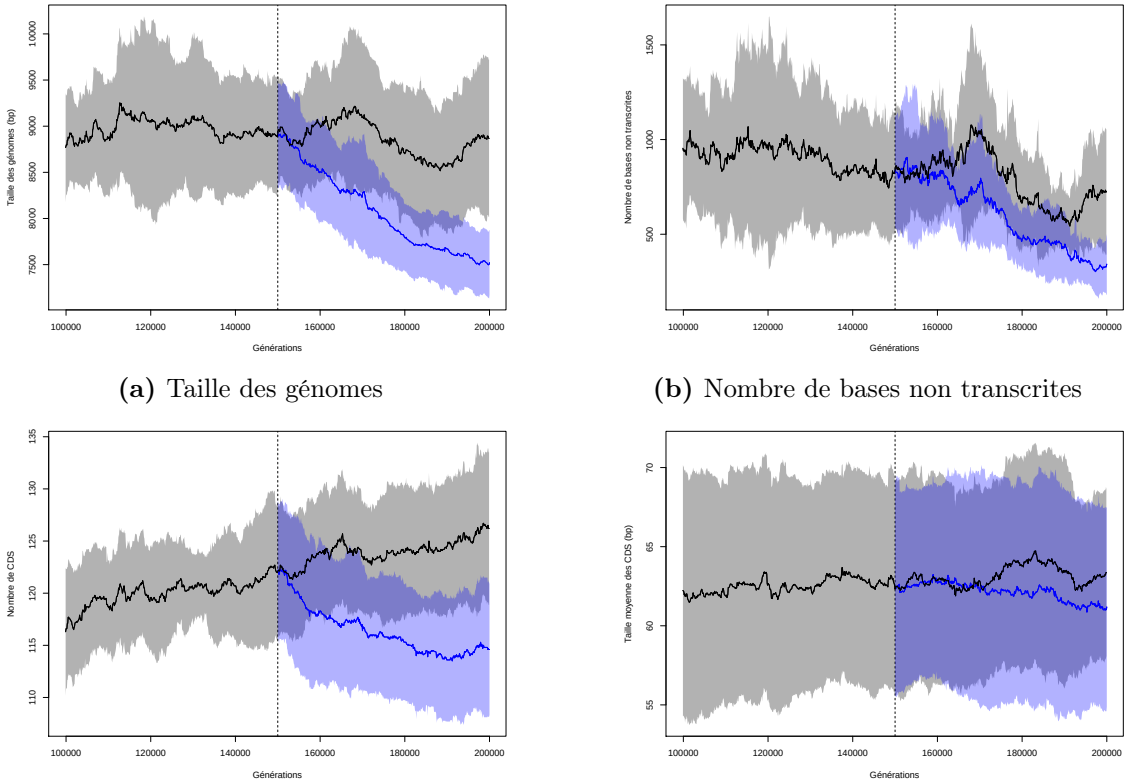
(a) Taille des génomes

(b) Nombre de bases non transcrites

(c) Nombre de CDS

(d) Taille moyenne des CDS

Figure B.1 – Évolution de certaines caractéristiques génomiques le long de la lignée du meilleur individu à la génération 200 000 pour les simulations de contrôle et celles du scénario de diminution de la taille des populations
Les lignes représentent les moyennes pour les 10 simulations de chaque scénario et la plage colorée à l'écart-type sur les 10 simulations.
Les simulations de contrôles sont en noir et celles du scénario de diminution de la taille des populations en bleu.



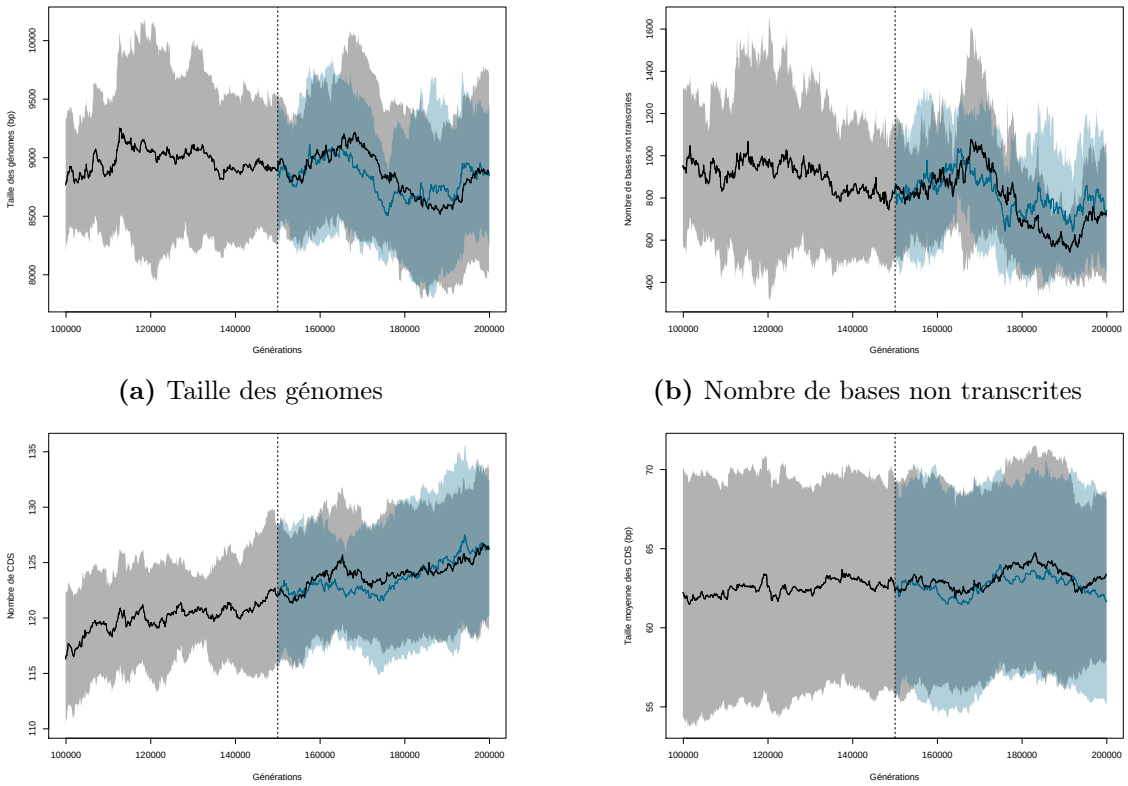
(a) Taille des génomes

(b) Nombre de bases non transcrites

(c) Nombre de CDS

(d) Taille moyenne des CDS

Figure B.2 – Évolution de certaines caractéristiques génomiques le long de la lignée du meilleur individu à la génération 200 000 pour les simulations de contrôle et celles du scénario de diminution de la pression de sélection
Les lignes représentent les moyennes pour les 10 simulations de chaque scénario et la plage colorée à l'écart-type sur les 10 simulations.
Les simulations de contrôles sont en noir et celles du scénario de diminution de la pression de sélection en bleu.



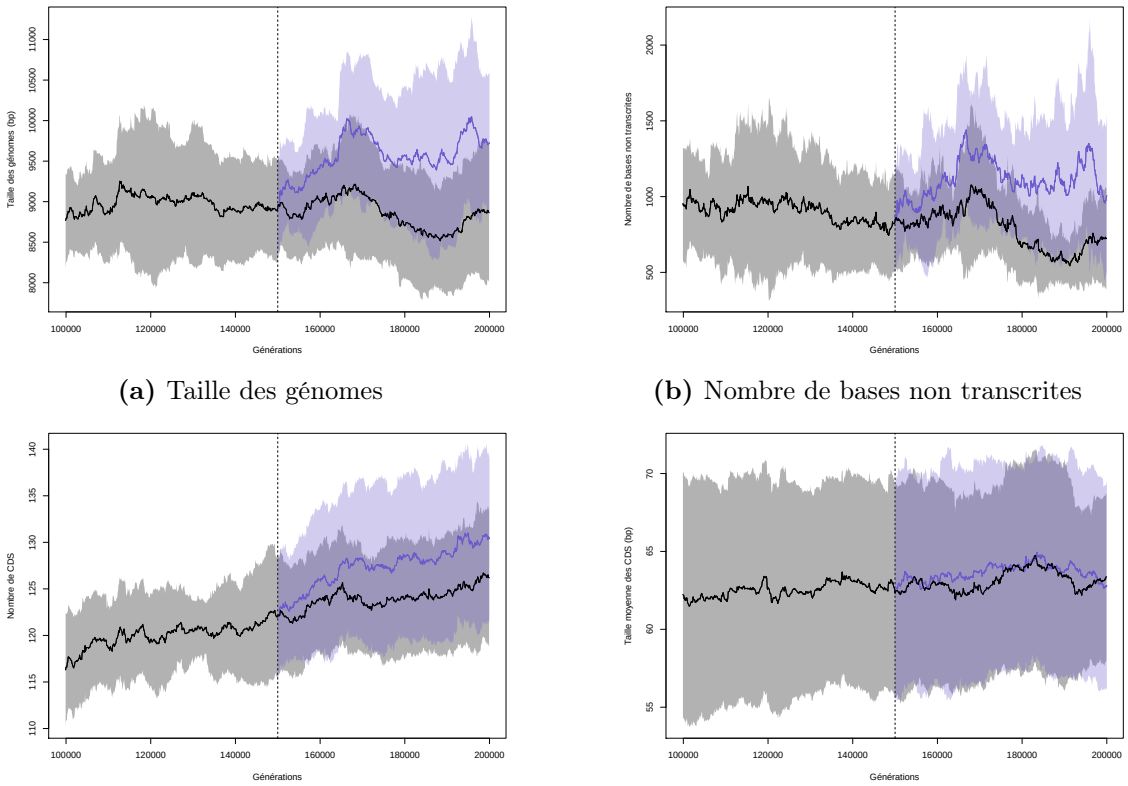
(a) Taille des génomes

(b) Nombre de bases non transcrites

(c) Nombre de CDS

(d) Taille moyenne des CDS

Figure B.3 – Évolution de certaines caractéristiques génomiques le long de la lignée du meilleur individu à la génération 200 000 pour les simulations de contrôle et celles du scénario de suppression de la recombinaison
Les lignes représentent les moyennes pour les 10 simulations de chaque scénario et la plage colorée à l'écart-type sur les 10 simulations.
Les simulations de contrôles sont en noir et celles du scénario de suppression de la recombinaison en bleu.



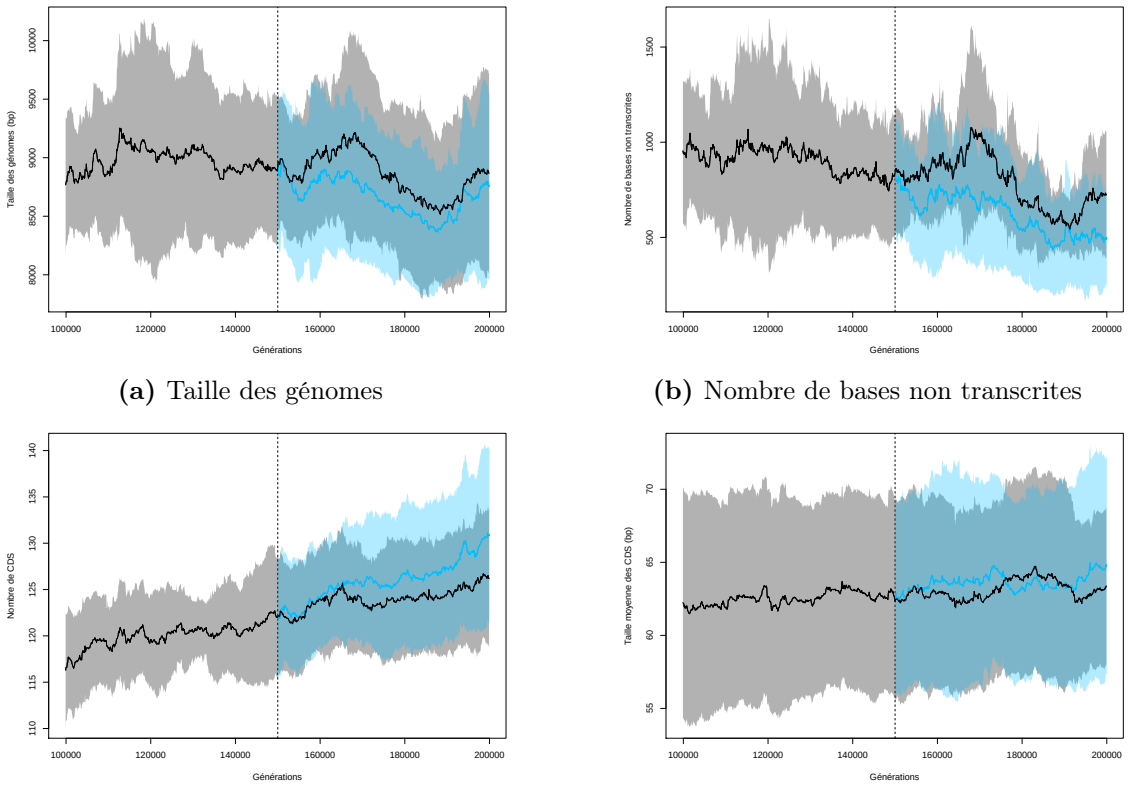
(a) Taille des génomes

(b) Nombre de bases non transcrites

(c) Nombre de CDS

(d) Taille moyenne des CDS

Figure B.4 – Évolution de certaines caractéristiques génomiques le long de la lignée du meilleur individu à la génération 200 000 pour les simulations de contrôle et celles du scénario d'augmentation de la pression de sélection
Les lignes représentent les moyennes pour les 10 simulations de chaque scénario et la plage colorée à l'écart-type sur les 10 simulations.
Les simulations de contrôles sont en noir et celles du scénario d'augmentation de la pression de sélection en bleu.



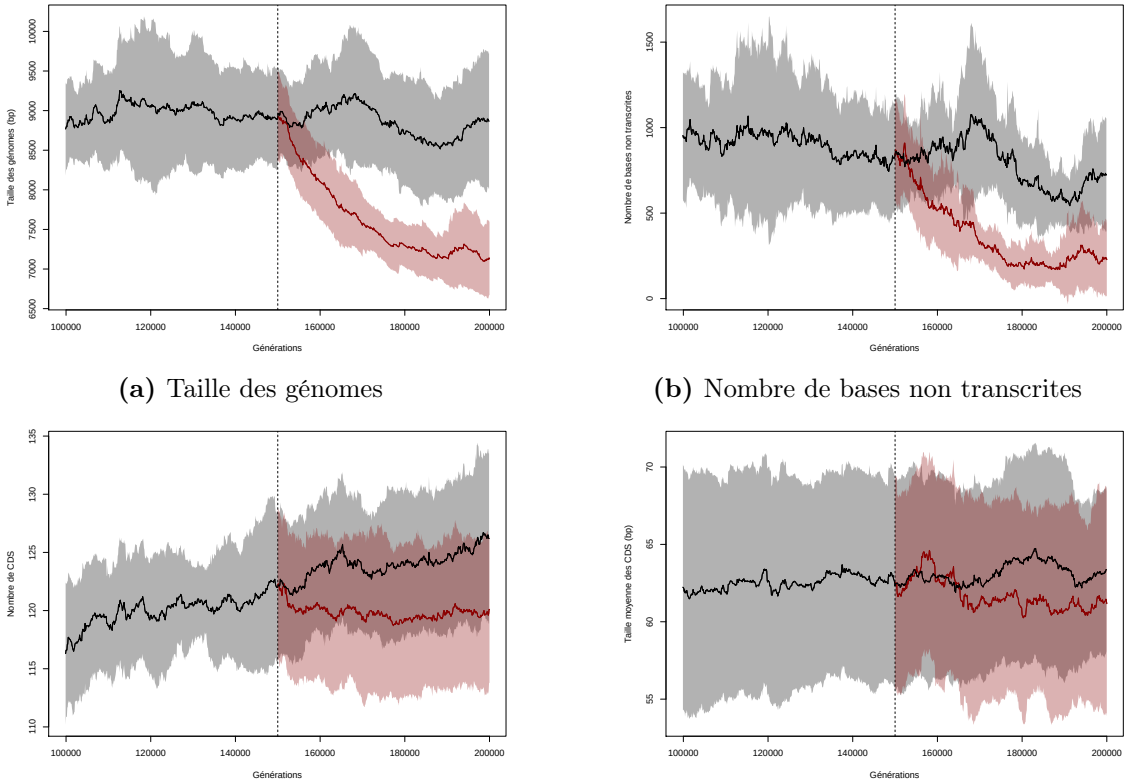
(a) Taille des génomes

(b) Nombre de bases non transcrites

(c) Nombre de CDS

(d) Taille moyenne des CDS

Figure B.5 – Évolution de certaines caractéristiques génomiques le long de la lignée du meilleur individu à la génération 200 000 pour les simulations de contrôle et celles du scénario d'augmentation de la taille de population
Les lignes représentent les moyennes pour les 10 simulations de chaque scénario et la plage colorée à l'écart-type sur les 10 simulations.
Les simulations de contrôles sont en noir et celles du scénario d'augmentation de la taille de population en bleu.



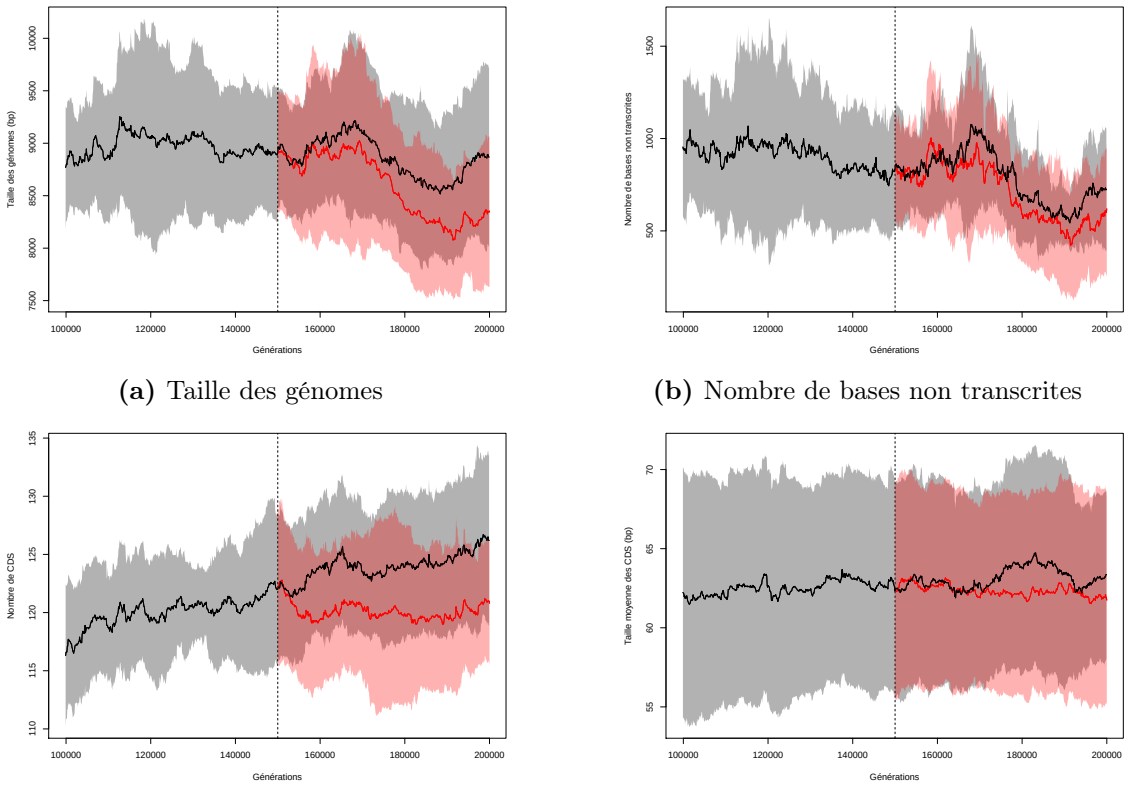
(a) Taille des génomes

(b) Nombre de bases non transcrites

(c) Nombre de CDS

(d) Taille moyenne des CDS

Figure B.6 – Évolution de certaines caractéristiques génomiques le long de la lignée du meilleur individu à la génération 200 000 pour les simulations de contrôle et celles du scénario d'augmentation des taux de mutation
Les lignes représentent les moyennes pour les 10 simulations de chaque scénario et la plage colorée à l'écart-type sur les 10 simulations.
Les simulations de contrôles sont en noir et celles du scénario d'augmentation des taux de mutation en rouge.



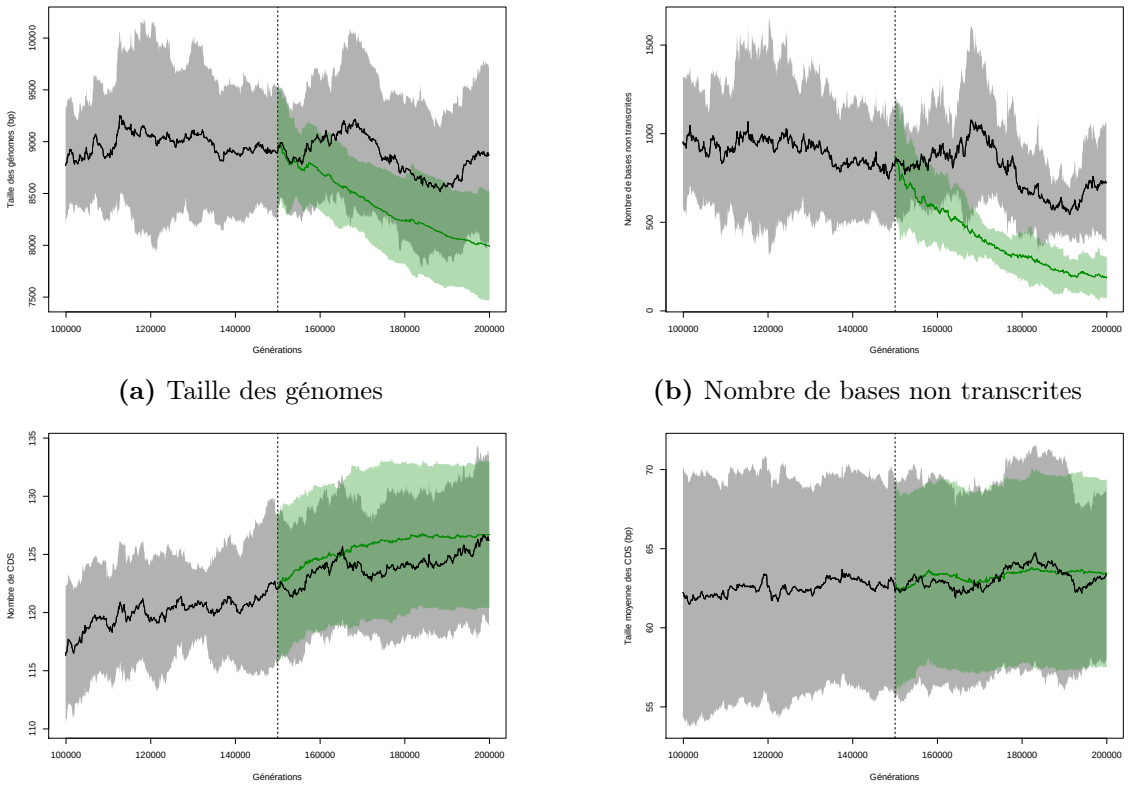
(a) Taille des génomes

(b) Nombre de bases non transcrites

(c) Nombre de CDS

(d) Taille moyenne des CDS

Figure B.7 – Évolution de certaines caractéristiques génomiques le long de la lignée du meilleur individu à la génération 200 000 pour les simulations de celles contrôle et du scénario d'augmentation des taux de réarrangement
Les lignes représentent les moyennes pour les 10 simulations de chaque scénario et la plage colorée à l'écart-type sur les 10 simulations.
Les simulations de contrôles sont en noir et celles du scénario d'augmentation des taux de réarrangement en rouge.



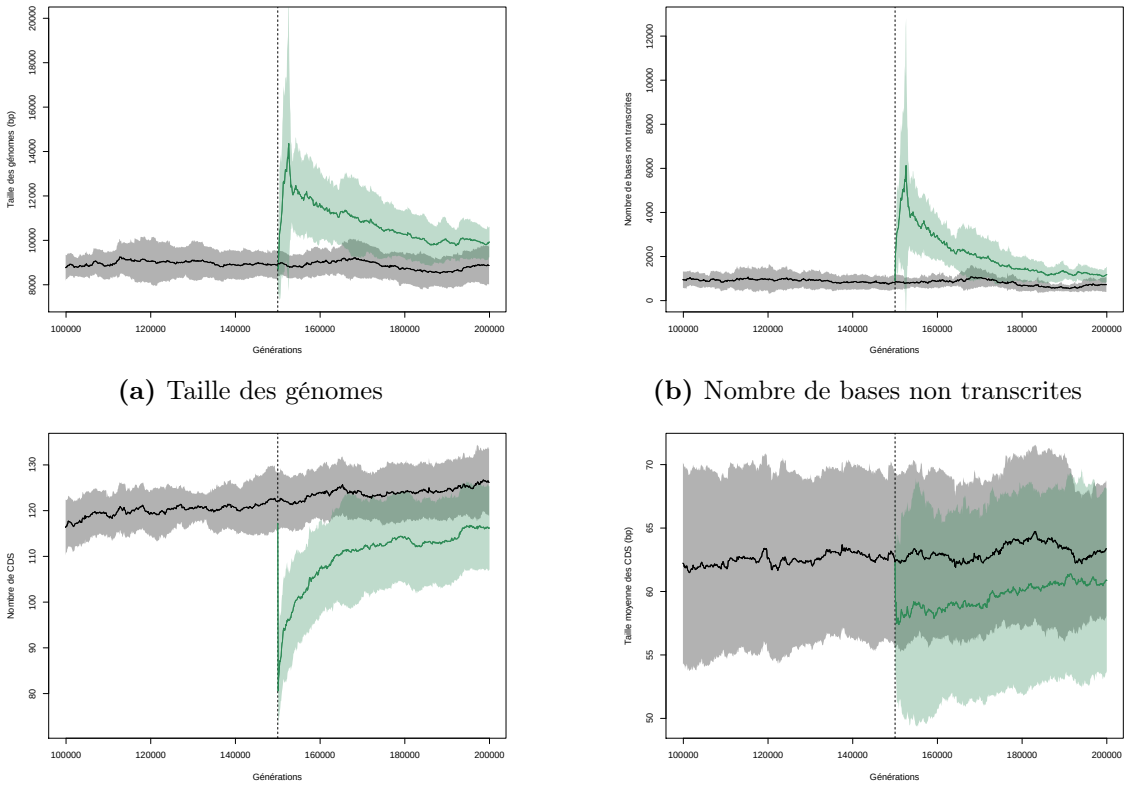
(a) Taille des génomes

(b) Nombre de bases non transcrites

(c) Nombre de CDS

(d) Taille moyenne des CDS

Figure B.8 – Évolution de certaines caractéristiques génomiques le long de la lignée du meilleur individu à la génération 200 000 pour les simulations de contrôle et celles du scénario de stabilisation de l'environnement
Les lignes représentent les moyennes pour les 10 simulations de chaque scénario et la plage colorée à l'écart-type sur les 10 simulations.
Les simulations de contrôles sont en noir et celles du scénario de stabilisation de l'environnement en vert.



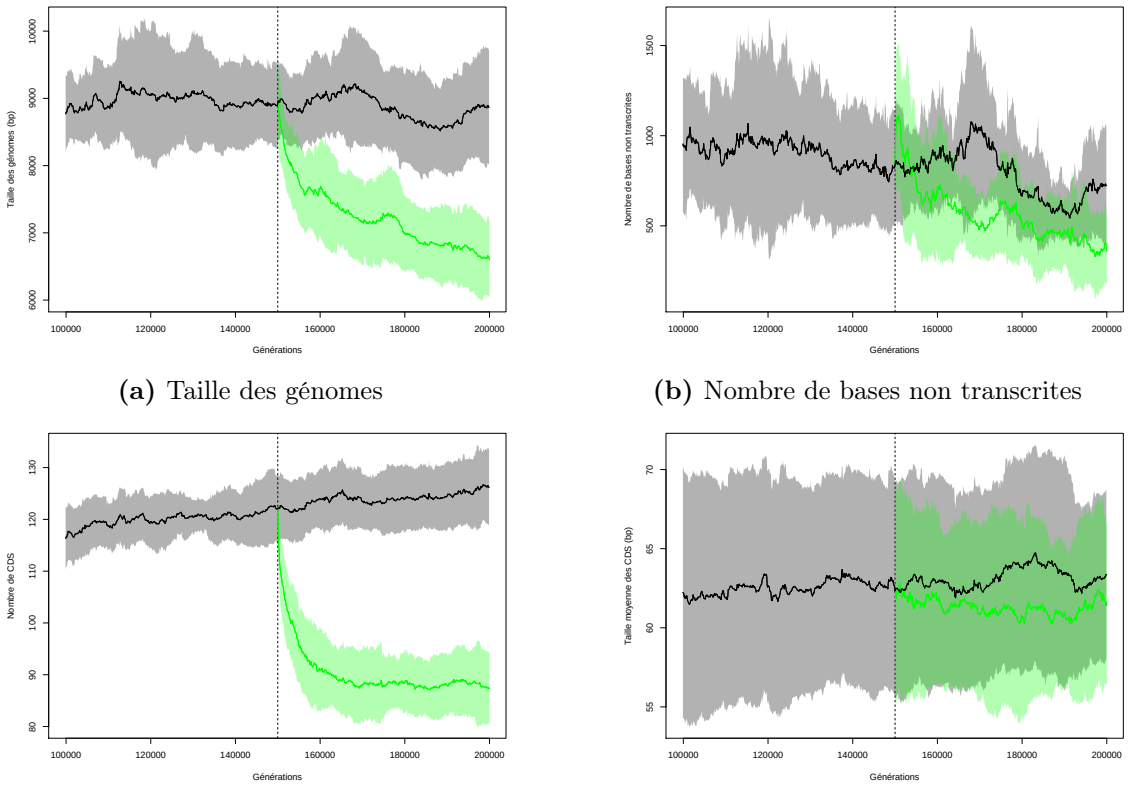
(a) Taille des génomes

(b) Nombre de bases non transcrites

(c) Nombre de CDS

(d) Taille moyenne des CDS

Figure B.9 – Évolution de certaines caractéristiques génomiques le long de la lignée du meilleur individu à la génération 200 000 pour les simulations de contrôle et celles du scénario de déplacement d'un lobe de l'environnement
Les lignes représentent les moyennes pour les 10 simulations de chaque scénario et la plage colorée à l'écart-type sur les 10 simulations.
Les simulations de contrôles sont en noir et celles du scénario de déplacement d'un lobe de l'environnement en vert.



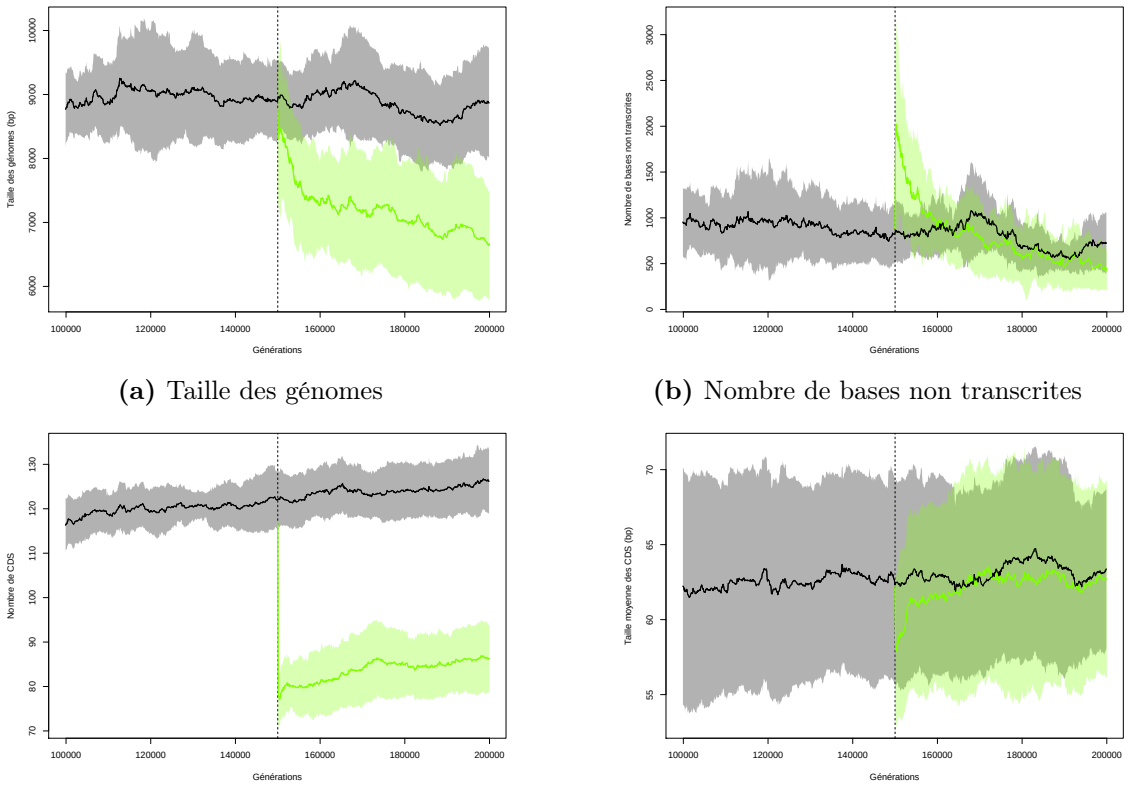
(a) Taille des génomes

(b) Nombre de bases non transcrites

(c) Nombre de CDS

(d) Taille moyenne des CDS

Figure B.10 – Évolution de certaines caractéristiques génomiques le long de la lignée du meilleur individu à la génération 200 000 pour les simulations de contrôle et celles du scénario de neutralisation d'un lobe de l'environnement
Les lignes représentent les moyennes pour les 10 simulations de chaque scénario et la plage colorée à l'écart-type sur les 10 simulations.
Les simulations de contrôles sont en noir et celles du scénario de neutralisation d'un lobe de l'environnement en vert.



(a) Taille des génomes

(b) Nombre de bases non transcrites

(c) Nombre de CDS

(d) Taille moyenne des CDS

Figure B.11 – Évolution de certaines caractéristiques génomiques le long de la lignée du meilleur individu à la génération 200 000 pour les simulations de contrôle et celles du scénario de suppression d'un lobe de l'environnement
Les lignes représentent les moyennes pour les 10 simulations de chaque scénario et la plage colorée à l'écart-type sur les 10 simulations.
Les simulations de contrôles sont en noir et celles du scénario de suppression d'un lobe de l'environnement en vert.

Annexe C

Récupération et traitements initiaux des séquences

Pour les analyses des caractéristiques génomiques, il est nécessaire de disposer des séquences des souches d'intérêt, en particulier les génomes complets, les séquences codantes pour des protéines et les séquences de gènes ribosomaux. De plus, afin de pouvoir comparer les souches ayant subi une évolution réductive aux souches ayant conservé un génome "long", il faut trouver des séquences comparables entre ces souches, c'est-à-dire des séquences ayant une histoire évolutive commune au sein des différentes souches et descendants d'un même gène ancestral. Ces séquences sont appelées séquences homologues. Le concept d'homologie regroupe deux notions (Figure C.1) : l'orthologie et la paralogie (Figure C.1). Deux gènes homologues sont orthologues s'ils ont acquis leur indépendance évolutive après un évènement de spéciation tandis qu'ils sont paralogues s'ils ont acquis leur indépendance évolutive après un évènement de duplication (Fitch, 1970).

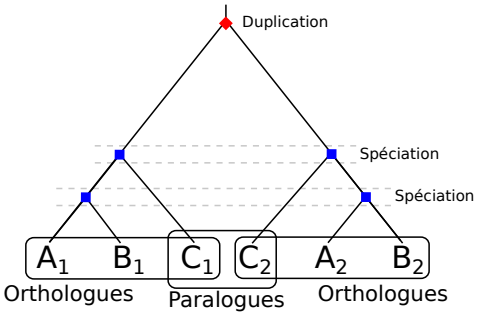


Figure C.1 – Concept d'homologie avec la notion d'orthologie et de paralogie
Un gène a subi une duplication chez l'ancêtre commun à trois espèces actuelles (A, B et C).
Les gènes A_1, B_1, C_1 sont orthologues entre eux tout comme A_2, B_2, C_2 : les nœuds les plus récents qui lient les gènes entre eux sont des nœuds de spéciation. Les gènes C_1 et C_2 sont paralogues, tout comme A_1 et A_2 ou B_1 et B_2 : les nœuds les plus récents qui lient les gènes entre eux sont des nœuds de duplication.

Afin de replacer les motifs observés dans un contexte phylogénétique, fonctionnel ou encore de structure génomique, un certain nombre de traitements doivent être effectués. Cette section présente les méthodes choisies pour la récupération et le prétraitement des données de séquences en vue de l'analyse de l'évolution réductive chez *Prochlorococucs*.

C.1 Récupération des séquences d'intérêt

Plusieurs bases de données regroupent les séquences de bactéries (les plus connues sont *GenBank*, *Ensembl* et *RefSeq*). La plupart de ces bases de données permettent de récupérer les données brutes sans traitement préalable. Or, dans notre cas, nous nous intéressons particulièrement aux séquences homologues. *Hogenom* 6 (Penel *et al.*, 2009) est une base de données rassemblant les gènes des génomes actuellement disponibles en familles au sein desquelles les séquences sont homologues. Utiliser cette base plutôt qu'une autre permet de limiter les traitements à effectuer dans la recherche de séquences comparables entre plusieurs souches. Afin d'unifier toutes les analyses, toutes les séquences nécessaires sont extraites de cette base, même lorsqu'elles n'ont pas vocation à être rassemblées en familles, comme les génomes complets. En Mars 2014, 12 génomes complets de *Prochlorococcus* sont disponibles dans cette base de données, ainsi que 6 souches de *Synechococcus* marines (Tableau C.1).

C.1.1 Génomes complets, séquences des gènes ribosomaux et séquences des gènes codant pour des protéines

Toutes les séquences ont été extraites d'*Hogenom* 6 (Penel *et al.*, 2009) à l'aide de scripts *Python* et *BioPython* (Cock *et al.*, 2009) interfacés avec *ACNUC* (Gouy et Delmotte, 2008), permettant l'interrogation de la base *Hogenom* 6 (Penel *et al.*, 2009).

Les séquences récupérées sont les génomes complets, les gènes ribosomaux et les gènes codant pour des protéines (CDS) (Tableau C.2).

C.1.2 Séquences des CDS orthologues à plusieurs souches

Au sein d'*Hogenom* 6, les gènes des différentes souches sont rassemblés en familles de gènes sur la base d'une homologie des gènes au sein d'une famille. Nous nous basons sur ce fait pour récupérer des familles de gènes homologues à certaines souches étudiées en utilisant les familles définies dans la base *Hogenom* 6.

Nous nous intéressons plus particulièrement aux gènes orthologues (Figure C.1) car ils reflètent plus exactement les évènements de spéciation et l'histoire évolutive des souches

		Nom complet	Nom d'usage	Taille du génome (Mb)
Prochlorococcus	HLI	*Prochlorococcus marinus* subsp. pastoris str. CCMP1986	MED4	1.66
		Prochlorococcus marinus str. MIT 9515	MIT9515	1.7
	HLII	*Prochlorococcus marinus* str. MIT 9301	MIT9301	1.64
		Prochlorococcus marinus str. AS9601	AS9601	1.67
		Prochlorococcus marinus str. MIT 9215	MIT9215	1.74
		Prochlorococcus marinus str. MIT 9312	MIT9312	1.71
	LLI	*Prochlorococcus marinus* str. NATL1A	NATL1A	1.86
		Prochlorococcus marinus str. NATL2A	NATL2A	1.84
	LLII	*Prochlorococcus marinus* subsp. marinus str. CCMP1375	SS120	1.75
	LLIII	*Prochlorococcus marinus* str. MIT 9211	MIT9211	1.69
	LLIV	*Prochlorococcus marinus* str. MIT 9313	MIT9313	2.41
		Prochlorococcus marinus str. MIT 9303	MIT9303	2.68
Synechococcus		*Synechococcus* sp. CC9311	CC9311	2.61
		Synechococcus sp. WH 7803	WH7803	2.37
		Synechococcus sp. CC9605	CC9605	2.51
		Synechococcus sp. CC9902	CC9902	2.23
		Synechococcus sp. WH 8102	WH8102	2.43
		Synechococcus sp. RCC307	RCC307	2.22

Table C.1 – Souches utilisées et tailles de génome correspondant
Les noms complets sont les noms utilisés pour récupérer les séquences correspondant aux souches dans la base de données *Hogenom*. Pour chaque souche, un nom d'usage, utilisé dans la suite du manuscrit, est associé. Il a été choisi afin de ne pas avoir de doublons dans les noms de souches. Les tailles de génomes correspondent aux tailles des souches dans la base de données du *Hogenom*.
Les couleurs des noms des souches symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit.

étudiées. Pour chaque souche, nous souhaitons ainsi conserver des CDS pour lesquelles la famille *Hogenom* correspondante compte exactement un gène dans la souche en question pour éliminer les cas supposés de paralogie. Une première phase de tri des séquences est ainsi effectuée pour ne conserver que les CDS dits non paralogues, en éliminant les familles de gènes pour lesquelles plusieurs séquences sont présentes dans les souches (Tableau C.2). Sont alors éliminés entre 20 et 30% familles de gènes pour *Synechococcus* et non réduites de *Prochlorococcus* et entre 10 et 20% pour les souches réduites de *Prochlorococcus*.

Pour trouver des séquences ayant une histoire évolutive commune reflétant celle des souches, il faut déjà définir les souches à comparer. Or celles-ci doivent être choisies en fonction de l'analyse et de la question posée. Ici, trois groupes de souches sont définis pour *Prochlorococcus* et *Synechococcus* (Tableau C.3). Pour un groupe donné, les familles retenues sont donc les familles *Hogenom* qui possèdent exactement une CDS dans chacune des souches du groupe étudié. Nous obtenons 871, 1242 et 1274 familles pour chacun des trois groupes (2^e phase dans le tableau C.3).

		Souche	Gènes	Gènes ribosomaux	CDS	CDS non paralogues
Prochlorococcus	HLI	MED4	1 762	55	1 717	1 524
		MIT9515	1 964	55	1 905	1 611
	HLII	MIT9301	1 962	53	1 906	1 673
		AS9601	1 965	53	1 920	1 681
		MIT9215	2 054	53	1 982	1 686
		MIT9312	1 856	56	1 810	1 577
	LLI	NATL1A	2 250	53	2 193	1 805
		NATL2A	2 228	54	2 162	1 798
	LLII/LLIII	SS120	1 930	54	1 883	1 561
		MIT9211	1 900	55	1 853	1 514
	LLIV	MIT9313	2 330	57	2 997	1 960
		MIT9303	3 136	58	2 269	1 848
Synechococcus		CC9311	2 944	55	2 832	2 027
		WH7803	2 586	54	2 533	2 056
		CC9605	2 758	55	2 645	2 037
		CC9902	2 357	53	2 306	1 960
		WH8102	2 581	55	2 519	1 978
		RCC307	2 582	54	2 531	1 844

Table C.2 – Nombre de gènes, gènes ribosomaux, CDS et CDS non paralogues dans les souches étudiées
Le nombre de gènes correspond au nombre de gènes recensés dans la base de données du NCBI pour chacune des souches. Le nombre de gènes ribosomaux est le nombre de séquences récupérées dans la base de données *Hogenom* en spécifiant "ribosomal" comme mot-clé de recherche. Le nombre de CDS est le nombre de séquences issues d'une requête pour des CDS complets dans la base de données *Hogenom*. Les CDS non paralogues sont les CDS pour lesquels la famille *Hogenom* correspondante est représentée une seule fois dans la souche étudiée.
Les couleurs des noms des souches symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit.

Pour récupérer ces séquences orthologues, seule l'appartenance à des familles prédéfinies compte. Les arbres phylogénétiques sous-jacents des familles de gènes ne sont pas utilisés et la recherche ne se fait pas selon un motif phylogénétique prédéfini. Ainsi, avec notre méthode, un grand nombre de séquences peuvent être récupérées, y compris des cas de transferts de gènes entre les souches, mais aussi des transferts issus d'autres espèces que les souches étudiées, qui peuvent alors ajouter du bruit dans les analyses. Ces cas peuvent être éliminés en considérant l'arbre phylogénétique complet soutenu par chaque famille de séquence (arbre dit "de gène" construit avec tous les exemplaires du gène présent dans *Hogenom*) en ne conservant que les familles qui donnent un arbre où les souches étudiées sont issues d'un même ancêtre commun dont tous les descendants sont exactement les souches étudiées. Cette phase de tri (la 3^e) se fait en prenant en compte les arbres complets des familles *Hogenom*. Elle est effectuée à l'aide de scripts *BioPython* (Cock *et al.*, 2009) et *Bio.Phylo* (Talevich *et al.*, 2012). Cette phase réduit le nombre de familles retenues de 15% (Tableau C.3).

Groupes	Souches concernées	2^e phase	3^e phase
Toutes les souches	MED4, MIT9515, MIT9301, AS9061, MIT9215, MIT9312, NATL1A, NATL2A, SS120, MIT9211, MIT9313, MIT9303, CC9311, WH7803, CC9605, CC9902, WH8102, RCC307	871	693
Souches *Synechococcus*	CC9311, WH7803, CC9605, CC9902, WH8102, RCC307	1 274	1 085
Souches HL	MED4, MIT9515, MIT9301, AS9061, MIT9215, MIT9312	1 242	1 133

Table C.3 – Groupes de souches utilisées pour la recherche de familles de gènes orthologues et nombre de famille de gènes trouvés dans les seconde et troisième phases de tri des familles de gènes La conservation des familles de gènes trouvées exactement une fois dans chacune des souches du groupe étudié est la deuxième phase de tri. Dans la troisième phase, ne sont conservées que les familles dont toutes les souches sont monophylétiques au sein des arbres de familles de gènes complets d'*Hogenom* 6.
Les couleurs des noms des souches symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit.

Cette méthode, volontairement stringente, élimine un grand nombre de familles de gènes et potentiellement des familles de gènes dont l'évolution pourrait apporter des pistes sur l'évolution réductive. Cependant, elle permet de ne retenir que des familles pour lesquelles toute observation de changements reflète un évènement évolutif fort.

C.2 Alignement des séquences

Étant donné le nombre de souches à comparer, les alignements multiples et non par paires de souches sont plus fiables. Un grand nombre d'outils d'alignement multiple sont disponibles tels *ClustalW* (Thompson *et al.*, 1994), *Muscle* (Edgar, 2004), *MAFFT* (Katoh *et al.*, 2005), *Prank* (Löytynoja et Goldman, 2005). Le choix de l'un ou l'autre des outils dépend de l'objectif à atteindre par l'alignement. En effet, les alignements de séquences peuvent être utilisés dans différents buts (Morrison, 2009). Les analyses phylogénétiques étant l'objectif principal de nos analyses, nous avons besoin d'un outil qui prenne en compte au mieux les évolutions de séquences dans l'alignement. C'est le cas de *Prank* (Löytynoja et Goldman, 2005) qui, lors de la construction de l'alignement, distingue les insertions et les délétions afin d'éviter la pénalisation répétée des insertions (contrairement aux autres outils d'alignement).

Prank permet un alignement fiable basé sur les codons. Or, comme nous nous concentrons sur les familles de gènes homologues qui sont des séquences codant pour des protéines, conserver la structure en codons pour l'alignement parait important. En effet, cela permet d'éviter le mauvais alignement de codons non homologues et donc de surestimer le nombre

de substitutions non synonymes et sous-estimer le nombre de substitutions synonymes, utiles pour la détection de sélection positive. Ainsi, *Prank* est considéré comme un des meilleurs outils dans la détection de sélection positive en diminuant le nombre de faux positif inférés (Jordan et Goldman, 2012; Fletcher et Yang, 2010).

Les séquences des familles de gènes orthologues conservées sont donc alignées à l'aide de *Prank* (Löytynoja et Goldman, 2005) avec un alignement multiple par codons.

C.3 Construction de concaténats

Dans notre jeu de données, l'alignement d'une famille de gènes donne un alignement d'au plus 4832 bases. Ceci est relativement court à l'échelle des génomes (entre 1.64 Mb pour *Prochlorococcus* MIT9301 et 2.68 Mb pour *Prochlorococcus* MIT9303). Pour certaines analyses, le signal sur des séquences si courtes pourrait ressembler davantage à du bruit qu'à un signal généralisable à l'échelle des génomes et de l'évolution réductive.

Pour contrer cet aspect et renforcer le signal contenu dans les séquences, les familles de gènes orthologues sont concaténées par génome dans un ordre aléatoire (identique pour chaque génome) avec les séquences alignées avec *Prank*. Nous obtenons ainsi au sein de chaque groupe, une séquence relativement longue pour chacune des souches (entre 83 844 et 1 131 243 bases).

Dans les alignements, certaines positions sont faiblement alignées. Ces régions peuvent refléter des régions divergentes, issues par exemple de transferts intragéniques. Si elles sont issues de génomes extérieurs, elles peuvent fausser certaines des analyses, en particulier en changeant localement le pourcentage de bases GC ou l'usage des codons. Il conviendrait alors de les éliminer pour les analyses nécessitant des régions fortement homologues. Pour ces analyses, un deuxième concaténat est construit sur les familles de gènes préalablement filtrées à l'aide de *GBlock* (Castresana, 2000), ce qui permettrait d'augmenter le signal phylogénétique (Talavera et Castresana, 2007).

C.4 Construction des arbres phylogénétiques

La réduction des génomes au sein du genre *Prochlorococcus* concerne plusieurs espèces[1]. Il est ainsi utile de mettre les changements génomiques observés dans un contexte phylogénétique.

Pour *Prochlorococcus* et *Synechococcus*, un certain nombre d'arbres phylogénétiques ont déjà été construits (Kettler *et al.*, 2007; Luo *et al.*, 2008, 2011; Sun et Blanchard, 2014).

[1]Les différentes souches de *Prochlorococcus* sont en fait tellement divergentes qu'elles devraient être considérées comme des espèces différentes (Thompson *et al.*, 2013)

Groupes		JC69 $k=1$	K80 $k=2$	F81 $k=3$	F84 $k=6$	HKY85 $k=5$	TN93 $k=6$	GTR $k=9$
Toutes les souches	L	-6.58×10^6	-6.51×10^6	-6.58×10^6	-6.51×10^6	-6.51×10^6	-6.50×10^6	-6.49×10^6
	AIC	6.58×10^6	6.51×10^6	6.58×10^6	6.51×10^6	6.51×10^6	6.50×10^6	$\mathbf{6.49\times10^6}$
Synechococcus	L	-4.85×10^6	-4.82×10^6	-4.8×10^6	-4.76×10^6	-4.76×10^6	-4.76×10^6	-4.75×10^6
	AIC	4.85×10^6	4.82×10^6	4.8×10^6	4.76×10^6	4.76×10^6	4.76×10^6	$\mathbf{4.75\times10^6}$
Souches HL	L	-3.40×10^6	-3.37×10^6	-3.31×10^6	-3.26×10^6	-3.26×10^6	-3.26×10^6	-3.25×10^6
	AIC	-3.40×10^6	-3.37×10^6	-3.31×10^6	-3.26×10^6	-3.26×10^6	-3.26×10^6	$\mathbf{-3.25\times10^6}$

Table C.4 – Vraisemblance et critère d'Akaiké (AIC) pour les différents groupes des différents modèles d'évolution de séquences utilisés pour la construction des arbres
k : nombre de paramètres du modèle, L : log-vraisemblance du modèle, AIC : critère d'Akaiké
JC69 (Jukes et Cantor, 1969), K80 (Kimura, 1980), F81 (Felsenstein, 1981), F84 (Felsenstein, 1993), HKY85 (Hasegawa *et al.*, 1985), TN93 (Tamura et Nei, 1993), GTR (Lanave *et al.*, 1984; Tavaré, 1986)

Cependant, la plupart de ces arbres ne prennent pas en compte toutes les séquences de notre étude (Kettler *et al.*, 2007; Luo *et al.*, 2008) ou alors ne sont pas basés sur le jeu de séquences d'intérêt (Luo *et al.*, 2011; Sun et Blanchard, 2014). Nous souhaitons ainsi construire un arbre phylogénétique pour chacun des groupes de souches étudiées en utilisant les concaténats de familles de gènes filtrés.

Différentes méthodes et outils permettent la reconstruction d'arbres phylogénétiques. Les méthodes basées sur le maximum de vraisemblance sont parmi les plus fiables. Elles recherchent dans l'espace des arbres un arbre qui maximise la vraisemblance, c'est-à-dire la probabilité d'observer les données étant donné l'arbre et les paramètres. De nombreux outils de reconstruction sont basés sur cette méthode : *PHYLIP* (Felsenstein, 2002), *MOLPHY* (Adachi et Hasegawa, 1996), *PAUP* (Swofford, 2003), *PhyML* (Guindon et Gascuel, 2003), *RAxML* (Stamatakis, 2006), *GARLI* (Zwickl, 2006), etc. Dans cette méthode, les suppositions du modèle d'évolution sont explicites et chacune peut être testée afin de choisir les paramètres qui maximisent la vraisemblance. Cette étape est importante car si le modèle est mal spécifié, l'arbre résultant a peu de propriétés statistiques (Yang et Rannala, 2012). Elle est cependant coûteuse en calcul.

Pour la construction des arbres, nous avons choisi d'utiliser *PhyML* (Guindon et Gascuel, 2003). L'algorithme sous-jacent part d'un arbre initial construit avec une méthode rapide basée sur les distances. Ensuite, il modifie l'arbre avec un algorithme d'optimisation qui ajuste simultanément la topologie de l'arbre et les longueurs de branches afin d'améliorer la vraisemblance (Guindon et Gascuel, 2003). Plusieurs modèles d'évolution des séquences nucléotidiques sont implémentés (JC69 (Jukes et Cantor, 1969), K80 (Kimura, 1980), F81 (Felsenstein, 1981), F84 (Felsenstein, 1993), HKY85 (Hasegawa *et al.*, 1985), TN93 (Tamura et Nei, 1993), GTR (Lanave *et al.*, 1984; Tavaré, 1986)). Nous avons choisi

cet outil pour sa rapidité, sa simplicité d'utilisation et la capacité de tester facilement plusieurs modèles d'évolution de séquences.

Comme nous ne souhaitons pas faire d'hypothèses *a priori* sur le modèle d'évolution nucléotidique à choisir, ils sont tous testés pour chacun des groupes. Le meilleur modèle doit combiner la vraisemblance maximale et le minimum de paramètres différents pour atteindre cette vraisemblance. Pour cela, le critère d'Akaiké ($AIC = 2k - 2L$) est utilisé pour discriminer les modèles (avec L la log-vraissemblance du modèle et k le nombre de paramètres) : le modèle conservé sera celui avec le critère le plus faible (Table C.4). Dans tous les groupes, le modèle GTR ou *General time reversible* (Lanave *et al.*, 1984; Tavaré, 1986) est systématiquement celui conservé. Ce modèle est le plus général, c'est-à-dire avec le moins d'*a priori* sur les paramètres, mais c'est aussi celui avec le plus de paramètres. Les arbres obtenus pour les différents groupes avec le modèle GTR sont présentés dans la figure C.2.

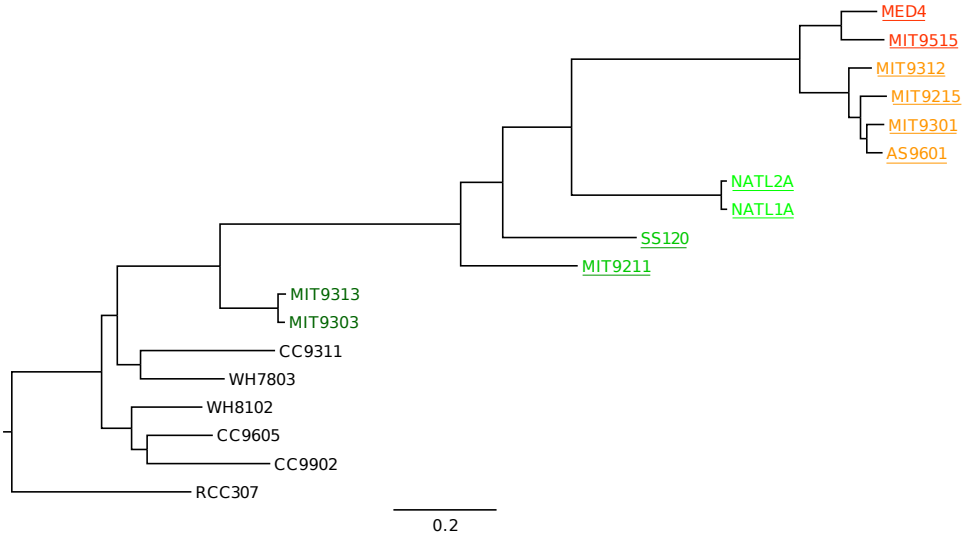
C.5 Catégorisation des familles de gènes

Hogenom (Penel *et al.*, 2009) fournit des annotations pour les gènes au sein des familles de gènes. Dans le cas des familles de gènes orthologues, il y a autant d'annotations pour une famille que de souches dans le groupe étudié. De plus, ces annotations sont souvent difficiles à interpréter. Il est donc difficile de faire ressortir des tendances d'annotations pour les familles de gènes en se basant seulement sur les annotations fournies par *Hogenom*.

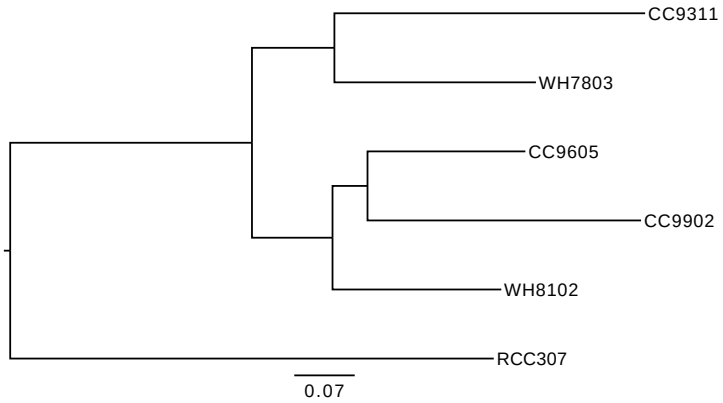
Les catégories COG (Tatusov *et al.*, 2000) permettent de pallier ce problème. L'activité cellulaire des organismes microbiens est ainsi décomposée en 25 catégories représentant chacune une gamme de fonctions (Tableau C.5). Actuellement cette base couvre 66 génomes unicellulaires avec 185 505 protéines prédites (Tatusov *et al.*, 2003).

Tous les génomes étudiés ici ne font pas partie des 66 génomes présents dans la base COG. De plus, nous souhaitons avoir l'information de catégorie COG pour chaque famille de gènes et non pour chaque gène au sein des génomes. Les séquences pour lesquelles nous cherchons les catégories COG correspondent à des séquences orthologues entre plusieurs souches, pour lesquelles les fonctions sont supposées proches au sein des différentes souches. Cette hypothèse permet d'assigner des catégories COG aux familles pour lesquelles au moins une catégorie est trouvée pour un des gènes constituant cette famille.

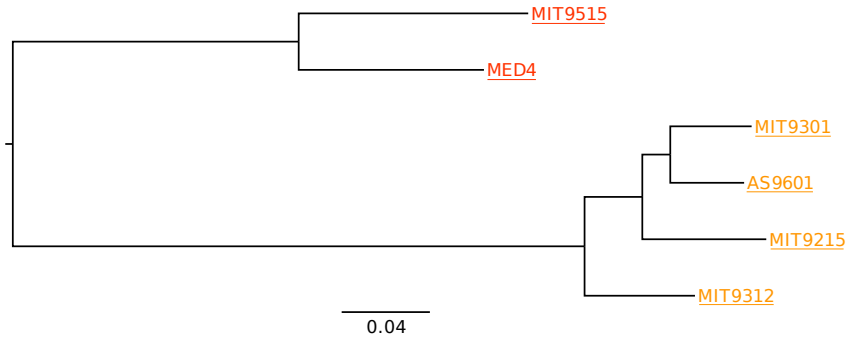
Les catégories COG sont disponibles pour un certain nombre de gènes de *Prochlorococcus* MED4 (Wang *et al.*, 2014). En extrapolant les catégories des gènes de *Prochlorococcus* MED4 aux familles de gènes présentes chez *Prochlorococcus* et *Synechococcus*, 991 familles sur les 3778 présentes dans au moins une souche peuvent être catégorisées. Les 2787 familles restantes ne peuvent être assignées directement à des catégories. Elles sont annotées à partir des données brutes des catégories COG et des familles *Hogenom*. Ainsi, les séquences d'une famille *Hogenom* sont comparées à toutes les séquences disponibles



(a) Toutes les souches de *Prochlorococcus* et de *Synechococcus*



(b) Souches de *Synechococcus*



(c) Souches de *Prochlorococcus* HL

Figure C.2 – Arbres phylogénétiques construits à l'aide de *PhyML* (Guindon et Gascuel, 2003) sur les concaténats de familles de gènes alignées et filtrées pour éliminer les régions non conservées
Les couleurs des noms des souches symbolisent les différentes clades avec en noir les souches de *Synechococcus*, en vert foncé les souches de *Prochlorococcus* LLIV, en vert les souches de *Prochlorococcus* LLII/LLIII, en vert clair les souches de *Prochlorococcus* LLI, en orange les souches de *Prochlorococcus* HLII et en rouge les souches de *Prochlorococcus* HLI. Les souches dont le nom est souligné sont les souches pour lesquelles le génome est réduit.

Catégories			*Prochlorococcus* et *Synechococcus*		
			Toutes les souches	*Synechococcus*	Souches HL
Information	Traduction, structure ribosomale et biogénèse	J	103	132	121
	Traitement et modification ARN	A	0	0	0
	Transcription	K	15	20	22
	Replication, recombinaison et réparation	L	32	48	55
	Structure et dynamique chromatinienne	B	0	1	0
	Total		150(26%)	192(24%)	195(23%)
Proc. cellulaires/Signalisation	Contrôle du cycle cellulaire, division cellulaire et partionnement chromosomique	D	8	10	10
	Structure nucléaire	Y	0	0	0
	Mécanismes de défence	V	2	3	5
	Mécanismes de transduction du signal	T	11	13	15
	Biogénèse cellulaire de membrane/enveloppe	M	30	43	44
	Mobilité cellulaire	N	0	0	0
	Cytosquelette	Z	0	0	0
	Structure extracellulaire	W	0	0	0
	Trafic, sécrétion et transport vésiculaire intracellulaire	U	9	12	12
	Modification post-traductionnelle, turnover protéiques et chaperonnes	O	27	37	39
	Total		86(15%)	116(14%)	125(15%)
Métabolisme	Production et conversion d'énergie	C	43	59	52
	Transport et métabolisme carbohydrate	G	27	47	40
	Transport et métabolisme acide aminé	E	67	79	89
	Transport et métabolisme nucléotidique	F	22	32	37
	Transport et métabolisme des coenzymes	H	66	89	84
	Transport et métabolisme des lipides	I	16	21	25
	Transport et métabolisme des ions inorganiques	P	24	33	34
	Biosynthèse, transport et catabolisme des métabolites secondaires	Q	7	11	11
	Total		269(46%)	369(45%)	371(45%)
	Prédiction de fonctions générales seulement	R	50	83	94
	Fonction inconnue	S	29	55	48
	Total		79(13%)	138(17%)	142(17%)
Total			584(84%) /693	814(75%) 1 085	832(75%) /1 113

Table C.5 – Catégories COG et nombre de familles de gènes dans ces catégories pour les différents groupes de souches (valeurs arrondies à l'unité)
Les catégories COG correspondent aux catégories de la classification de Tatusov *et al.* (2000).

dans la base COG. Lors d'une correspondance exacte entre une séquence d'une famille *Hogenom* et une séquence d'une catégorie COG, la catégorie COG est attribuée à la famille *Hogenom*. Quand une famille *Hogenom* correspond à plusieurs catégories COG, chacune des catégories COG trouvées est pondérée de telle façon que le poids soit proportionnel au nombre de fois où une catégorie est trouvée et que la somme des poids soit égale à 1.

Les différentes catégories obtenues pour les familles de gènes correspondant aux groupes de souches étudiés sont résumées dans le Tableau C.5. Environ 80% des familles de gènes

orthologues au sein des souches de *Prochlorococcus* et de *Synechococcus* possèdent une ou plusieurs annotations (Tableau C.5).

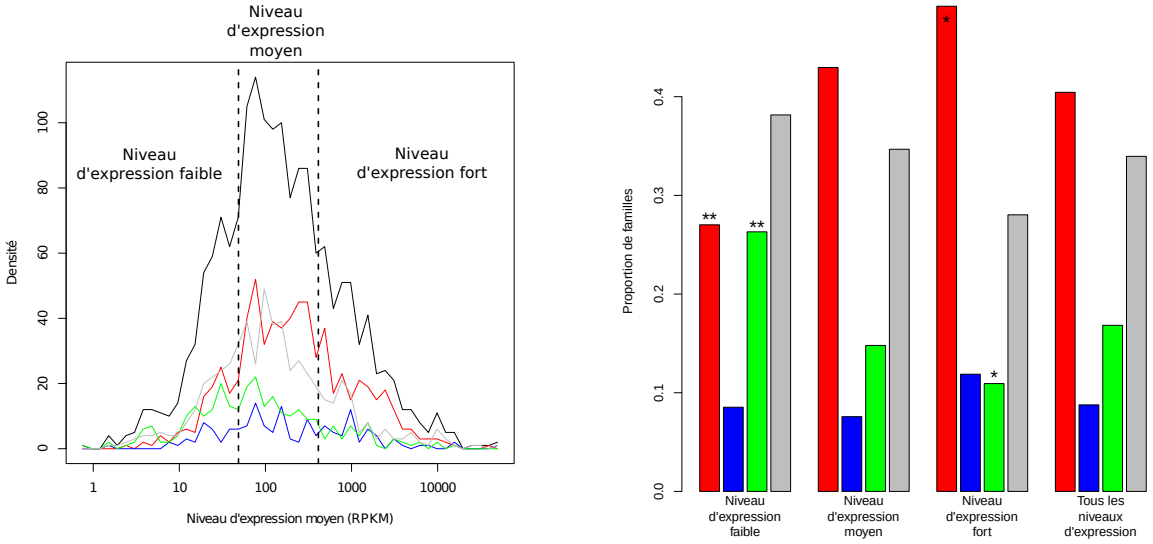
Au sein des familles annotées, les proportions dans les différentes catégories COG sont relativement bien conservées entre les différents groupes de *Prochlorococcus* et de *Synechococcus*. 25% des familles catégorisées le sont pour des fonctions liées au stockage et au traitement de l'information, 15% pour des fonctions liées aux processus cellulaires et de signalisation, 45% au métabolisme et 15% à des fonctions peu caractérisées (Tableau C.5). Cette répartition de catégories reflète-t-elle la réalité des processus cellulaires chez *Prochlorococcus* et *Synechococcus* ? Il faut être prudent. En effet, les familles étudiées correspondent à des familles conservées au cours de l'évolution et ne prennent donc pas en compte les spécificités de chacunes de souches.

Il est important de noter que la dernière mise à jour des catégories COG et surtout les gènes utilisés comme référence datent de 2003 (Tatusov *et al.*, 2003). Depuis, un grand nombre de génomes ont été séquencés et annotés. L'annotation de certains des gènes utilisés pour les catégories a pu être améliorée. Il faut ainsi être prudent avec les catégories trouvées dans notre cas. En outre, pour une famille *Hogenom*, plusieurs catégories COG peuvent être trouvées dont certaines liées potentiellement à des fonctions très différentes les unes des autres. Ceci est en partie dû à la recherche de correspondance entre n'importe quelle séquence d'une famille de gènes et n'importe quelle séquence d'une catégorie COG. Les séquences *Hogenom* utilisées peuvent être différentes des séquences d'intérêts, en particulier en cas de transfert au sein de la famille *Hogenom*. Étant données ces limites, les catégories COG pour les familles de gènes qui nous intéressent doivent être utilisées avec prudence.

C.6 Données d'expression

Les protéines indispensables à la survie des organismes sont souvent codées par des gènes fortement exprimés et tout changement de séquence peut modifier les fonctions. Ainsi, les gènes fortement exprimés évoluent plus lentement (Wang *et al.*, 2014; Pál *et al.*, 2001; Brawand *et al.*, 2011; Drummond et Wilke, 2008), du fait d'une forte pression de sélection pour minimiser les coûts de mauvaise traduction ou d'un mauvais repliement. Les niveaux d'expression peuvent ainsi être utilisés comme estimateur des taux d'évolution des séquences et des degrés relatifs de sélection purificatrice contre les changements des séquences des gènes.

Chez *Prochlorococcus*, les données d'expressions des gènes sont peu disponibles, et ces données sont principalement issues de conditions de laboratoire (Tolonen *et al.*, 2006; Wang *et al.*, 2014). Dans l'étude de Tolonen *et al.* (2006), des données d'expressions en présence et en absence d'azote sont comparées mais ne sont pas exploitables directement, car les valeurs brutes ne sont pas fournies. Wang *et al.* (2014) ont mesuré l'expression des gènes de la souche *Prochlorococcus* MED4 dans 10 conditions différentes et leurs données



(a) Densité des niveaux d'expression en fonction de l'orthologie des familles de gènes

(b) Proportion des différentes types de gènes au sein des catégories de niveaux d'expressions

Figure C.3 – Caractéristiques des niveaux d'expression selon les familles de gènes chez *Prochlorococcus* MED4
Les différents types d'orthologies sont symbolisées par les couleurs : en noir, toutes les familles de gènes présentes, en rouge, les familles de gènes présentes chez toutes les espèces de *Prochlorococcus* et *Synechococcus* (693 familles) ; en bleu, les familles de gènes spécifiques aux souches réduites de *Prochlorococcus* ; en vert, les familles de gènes spécifiques aux souches de *Prochlorococcus* HL ; en gris, les familles de gènes spécifiques à MED4, sans orthologues dans aucune autre souche.
Pour la Figure C.3a, les lignes verticales en pointillés délimitent les catégories de niveaux d'expression. Les seuils ont été choisis pour que 25% des familles de gènes aient des faibles niveaux d'expression, 50% des niveaux d'expression moyens et 25% des niveaux des forts niveaux d'expression.
Sur la Figure C.3b, les étoiles correspondent à la significativité des tests de χ^2 de comparaison des proportions des différents types d'orthologie au sein de chacune des catégories par rapport à la proportion des différents types d'orthologie au sein de toutes les familles de gènes. Une étoile signifie une p-value comprise entre 5% et 1% et deux étoiles à une p-value inférieure à 1%.

brutes sont disponibles. Ce sont celles-ci que nous utilisons dans le manuscrit.

La gamme des niveaux d'expression des gènes de MED4 est très large (Figure C.3a). Trois catégories de niveaux d'expressions sont définies : faible, moyen et fort. Les seuils ont été choisis pour que 25% de toutes les familles de gènes aient des faibles niveaux d'expression, 50% des niveaux d'expression moyen et 25% des forts niveaux d'expression. Comme le montre la Figure C.3a, les niveaux d'expression moyens sont ceux compris entre 49 et 412 RPKM (*reads per kb per million*, unité de mesure utilisée pour les données *RNA-seq*).

Les familles de gènes ne sont pas réparties de la même façon dans les catégories de niveaux d'expression selon que les familles sont orthologues ou non (Figure C.3b). Ainsi, dans la catégorie des faibles niveaux d'expression, les familles présentes chez toutes les souches sont sous-représentées ($P = 5.029 \cdot 10^{-3}$, test de χ^2 à un degré de liberté), alors que les familles spécifiques aux souches HL sont surreprésentées ($P = 3.901 \cdot 10^{-3}$, test de χ^2 à un degré de liberté). La tendance est inversée dans la catégorie de forts niveaux d'expression :

surreprésentation des familles présentes chez toutes les souches ($P = 0.04711$, test de χ^2 à un degré de liberté) et sous-représentation des familles spécifiques aux souches HL ($P = 0.01187$, test de χ^2 à un degré de liberté). Ces observations sont cohérentes avec l'idée que les gènes les plus exprimés sont aussi les gènes les plus conservés. Cependant, aucune tendance ne se dégage des familles spécifiques aux souches réduites ni celles spécifiques à MED4. test